UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

Centro de Biotecnologia da UFRGS

Departamento de Biologia Molecular e Biotecnologia

ESTUDO DA ATUAÇÃO DE SUBSTÂNCIAS DE ABUSO DURANTE O DESENVOLVIMENTO EMBRIONÁRIO POR MEIO DA QUÍMIO-BIOLOGIA DE SISTEMAS

BRUNO CÉSAR FELTES

Dissertação submetida ao Programa de Pós-Graduação em Biologia Celular e Molecular (PPGBCM) da UFRGS como requisito parcial para obtenção do grau de Mestre em Biologia Celular e Molecular.

Orientador: Prof. Dr. Diego Bonatto

Porto Alegre

Outubro de 2013

INSTITUIÇÕES E FONTES FINANCIADORAS

Agências financiadoras

Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq; outorga no. 474117/2010-3);

Programa Institutos Nacionais de Ciência e Tecnologia (INCT de Processos Redox em Biomedicina-REDOXOMA; outorga no. 573530/2008-4);

Fundação de Amparo a Pesquisa do Rio Grande do Sul FAPERGS (PRONEM outorga no. 11/2072-2);

CAPES (Coordenação de Aperfeiçoamento de Pessoal do Ensino Superior).

Instituição de origem

Laboratório de Radiobiologia Molecular, sala 219

Centro de Biotecnologia da UFRGS

Universidade Federal do Rio Grande do Sul

“A man is but the product of his thoughts. What he thinks, he becomes.”

-Mahatma Gandhi

AGRADECIMENTOS

Ah! A lista de agradecimentos! Aquele momento em nossas conquistas, que colocamos por escrito, em que tentamos recordar e organizar os pensamentos em forma de gratidão. Mas sejamos francos, pequenas frases jamais expressarão toda gratidão e apreço por aqueles que moldam nosso caminho ao longo da vida, ou durante os períodos conturbados da ascensão profissional. Contudo, darei o meu melhor para expressar em simples frases, criadas por pessoas infinitamente mais sábias que eu, o que sinto por todos que fizeram parte do meu caminho até então. Apenas saibam que tudo aqui expresso é apenas uma fração mínima da imensidão dos meus sentimentos.

*"*Se *eu vi mais longe, foi por estar de pé sobre ombros de gigantes" – Sir* Isaac Newton

Esta frase é para meus pais, Paulo Cesar Feltes e Heloísa Pedroso de Moraes Feltes (os meus gigantes), que sempre olham para seu filho e sentem orgulho; e que constantemente inundam meu coração de amor dizendo sou e serei um grande homem. Todos seus ensinamentos e sua dedicação me tornam não só mais forte, mas me dão coragem para enfrentar desafios. Se vocês acham que eu vejo mais longe, como sempre dizem, não esqueçam que é porque vocês me proporcionaram isso também. Minhas conquistas não são apenas minhas, são suas também.

"Amor não é amor se se altera quando encontra alterações ou se curva diante de força contrária" – William Shakespeare

Esse pensamento é para Joice, que assim como eu, jamais alterou seu amor, ou se deixou magoar pelas alterações ao longo do caminho. Que me ajudou e ajuda diariamente a ser uma pessoa melhor, e a arranjar coragem para seguir os caminhos tortuosos que desbravo. Da mesma forma ela me presenteia com o presente mais difícil de encontrar no universo: a felicidade.

"Desejar ser amigo é um trabalho rápido, mas a amizade é um fruto de amadurecimento lento" - Aristóteles

Esse pensamento é para todos meus amigos que ao longo dos anos moldaram meus pensamentos e minhas ações através do seu apoio, suas risadas e, acima de tudo, sua presença. São muitos amigos para nomear e não quero ser injusto em esquecer alguém. Contudo, aqueles que sempre estiveram presentes sabem do meu apreço, pois eu nunca escondi meus sentimentos. A amizade verdadeira é um tesouro raríssimo e eu posso dizer com orgulho que encontrei. Mesmo através de tempos ruins e desentendimentos, as amizades verdadeiras não desaparecem. Obrigado a todos que cruzaram e ainda atravessam as tempestades comigo.

"Recompensa mal um mestre aquele que se contenta em ser discípulo" – Friedrich Nietzsche

Essa frase é para meu orientador desde a iniciação científica, Diego Bonatto. Dentre as muitas conversas e interações que ocorreram ao longo dos anos, e o caminho científico que ele sempre passou adiante, eu com certeza assimilei e aprendi diversas lições, diretas ou indiretamente. Dentre elas, está o que agora é uma mania, de aperfeiçoar tudo aquilo que faço; ou seja, "se for para fazer, faça além!" Essa lição eu carregarei comigo para onde eu for e em todos os trabalhos que pretendo fazer, sejam eles relacionados à ciência, ou não. Essa lição é a que move um aluno a se tornar mais que um discípulo – o move a se tornar um mestre, assim como o mestre que o ensinou.

"Então nós crescemos juntos como uma dupla cereja, parecendo separadas, mas ainda assim uma união de partição, duas lindas frutas moldadas em um único ramo." – William Shakespeare

Um obrigado em especial deve ser dado a todos os meus colegas de trabalho. Sem eles para me alegrar em dias nublados, para me ajudar a pensar, sanar minhas dúvidas e me ajudar a crescer no ramo científico, o sucesso de terminar o mestrado não teria sido alcançado com tanto júbilo sem eles. Sabemos que crescemos como profissionais no ramo molecular e um dia iremos nos partir em diversos caminhos

diferentes. Contudo, todos os conhecimentos e experiências trocadas jamais sairão da minha mente.

"(...) algumas vezes as pessoas se sentem mais livres em falar com estranhos do que com uma pessoa conhecida. Por que isso? – Provavelmente porque um estranho nos vê do jeito que somos, e não como ele gostaríamos que fossemos" — Carlos Ruiz Zafón

Esse pensamento é dedicado a todos que cruzaram meu caminho até então. Pois todos esses estranhos, mesmo que de uma maneira ínfima, alteraram meu jeito de ser. Mesmo fornecendo bons ou maus exemplos, minha mente e minhas ações não foram simplesmente todas originais. As ações de todos nesse pequeno mundo foram baseadas, ou inspiradas, ou moldadas por alguma ação anterior. Somos originais do modo que as aplicamos, como as adaptamos a nossa vontade, e como as aperfeiçoamos, mas não podemos ser arrogantes, presunçosos ou ingênuos o suficiente para acharmos que elas não tiveram uma base.

ESTRUTURA DA DISSERTAÇÃO

Esta dissertação de mestrado é dividida em uma introdução geral, dois capítulos redigidos em forma de artigo, sendo o primeiro deles já publicado, uma discussão geral e conclusões. Da mesma forma, é incluído um item de adendos contendo outras produções durante o período de mestrado.

A introdução geral consiste em uma breve apresentação histórica da bioinformática e o nascimento da biologia de sistemas como ferramenta de análise, seguido da explicação sobre tipos de redes, principais elementos constituintes de um interatoma e parâmetros de análise topológica. Esta introdução à análise de redes de interação é seguida de uma breve introdução sobre a problemática abordada neste trabalho sobre o abuso de substâncias tóxicas durante o desenvolvimento embrionário.

O capítulo 1 consiste em uma análise sobre os efeitos toxicológicos de diferentes substâncias carcinogênicas durante o desenvolvimento embrionário em mulheres fumantes. Neste artigo é analisado com ferramentas de químio-biologia de sistemas e análise transcriptômica os possíveis efeitos, caminhos moleculares e alvos proteicos de moléculas pouco estudadas no seu potencial teratogênico. Deste modo, é proposta uma visão de quais caminhos moleculares poderiam afetar o desenvolvimento do feto. Esse artigo foi publicado no periódico *PLoS ONE*.

O capítulo 2 apresenta uma análise de químio-biologia de sistemas sobre como o etanol pode afetar o metabolismo de vitaminas e o neurodesenvolvimento em *Mus musculus*. Neste estudo, são avaliadas as mudanças moleculares observadas pré-natais e pós-natal de quatro diferentes conjuntos de dados transcriptômicos de tecido neural de ratos expostos ao etanol em diversos momentos do desenvolvimento e na fase adulta, buscando mecanismos pouco conhecidos associados com a Síndrome Alcoólica Fetal (SAF). Este artigo encontra-se em fase de revisão no periódico *Toxicological Sciences*.

Os dois capítulos são seguidos de uma discussão geral dos temas e dados abordados, incluindo novos dados e são finalizados com uma conclusão sobre a importância dos estudos realizados nessa dissertação de mestrado.

A dissertação é finalizada com um item de “Adendos”, onde é mostrado um terceiro artigo que prospecta novas combinações de pequenas moléculas e alvos proteicos para serem inibidos por fármacos visando o aprimoramento da reprogramação celular. Neste estudo foram utilizados das ferramentas de análise topológica para propor novas combinações de pequenas moléculas para a geração de células pluripotentes induzidas. Nesse mesmo trabalho também foram avaliados, por meio de análises de centralidades e de interferência, novos alvos proteicos para serem inibidos por pequenas moléculas e, assim, promover os mecanismos de indução ou de manutenção do estado-tronco pluripotente. Este artigo encontra-se aceito para publicação no periódico *Molecular BioSystems* e já se encontra disponibilizado online.

No item “Adendos” também se encontra um capítulo de livro que mostra a história, teoria e aplicações da biologia de sistemas.

SUMÁRIO

LISTA DE ABREVIATURAS

AR – Ácido Retinóico

BING2 – *Death-Domain Associated Protein*

CPI – *Chemical-Protein Interaction* **–** Interação Químico-Proteína

CYP – *Cytocromome* P450

EGEP-*Network* – *Early-Gestation-Exposed-Postnatal-Network*

FAS – *Fetal Alcohol Syndrome*

FDFT1 – *Farnesyl-Diphosphate Farnesyltransferase* 1

FDPS – *Farnesyl Diphosphate Synthase*

Fgf – *Fibroblast Growth Factor*

Fgfr1op – FGF *receptor 1 Oncogene Partner*

Gas7 – *Growth Arrest-Specific Protein* 7

GEO – *Gene Expression Omnibus*

GST – *Glutathione S-transferase*

GSTM1 – *Glutathione S-transferase Mu* 1

HG – *Hub*-gargalo

HPRT1 – *Hypoxanthine Phosphoribosyltransferase*

Hox – *Homeobox*

ITPA – *Inosine Triphosphatase Protein*

LGEP-*Network* – *Late-Gestation-Exposed-Postnatal-Network*

MDM2 – P53 E3 *Ubiquitin Protein Ligase Homolog*

P53 – *Tumor protein* p53

PAH – *Polycylic Aromatic Hydrocarbons* – Hidrocarbonetos Policíclicos Aromáticos

PE-*Network* – *Prenatally-Exposed-Network*

PGH – Projeto Genoma Humano

PLM – *Phospholemman*

POMC – *Proopiomelanocortin*

PPI – *Protein-Protein Interaction* – Interação Proteína-Proteína

PSE-*Network* –*Postnatal-Exposed-Network*

SAF – Síndrome Alcoólica Fetal

SNC – Sistema Nervoso Central

StAR – *Steroidogenic Acute Regulatory Protein*

TC – *Tobacco Component* – Componente do Tabaco

UGT – UDP *glucuronosyltransferase*

USP2 – *Ubiquitin Specific Peptidase* 2

RESUMO

Muitos caminhos bioquímicos e interações moleculares ainda são pouco conhecidos para as ciências biomédicas, dentre elas a ação de pequenos compostos tóxicos no desenvolvimento embrionário de diferentes modelos biológicos. Neste sentido, dois cenários críticos, de amplo interesse clínico e de impacto social, se destacam: o abuso de tabaco e de bebidas alcoólicas. Sabe-se que os derivados químicos do tabaco e do etanol são capazes de alterar o funcionamento de diferentes vias bioquímicas, levando a modificações. Por exemplo, crianças nascidas de mulheres fumantes expostas ou usuárias de tabaco mostram inúmeras alterações morfológicas e funcionais em diferentes tecidos do seu organismo. Da mesma forma, o abuso de álcool durante a gravidez é responsável por danos ao tecido neural do feto que, ao nascer, apresenta problemas cognitivos, motores e de aprendizado que se agravam ao longo da vida. Esse quadro patológico é chamado de Síndrome Alcoólica Fetal (SAF). Infelizmente, devido à complexidade inerente dos sistemas bioquímicos, os mecanismos moleculares subjacentes a ambos cenários são escassos.

Assim, essa dissertação de mestrado visa aplicar diversas ferramentas de químio-biologia de sistemas para elucidar os possíveis alvos e caminhos moleculares relacionados às anomalias geradas pelo abuso de tabaco e à SAF. Para tanto, análises topológicas globais e locais de redes de interação foram empregadas juntamente com informações transcritômicas para ambas as condições de uso de tabaco e de etanol em modelo humano e murino (*Mus musculus*). As análises dos efeitos do tabaco no desenvolvimento mostram que o abuso deste resulta na alteração na biossíntese de prostaglandinas e leucotrieno, assim como na regulação negativa de genes HOX e receptores de ácido retinóico. Da mesma forma, foi possível identificar diversas proteínas relacionados a diferenciação celular e formação do tecido ósseo. Por fim, as análises dos efeitos do etanol no neurodesenvolvimento indicam que o etanol afeta a diferenciação neural e importantes processos como a via de glutamato e o metabolismo de diferentes vitaminas. As análises indicam que o etanol pode causar graves quadros de neuroinflamação. Também se observou que diversas vitaminas têm a sua biossíntese e o seu metabolismo alterado pelo etanol, com importantes implicações no neurodesenvolvimento.

ABSTRACT

Many pathways and molecular interactions are still poorly described in biomedical sciences. Among these pathways, the knowledge related to the action of toxic compounds during embryonic development is largely unknown. In this sense, two scenarios, of broad clinical and social impact stand out: the abuse of tobacco and alcohol in the form of fermented or distillates. It is known that the chemical derivatives of tobacco and ethanol are capable of alter the functionality of different biochemical pathways. For example, children born from smoking abusing women or exposed to tobacco smoke show innumerous morphological alterations in different tissues. In addition, the abuse of alcohol during pregnancy is responsible for damages in the fetus neural tissue. Those fetuses, after birth and during growth, present cognitive, motors and learning problems that aggravate in the course of life. This pathology is called Fetal Alcohol Syndrome (FAS). Unfortunately, due to the inherent complexity of biochemical systems, the molecular mechanisms underlying both scenarios are scarce and poorly understood.

Thus, this master's degree dissertation aim to apply different chemo-systems biology tools to elucidate the possible molecular pathways and potential targets related to the tobacco and SAF-related anomalies. For such, local and global topological analyses from interaction networks were employed together with transcriptomic information to both conditions of tobacco and ethanol abuse in human and mice (*Mus musculus*).

The analysis of the effect of tobacco during development shows that the abuse of this drug results in the alteration of prostaglandin and leukotriene biosynthesis, as well for a negative regulation of HOX gene receptors and retinoic acid. Moreover, it was possible to identify different proteins related to osteogenesis. Finally, the analysis of the effects of ethanol in neurodevelopment indicate that ethanol impair neural differentiation and essential process, such as glutamate pathway and the metabolism of different vitamins. The gathered data also propose a model where ethanol can promote severe neuroinflammation. In addition, was observed that multiple vitamins had their biosynthesis and metabolism impaired by ethanol, with crucial implications for neurodevelopment**.**

1. INTRODUÇÃO

1.1. A era pós-genômica e a bioinformática

Em busca do entendimento da organização e função do genoma humano e de outros organismos, em 1990 foi lançado pelo *U.S Department of Energy* e o *National Institute of Health* dos Estados Unidos da América o Projeto Genoma Humano (PGH). O projeto tinha como objetivo mapear o genoma de humanos em busca de obter plataformas de comparação da organização genômica e avançar na área biomédica (Lander *et al*., 2001; Consortium, International Human Genome Sequencing, 2004).

Dentre o vasto volume de dados gerados pelo próprio PGH e também aqueles resultantes de análises posteriores de larga escala destacam-se: (i) a caracterização e a anotação funcional de genes; (ii) diferenças na composição dos genomas, como o número de sequências codificantes e conteúdo de cada nucleotídeo presente no DNA; (iii) relação detalhada entre recombinações e distâncias físicas entre os genes e (iv) arquitetura combinatória das proteínas, identificando regiões multidomínios e múltiplas subunidades (Consortium, International Human Genome Sequencing, 2004). Desta forma, diferentes tipos de bancos de dados foram criados para estocar as mais diversas informações biológicas, onde ferramentas computacionais se tornaram fundamentais para comparar e analisar as múltiplas sequências genômicas obtidas pelo PGH, bem como outras informações complexas. Sendo assim, a quantidade maciça de dados gerados pelo PGH e outros projetos genômicos se tornaram disponíveis para a comunidade científica, permitindo suas análises em diversos campos das biociências.

As ferramentas computacionais aplicadas para a análise e a preparação de dados provindos de sistemas biológicos são objeto de estudos da área da Bioinformática. Mesmo que o termo “bioinformática” tenha sido cunhado na década de 1970 com o propósito de ser o “estudo dos processos informativos dentro de um sistema biológico”, a bioinformática foi classicamente ligada a análise computacional de comparações genômicas (Hogeweg, 2011). Contudo, a necessidade emergente de fazer sentido à grande abundância de dados gerados pelo PGH, contribuiu para o aperfeiçoamento de ferramentas computacionais cada vez mais especializadas para lidar com os diferentes tipos de dados.

O desenvolvimento dessas ferramentas de análise também foi fundamental para a contextualização das informações biológicas no sentido de possibilitar a descoberta de como os mecanismos moleculares funcionam em um determinado organismo (Sobral, 1999).

Com a expansão das técnicas de análise de larga-escala (por exemplo, microarranjos de DNA, RNAseq e análises proteômicas), a demanda de ferramentas para análise de dados cresceu, uma vez que esses dados encontravam-se fragmentados em bancos de dados. Neste sentido, a bioinformática também ganhou papel fundamental para geração das "anotações genômicas" (Ouzounis, 2012). Essas anotações, chamadas de ontologias gênicas, provinham de experimentos de larga-escala e compilavam informações sobre qual a função do produto genômico, sua localização celular e de qual processo biológico ele participava (Consortium, Gene Ontology, 2001). Assim, foi iniciado o *Gene Ontology (GO) Project* que visava compilar, tratar e armazenar as anotações referentes a processos biológicos em diferentes organismos (Consortium, Gene Ontology, 2001). Finalmente, as técnicas de análise de larga-escala não exigiam análises individuais de dados, mas sim um procedimento onde fosse possível contemplar dados fragmentados e os processos biológicos associados às amostras estudadas, dando sentido ao que biologicamente estava ocorrendo nos experimentos. Essa demanda foi a principal responsável pelo surgimento de uma ramificação da bioinformática dedicada, em parte, a esse propósito e que foi denominada de Biologia de Sistemas.

1.2. A biologia de sistemas e o estudo da complexidade biológica

Durante muitos anos a visão reducionista foi o foco das ciências biológicas, gerando informações sobre as partes individuais de uma célula (Palsson, 2000). Entretanto, dados fragmentados de análises de pequena escala (por exemplo, interação entre dois ou mais genes/proteínas), embora fundamentais, não conseguiam expressar a complexidade da organização celular, uma vez que essa complexidade emerge da interação dos componentes celulares como um todo (Barabási & Oltavai, 2004).

Neste sentido, a biologia sofreu uma gradual mudança durante a era pós-genômica, pois o crescimento de dados "ômicos" (isto é, genômica, proteômica,

trascritômica, metabolômica, entre outros) exigiu ferramentas para a análise global desses dados. Sendo assim, dentre as formas de análise está a geração e o desenho de redes de interação (interatomas) que derivam da inclusão dos diferentes tipos de dados biológicos (por exemplo, gene-proteína, proteína-metabólito, transcrito-proteína, entre outros).

Essas redes de interações ou grafos seguem propriedades matemáticas bem definidas cuja compreensão é aplicável para a correta compreensão dos dados biológicos.

1.2.1. Elementos básicos de redes, as suas topologias e as suas classificações

As redes ou grafos de interação são constituídos de dois elementos principais: (i) vértices ou nós (*nodes*) (**Fig. 1**), que representam as partes interagentes do sistema (Barabási & Oltavai, 2004; Newman, 2003) e sendo estas formadas por proteínas, genes, metabólitos, ácidos nucleicos, RNAs, carboidratos ou qualquer outro componente celular, moléculas sintéticas (por exemplo, fármacos), elementos inorgânicos (isto é, íons metálicos) ou quaisquer moléculas de importância biológica e (ii) conectores (*edges*) (**Fig. 1**), que indica o tipo de conexão entre os nós (Barabási & Oltavai, 2004; Newman, 2003). Os conectores podem significar os domínios interatores de proteínas ou até mesmo uma modificação pós-traducional passada de uma proteína para outra.

Um terceiro fator que deve ser levado em consideração para a análise dos dados de biologia de sistemas é o ambiente amostral. Neste caso, é necessário e fundamental entender o contexto biológico em que a rede interatômica está inserida. Por exemplo, é esperado que proteínas relacionadas à diferenciação neural se encontrassem em uma rede focada em neurodesenvolvimento.

Por outro lado, a topologia de qualquer tipo de rede reflete diretamente na sua funcionalidade (Scardoni & Laudanna, 2012). Por exemplo, a topologia de redes ferroviárias e rodoviárias influencia diretamente o trafego que passam por elas (**Fig. 1**). Da mesma forma, as ligações interpessoais em grupos sociais influenciam a propagação de informações e doenças. Sendo assim, o entendimento da estrutura de uma rede se torna crucial para o entendimento de como suas partes funcionam como um todo.

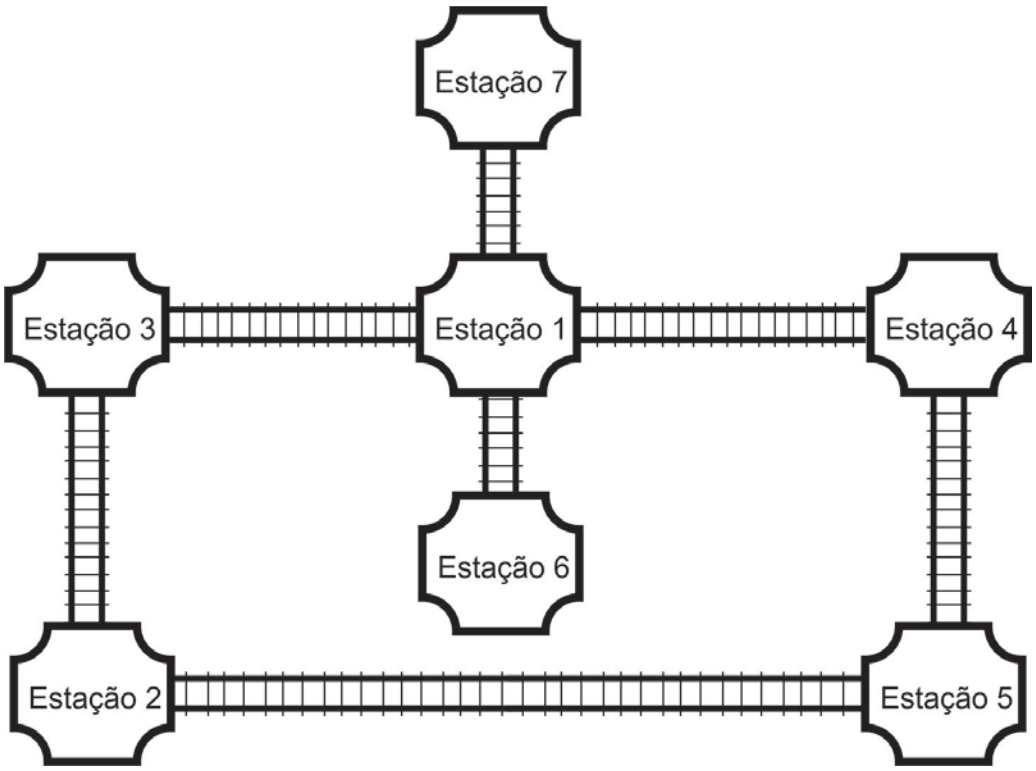


Figura 1. Exemplo de uma topologia de redes ferroviárias. A topologia influencia diretamente a funcionalidade de uma rede. Na figura é possível ver que a Estação 1 é o ponto de partida e chegada de quatro estações diferentes (Estações 3, 4, 6 e 7). Desta forma, é esperado que o tráfego na Estação 1 seja maior e que haja a necessita de um controle mais rígido de horários de partidas e chegadas quando comparado com outras estações.

Para compreender as redes de interação é necessário, primeiramente, definir os tipos de conectores existentes. Conforme a interação, uma rede pode ser do tipo "redes dirigidas" ou "redes bidirecionais" (**Fig. 2**).

A rede dirigida mostra que a informação se propaga apenas no sentido indicado pelos conectores (**Fig. 2A**). Já as redes bidirecionais não mostram o sentido que cada conector segue e ilustram que a informação se propaga em ambas as direções (**Fig. 2B**). De uma forma geral, a rede direcionada possui uma menor plasticidade, ou seja, é mais propensa a sofrer grandes modificações caso alguma de suas partes sofra alguma interferência negativa, do que a rede bidirecional.

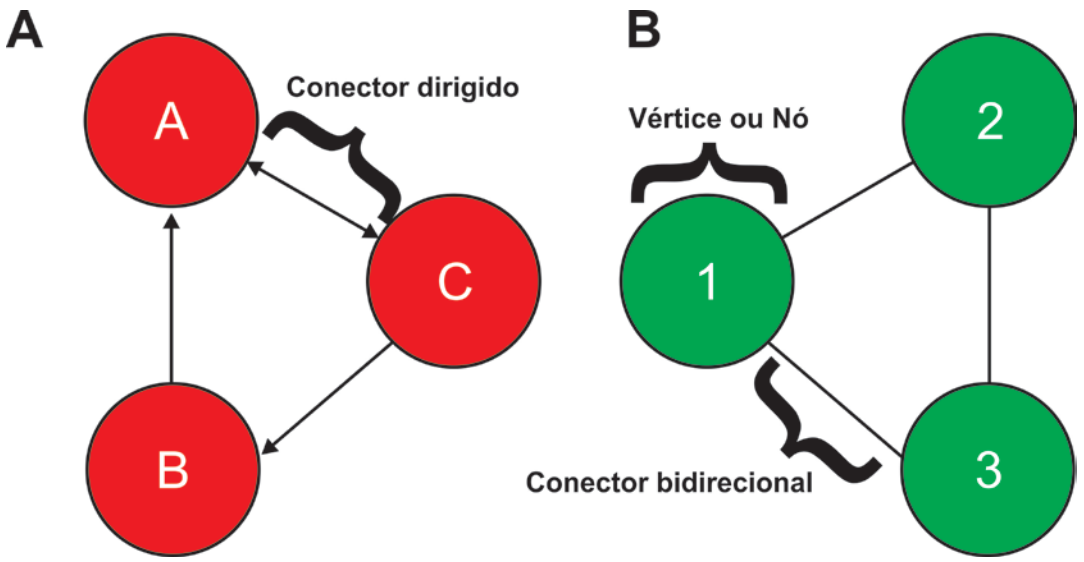


Figura 2. Elementos de uma rede e os dois principais tipos de redes encontradas em sistemas complexos. Em (**A)** é mostrada uma rede dirigida enquanto que em (**B)** é mostrada uma rede bidirecional. Na rede **(A)** os conectores mostram o sentido em que a informação se propaga. Desta forma, o nó "B" apenas pode propagar informação para "A", enquanto "A" só pode propagar para "C", mas "C" pode propagar para "A" e "B". Já na rede (B) os conectores não mostram o sentido da informação, indicando que ela se propaga nos dois sentidos para cada conjunto de nós.

Sabendo que uma rede pode ser dirigida ou bidirecional, e tendo em vista os seus principais elementos, existem três modelos definidos de redes de interação: (i) as redes de livre escala, que são definidas por uma lei de potenciação, que indica que a rede possui uma dinâmica de estruturação que permite o crescimento da rede pela adição de novos nós. Desta forma, as redes consistem de um sistema aberto que inicia com um pequeno grupo de nós e aumenta de tamanho exponencialmente no tempo devido à inserção de novos nós (Barabási & Oltavai, 2004) (**Fig. 3A**). A lei de potenciação indica que os novos nós adicionados terão mais probabilidade de se ligar a outros nós que possuem um maior número de conexões (Barabási & Oltavai, 2004). Desta forma, haverá uma quantidade menor de nós com uma conectividade acima da média de conectividade da rede, e uma maior quantidade de nós com baixa conectividade; (ii) rede aleatoriamente conectada (randômica), onde todos os nós possuem uma probabilidade semelhante de conter o mesmo número de ligações e não é regida por uma lei de potenciação (Barabási & Oltavai, 2004) (**Fig. 3B**) e (iii) rede

hierárquica, que demonstra que a maior parte dos nós terá uma conectividade semelhante e na sua arquitetura ela mostra de uma forma evidente os graus de hierarquia (Barabási & Oltavai, 2004) (**Fig. 3C**).

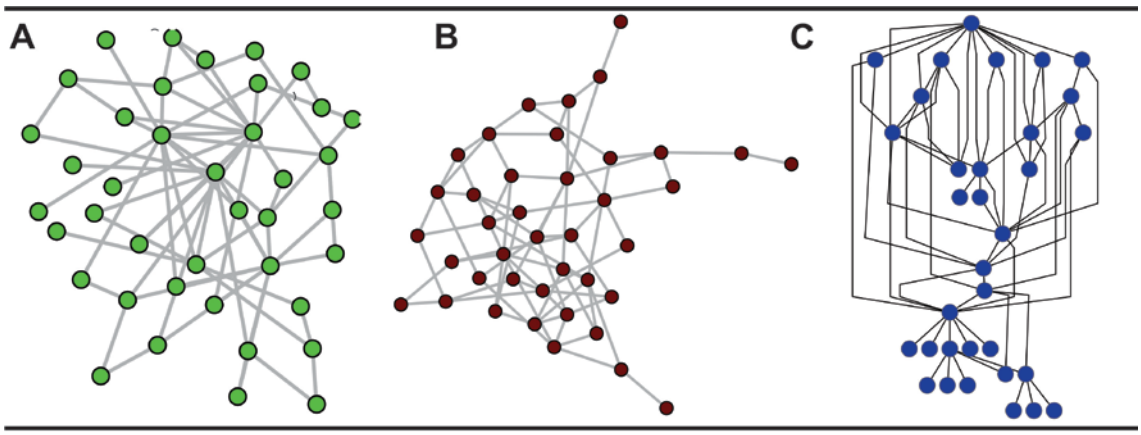


Figura 3. Os três principais modelos de redes. Em **(A)** rede de livre escala, em **(B)** rede randômica e **(C)** rede hierárquica. É possível observar que a rede randômica apresenta um número aproximadamente igual de conexões por nó (aproximadamente três conexões por nó), enquanto a rede de livre-escala apresenta apenas alguns nós mais conectados. A rede hierárquica claramente mostra os níveis de importância de cada nó em sua arquitetura.

Redes randômicas e hierárquicas não são topologicamente adequadas para modelar conexões em um sistema biológico, uma vez que as interações proteínas-proteínas e genes-proteínas não são aleatórios em sistemas vivos (Newman, 2003). Da mesma forma, a organização molecular da célula não é hierárquica. Sendo assim, as redes de transmissão de energia elétrica, ferrovias e as observadas *in vivo* são de livre-escala (Barabási & Oltavai, 2004; Newman, 2003).

Uma vez definido o que é uma rede e seus elementos, assim como os tipos de redes existentes, é importante analisar como seus componentes se organizam de forma topológica. Neste sentido, é especialmente importante analisa-las com o que diz respeito à modularização e na descrição dos seus elementos centrais.

1.3 Parâmetros de análise de redes de livre-escala

1.3.1 Modularização ou clusterização

O parâmetro topológico de clusterização, também chamado de modularidade, se baseia em um princípio de agrupamento entre partes individuais de um sistema, ou seja, na formação de regiões altamente conectadas (Wagner *et al*., 2007). Este princípio de organização intrínseco de um determinado sistema é observado evolutivamente em diversas ocasiões como, por exemplo, na interação entre organismos da mesma espécie, como a mosca chamada de *cluster fly* (**Fig. 4**), que ganhou esse nome pelo fato de formar agrupamentos compostos por organismos da mesma espécie. A modularidade também é vista na organização de itens por sua funcionalidade, como uma caixa de ferramentas ou uma gaveta de meias, ou pela sua semelhança (como uma caixa de pregos, por exemplo). Ela também é observada na nossa tendência de formarmos grupos sociais por afinidade (isto é, afinidade por pensamentos, crenças ou inclinações profissionais). A tendência de formar grupos altamente conectados também é relacionada à sobrevivência, onde muitos animais se organizam em grandes grupos visando afastar predadores.

Figura 4. Organismos da espécie *Pollenia rudis* (*cluster fly*) agrupados. Fonte: http://www.thesuffolkpestcontrolcompany.co.uk/pest-help/cluster-flies

Evolutivamente, todos os organismos também possuem sistemas moleculares organizados que proporcionam interação entre suas biomoléculas e entre o organismo e o seu ambiente (Wagner *et al*., 2007). Esses sistemas são também chamados de

clusters ou módulos e influenciam diretamente na formação das redes moleculares e/ou na dinâmica evolutiva da espécie (Wagner *et al*., 2007). Esses módulos podem ser conservados entre diferentes espécies, como as proteínas que compõe o citoesqueleto em células eucarióticas e os complexos de manutenção genômica de replicação e transcrição, por exemplo. Os elementos que compõe módulos estáveis e vantajosos para o organismo, no ponto de vista de adaptação, tendem a sofrem maior pressão seletiva entre gerações (Wagner *et al*., 2007).

Deste modo, a análise de *clusters* se torna fundamental no entendimento da dinâmica molecular e organização do sistema. Esse objetivo é possível pela aplicação de programas que calculam os graus de modularização representada pela equação 1:

$$C_i = \frac{2n}{k_i(k_i - 1)} \quad (1)$$

Onde k_i é o tamanho da vizinhança de vértices (nós) do vértice *i* e *n* o número de conectores na vizinhança. Desta forma, quanto maior for C_i, mais conectado é o *cluster* (**Fig. 5**).

A vantagem do estudo da modularização em um interatoma é a possibilidade de observar novos grupos de elementos altamente conectados e identificar novas funções para biomoléculas, assim como elucidar novas vias bioquímicas que poderiam ser fundamentais para a estabilidade de algum processo.

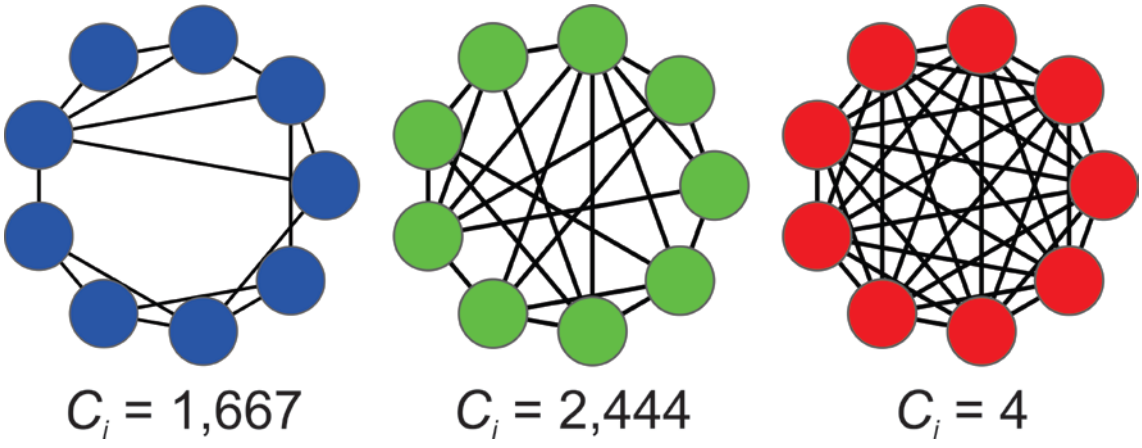


Figura 5. Módulos e seus respectivos valores de C_i. A figura ilustra que a medida que um módulo aumenta a conectividade entre seus nós, o valor de C_i aumenta.

1.3.2 Centralidades

1.3.2.1 Grau de nó

Um dos parâmetros básicos para o entendimento de análises topológicas é o parâmetro de grau de nó (*node degree)*. Neste parâmetro é calculado o número de nós diretamente conectadas a outro nó (Scardoni & Laudanna, 2012). Sendo assim, o grau de nó leva em consideração os conectores diretos que incidem e partem de um determinado nó.

Os nós que possuírem um valor de grau maior que a média de grau de nó da rede são chamados de *hubs* (Scardoni & Laudanna, 2012) (**Fig. 6**). Este parâmetro é dado pela equação 2:

$$Deg(v) = \sum E_i \qquad (2)$$

Onde *Deg(v)* é à medida que indica o número de conexões (*Ei*) que passam pelo nó (*v*).

Os *hubs* podem ser exemplificados como indivíduos formadores de opiniões em grupos e redes sociais ou supervisores e chefes de empresas que possuem muitos empregados. Esses indivíduos possuem uma ampla gama de contatos, seguidores e/ou apreciadores, se tornando pontos críticos da formação de grupos. Deve-se ter em mente que o *hub* não indica a importância de um nó e sim a sua popularidade.

Em termos biológicos é possível exemplificar os *hubs* como proteínas que possuem um grande número de biomoléculas como parceiras diretas como, por exemplo, fatores de transcrição que pode interagir com uma ampla gama de proteínas (Ferecatu *et al*., 2009; Scardoni & Laudanna, 2012) e mediar a ativação de distintos bioprocessos e a transcrição de diversos genes.

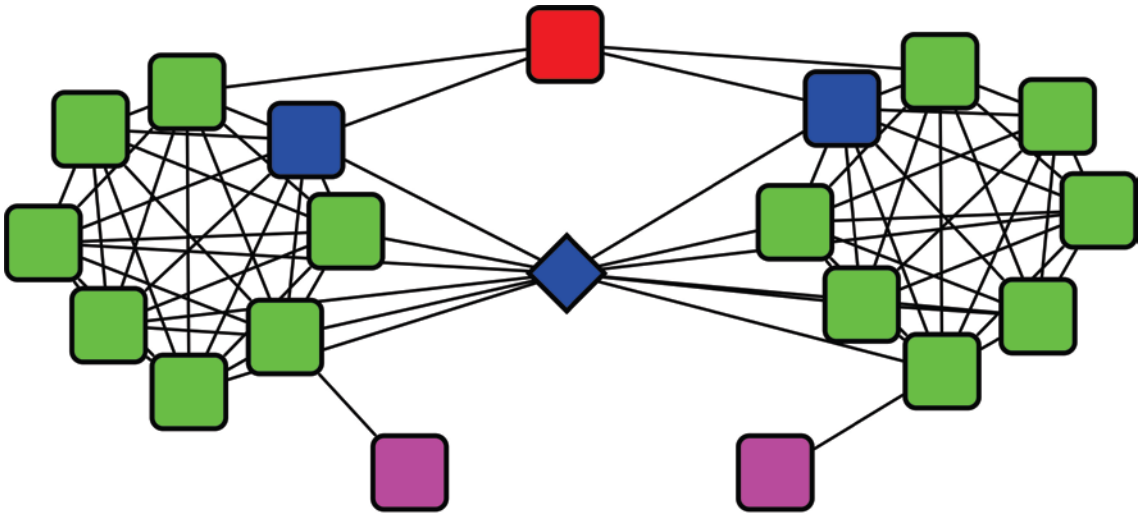


Figura 6. Centralidades em uma rede de livre-escala. Na figura é possível observar os tipos principais de centralidades. Em verde encontram-se os nós classificados como *hubs*, ou seja, todos aqueles que possuem uma conectividade maior do que a média de conectividade da rede. Já o nó em vermelho indica que o mesmo possui um alto valor de *betweenness* e é classificado como gargalo. Sendo assim, ele se torna essencial para a ligação entre diferentes grupos de nós. Os nós em azul são classificados como *hubs*-gargalos (HG) possuindo, assim, um alto valor de *betweenness* e grau de nó. Deste modo, os HGs são os nós mais cruciais para um sistema, topologicamente falando. Contudo, o nó com a forma de diamante é o único com um valor de *closeness* acima da média, ou seja, é o nó mais próximo de todos os outros nós da rede. Isso é facilmente visto, pois os outros nós azuis, embora do tipo HGs, não interconectam os módulos como o nó em forma de diamante. Por fim, os nós em roxo são aqueles nomeados de não-*hub* não-gargalo, pois não possuem um valor acima da média para ambos parâmetros.

1.2.2.3 *Betweenness*

O parâmetro de *betweenness* é referente ao trafego de informação que passa por um determinado nó (Scardoni & Laudanna, 2012). Ou seja, o *betweenness* considera os caminhos mais curtos que passam entre dois nós, como é dado pela equação 3:

$$Bet(v) = \sum_{s \neq v \neq t} \frac{\sigma_{sw}(v)}{\sigma_{sw}}$$

(3)

Onde σ_{sw} é o número total de caminhos mais curtos que passam do nó *s* para o nó *w* e σ_{sw} *(v)* é o número desses caminhos que passam por aquele nó.

Sendo assim, estes nós com alto valor de *betweenness* são cruciais para ligar diferentes módulos (**Fig. 6**) e são chamados de gargalos ou *bottlenecks*. Os gargalos não necessariamente possuem muitas conexões com outros nós, mas podem também ser *hubs* (isto é, possuir um grau de nós acima da média de grau de nós da rede). Neste caso, quando um gargalo também é um *hub* ele ganha a denominação de *hub*-gargalos (HGs) (**Fig. 6**). Os HGs são os nós com maior probabilidade de possuir grande importância topológica em uma rede de interação, pois eles combinam a característica de muitas conexões do *hub*, com a capacidade de ser um nó comunicante entre vários processos, que é característica do gargalo (Yu *et al*., 2007).

Um alto valor no parâmetro de *betweenness* pode ser exemplificado de uma maneira prática como as vias de transito. Neste sentido, se um carro saiu da cidade A em direção à cidade B, as vias mais utilizadas serão aquelas que forem o meio de chegada mais rápido e com menos paradas entre as duas cidades. A demolição dessas vias rápidas causaria transferência de transito para outras vias, causando um aumento de tráfego (**Fig. 7**).

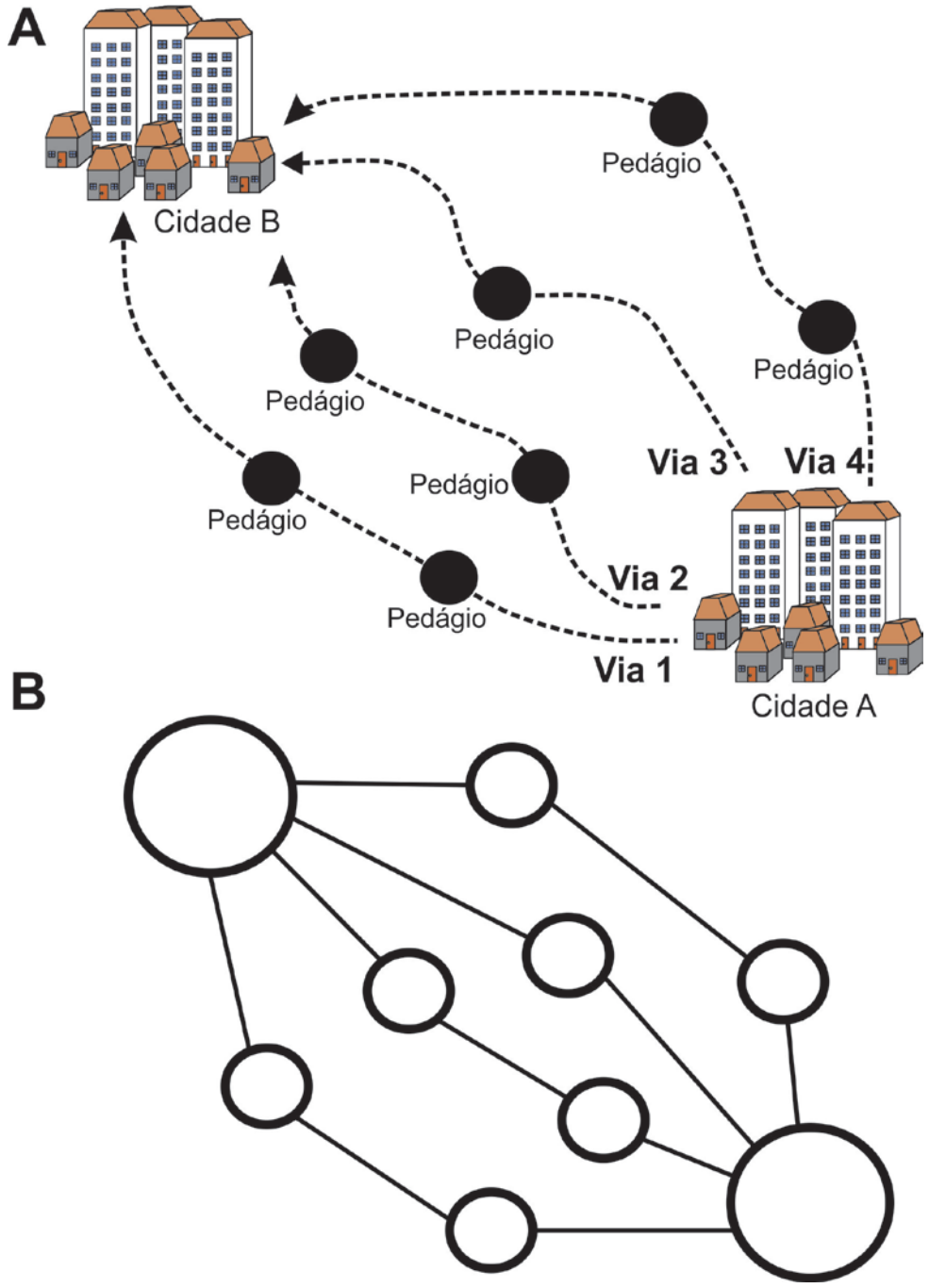


Figura 7. Exemplo prático de *betweenness*. **A)** A figura mostra as vias hipotéticas que interligam duas cidades. A via mais utilizada para viajar entre as duas cidades será a via que possuí menos pedágios (Via 3). Consequentemente, o fluxo de carros que passará no pedágio da Via 3 será maior do que nas outras vias. Deste modo, este pedágio terá um *betweenness* maior do que os outros pedágios. **B)** A mesma visão das cidades em **A**, porém na visão de redes.

Em redes biológicas, o parâmetro de *betweenness* pode ser ilustrado como uma proteína capaz de interligar diferentes processos biológicos distintos. Nesse sentido, a proteína p53 pode ser usada como exemplo, pois a mesma participa como uma ponte entre processos, como ciclo celular, reparo de DNA e apoptose (Ferecatu *et al*., 2009), onde diversos sinais, provindos de proteínas diferentes, farão com que p53 dirija o destino da célula para processos distintos (**Fig. 8**). Sua deleção levaria a inativação ou defeitos em todas as vias que ela participa.

BOX1

P53: fator de transcrição envolvido na resposta a danos de DNA em G_1/S e G_2/M, através da sua ativação e recrutamento de diversas proteínas ativadoras de apoptose (por exemplo, Puma, Bax) e repressão de genes anti-apoptóticos (por exemplo: Bcl2) (Ferecatu *et al*., 2009). Em condições normais p53 é degradada por Mdm2 através de poli-ubiquitinação que leva à degradação proteossomal. (Ferecatu *et al*., 2009).

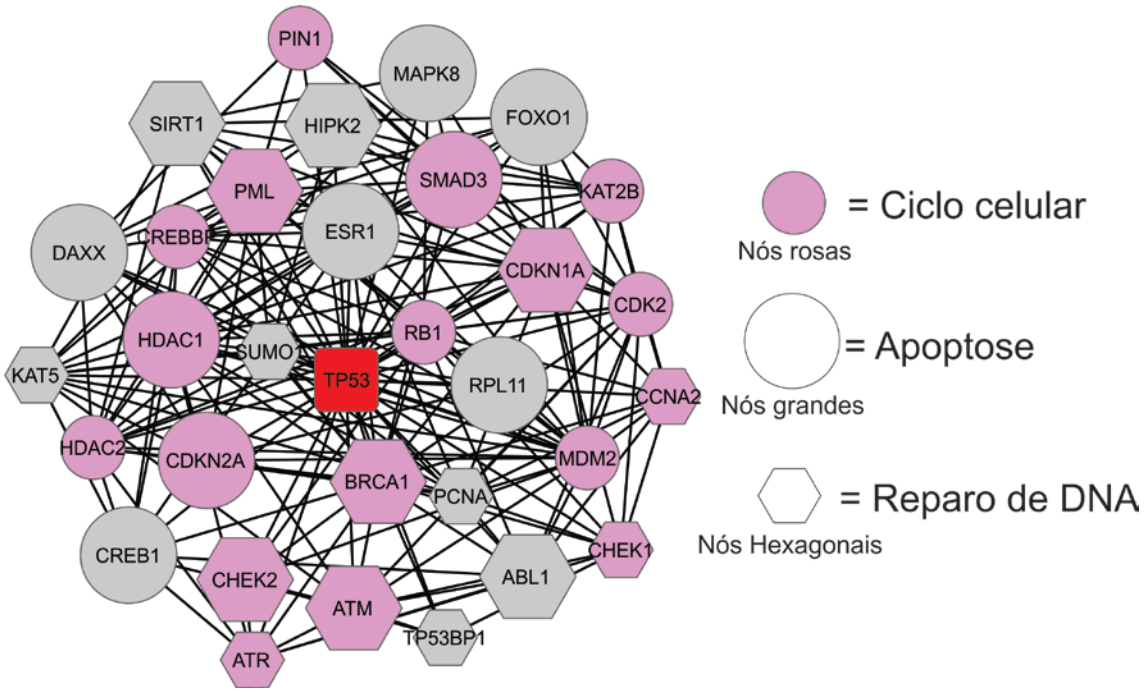


Figura 8. Rede de interação proteína-proteína mostrando alguns dos processos biológicos em que p53 participa. Todas as proteínas na figura se conectam a p53 e participam de diversos processos biológicos. A rede foi prospectada no programa STRING 9.05 [http://string-db.org/] e montada no programa Cytoscape 2.8.3 [http://www.cytoscape.org/].

1.2.2.4 *Closeness*

O *closeness* leva em consideração os caminhos mais curtos entre um vértice e todos os outros vértices da rede, sendo classificado como um parâmetro de isolamento ou aproximação de um nó (Scardoni & Laudanna, 2012). Desta forma, um alto valor de *closeness* indica que um determinado vértice é "próximo" de um grande número de outros vértices, enquanto um baixo valor de *closeness* mostra que essa proteína é afastada da maior parte dos outros vértices de um sistema, encontrando-se, de certa forma, isolada (Scardoni & Laudanna, 2012). Sendo assim, quanto maior o valor de *closeness* de um determinado vértice, mais relevante topologicamente para a propagação de informação entre nós vizinhos ele se torna (**Fig. 6**). Este parâmetro é dado pela seguinte equação 4:

$$Clo(v) = \frac{1}{\sum w \in v^{dist(v,w)}} \quad (4)$$

Onde o valor de *closeness* de um vértice *v* [*Clo(v)*] é calculado através da soma de caminhos mais curtos entre um vértice *v* e todos os outros vértices *w* [*dist(v,w)*] em um interatoma.

No quesito biológico podemos usar mais uma vez a proteína p53 como exemplo, pois uma vez ativa ela coordenará diversos processos e proteínas simultaneamente ou em diferentes momentos (Ferecatu *et al*., 2009) (**Fig. 9**). Ela não só coordena as proteínas diretamente transcritas por ela, como possui um impacto direto em diversos processos biológicos.

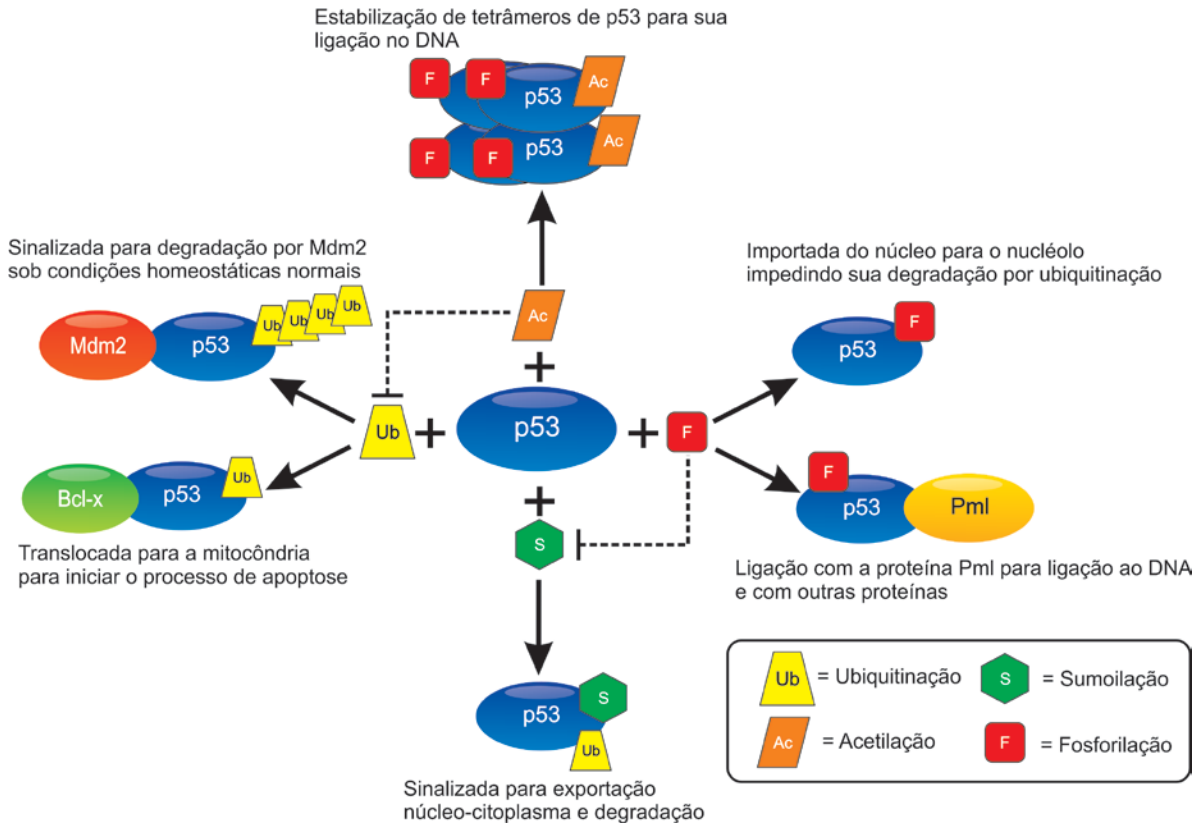


Figura 9. Exemplo de *closeness*. A proteína p53, dependendo da sua modificação pós-traducional atua em diferentes funções ou é levada para diversos destinos. No esquema acima é visto que p53 pode ser tanto translocada para a mitocôndria se ligando a Bcl-x para começar o extravasamento de citocromo-c e o processo de apoptose, quanto pode ser poli-ubiquitinada e ser encaminhada para degradação proteossomal. Por outro lado, ela também pode ser translocada do núcleo para o citoplasma por sumoilação, um processo que pode ser inibido pela fosforilação de p53. Ademais, a proteína p53, uma vez acetilada e fosforilada é estabilizada na forma de tetrâmetros para sua ligação no DNA como fator de transcrição. Por fim, p53 fosforilada tanto estabiliza a proteína quanto impede sua degradação e também permite sua ligação com Pml que, ao formar o heterodimero com p53, serve como uma plataforma para ligação de outras proteínas que irão transcrever genes alvo.

BOX2

MDM2: em condições normais MDM2 se encontra no núdeo e é translocada para o citoplasma onde promove a degradação de proteínas-alvo, promovendo poli-ubiquitinação e subsequente degradação por proteassomas (Nag *et al*., 2013).
BCL-X: proteína anti-apoptótica que quando ligada a p53 tem sua atividade inibida (Ferecatu *et al*., 2009).

Vale ressaltar que nem todo HG possui um valor alto de *closeness*, como é visto na **Fig. 6**. Na figura é possível observar que apenas o nó central possui um valor do parâmetro de *closeness* acima da média, tornando-se o nó "mais próximo" de todos os outros.

1.4. Abuso de tabaco e etanol durante o desenvolvimento embrionário em humanos e modelo murino

A biologia de sistemas tem sido aplicada amplamente para a prospecção de fármacos para serem utilizada no tratamento de doenças (Csermely *et al*., 2013; Chandra & Padiapu, 2013; Rosado *et al*., 2011), na elucidação de mecanismos moleculares associados a neurogênese e envelhecimento (de Faria Poloni *et al*., 2011; de Magalhães & Toussaint, 2004; Feala *et al*., 2013; Feltes *et al*., 2011) e para aplicações biotecnológicas (da Hora Junior, *et al*., 2012; Kildegaard *et al*., 2013). Ou seja, todo e qualquer cenário em que haja dificuldades na compreensão de mecanismos moleculares ou que necessitam de informações e gerações de hipóteses para um avanço rápido da área se beneficiam da análise de redes.

Neste sentido, um dos cenários de relevância biomédica que ainda carece de informações e que potencialmente pode ser favorecido por este tipo de análise é a atuação de substâncias narcóticas durante o desenvolvimento embrionário. Infelizmente, os dados referentes aos mecanismos moleculares associados ao abuso de tabaco e o álcool durante o a gravidez são escassos e os estudos focam-se mais profundamente a nível morfológico e sintomático. Por outro lado, devido aos problemas éticos e práticos para segurança do feto e da mãe, o estudo molecular subjacente à exposição destas substâncias continua confuso e pouco descrito.

Neste sentido, entendendo a necessidade de avançar no estudo das patologias e anormalidades morfológicas causadas pela exposição dos fetos a essas substâncias, tanto pelo abuso de álcool, quanto pelo consumo de tabaco durante a gravidez, ambas as problemáticas serão abordadas para as análises de biologia de sistemas na presente dissetação

1.4.1 Abuso de tabaco

O tabaco e os seus derivados de combustão contém mais de 4.800 substâncias tóxicas, onde muitas são consideradas carcinogênicas (Pfeifer, *et al*., 2002). Dentre essas substâncias destacam-se os óxidos de nitrogênio, o butadieno, o isopreno, o formaldeído, o benzeno, o estireno, o acetaldeido, a acroleína e o furano, todos considerados agentes carcinogênicos (Pfeifer, *et al*., 2002). Da mesma forma, é possível encontrar hidrocarbonetos policíclicos aromáticos (PAH, *Polycylic Aromatic Hydrocarbons*), *N*-nitrosaminas e metais pesados, além da nicotina, que é o composto psicoativo do cigarro que gera dependência (Pfeifer, *et al*., 2002).

Os malefícios do tabaco vão desde doenças cardiovasculares e respiratórias, doenças cerebrais até quadros graves de câncer de pulmão, estomacal e de garganta (Benowitz, 2010). Seu uso também é considerado um fator de risco para sepse e o desenvolvimento de patologias como osteoporose, doenças do trato reprodutivo e diabetes (Benowitz, 2010). Contudo, os agentes carcinogênicos do tabaco não afetam somente o usuário, mas também aqueles expostos sua fumaça, como no caso de mulheres grávidas. Nos Estados Unidos, 18-25% das mulheres fumantes param de fumar durante a gravidez, sendo que 13-25% fumam até o fim do primeiro trimestre de gestação (Rogers, 2008). Em mulheres fumantes, a concentração de progesterona é diminuída, assim como a quantidade de ferro e zinco no sangue (Piasek *et al*., 2001). O tabaco também afeta os níveis de estrogênio, estradiol e gonadotrofina coriônica, o que pode acarretar em abortos espontâneos (Piasek *et al*., 2001).

Mesmo com o crescimento de propagandas e campanhas que visam cessar o consumo de tabaco em mulheres grávidas, muitas dessas ainda continuam a fumar ou são expostas ao tabaco de forma passiva, assim expondo o feto em desenvolvimento as mesmas substâncias tóxicas que a mãe (Hackshaw *et al*., 2011). Os fetos e recém-nascidos de mulheres grávidas fumantes ou que foram expostas a fumaça do tabaco apresentam graves defeitos morfológicos ao longo do desenvolvimento, especialmente no tecido ósseo e neural (Hackshaw *et al*., 2011). Da mesma forma, já foi visto que os defeitos morfológicos associados ao tabaco atuam em uma ampla gama de órgãos e estruturas corporais e não podem ser previstos de forma precisa (Hackshaw *et al*., 2011). Contudo, os defeitos no desenvolvimento dos membros e

órgãos, assim como a diminuição de peso do recém-nascido, também são quadros clínicos comuns (Jauniaux & Burton, 2007).

Sendo o tabaco composto por milhares de diferentes substâncias tóxicas, o entendimento dos mecanismos moleculares que dão origem a essas anormalidades ainda precisa ser elucidado.

1.4.2 Abuso de álcool etílico

Quando ingerido durante a gravidez, o álcool etílico pode levar a um quadro patológico amplamente estudado chamado de Síndrome Alcoólica Fetal (SAF) (O'Leary, 2004). As crianças diagnosticadas com SAF apresentam diversas anormalidades faciais (**Fig. 10**), baixo peso pós-natal, taxa de crescimento diminuída, microcefalia, hidrocefalia e deficiências intelectuais tanto de aprendizado e comportamento, quanto motoras e sociais (O'Leary, 2004). Um dos grandes problemas de tratar crianças com SAF é pelo fato que os sintomas surgem em diferentes idades e os quadros de problemas cognitivos podem perdurar por muitos anos e até se tornar mais severos (O'Leary, 2004).

O consumo de etanol durante a gravidez pode levar a teratogênese e afetar o desenvolvimento do sistema nervoso central (SNC) (O'Leary, 2004), onde o etanol já foi visto em promover morte neuronal (Genetta *et al*., 2007) e alterar a expressão gênica no desenvolvimento da crista neural em ratos (Wentzel & Eriksson, 2009). Desta forma, entender os mecanismos moleculares subjacentes ao desenvolvimento de SAF de tornam cruciais para o tratamento ou até intervenções farmacológicas dessa doença, visando atenuar seus sintomas.

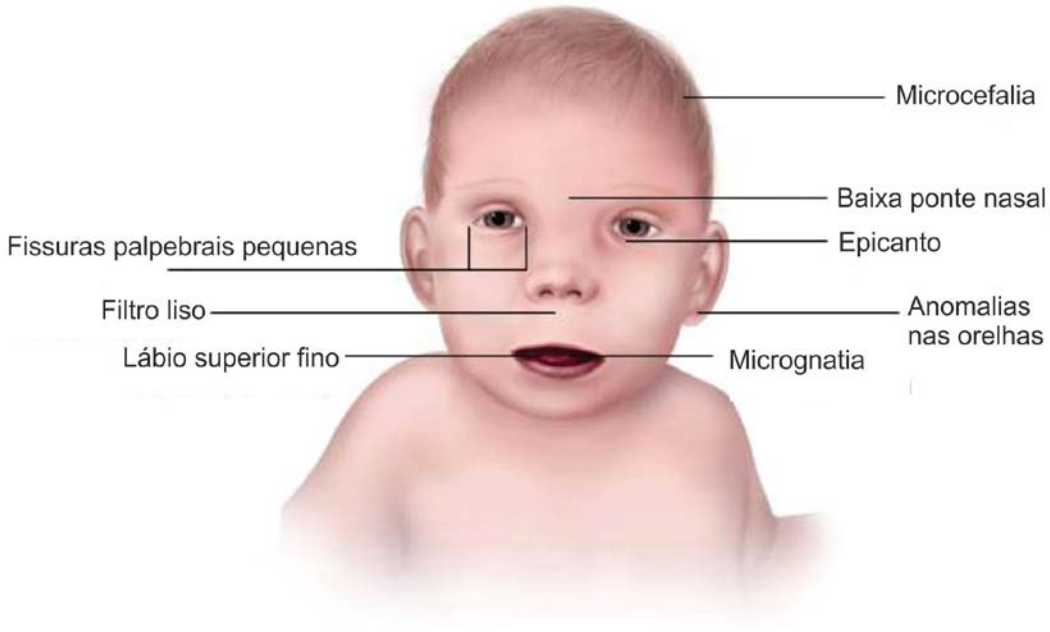


Figura 10: Anomalias faciais apresentadas por uma criança nascida com Síndrome Alcóolica Fetal (SAF). A imagem foi adaptada de http://www.aap.org/en-us/advocacy-and-policy/aap-health-initiatives/fetal-alcohol-spectrum-disorders-toolkit.

Já foi observado que o etanol causa deficiências em vitamina A (Goez *et al*., 2011), vitamina B_1 (Ke *et al*., 2009), vitamina B_9 (Hewitt *et al*., 2011) e vitamina E (Bjørneboe *et al*., 1987). Essa associação é necessária para entender a relação do abuso de álcool e SAF, pois as vitaminas possuem um papel central na neurogênese. Por exemplo, a vitamina A já foi vista em estar relacionada com diferenciação do tubo neural, desenvolvimento do hipocampo e de regiões prosencefálicas (Jiang *et al*., 2012; Rhinn & Dollé, 2012) onde a vitamina C já foi vista como necessária para o desenvolvimento do hipocampo e induzir a diferenciação de astrócitos e neurônios (Lee *et al*., 2003; Tveden-Nyborg *et al*., 2012). A vitamina E, por sua vez, já foi vista como essencial para formação do crânio, espinha e tubo neural (Kirsch *et al*., 2013; Leung *et al*., 2013; Morse, 2012), onde a vitamina B1 também já foi relacionada ao desenvolvimento do hipocampo (Ba *et al*., 1996).

Entretanto, os detalhes moleculares de como o etanol está envolvido molecularmente com o metabolismo de vitaminas e como isso leva a má formação do cérebro em pacientes com SAF é pouco estudado.

2. OBJETIVOS

2.1. Objetivo geral

Verificar, por meio de ferramentas de biologia de sistemas, como o desenvolvimento embrionário, utilizando os organismos modelo de *Homo sapiens* e *Mus musculus*, pode ser afetado através do abuso de substâncias tóxicas, tal como o etanol e os diferentes compostos presentes no tabaco.

2.2. Objetivos específicos

- Verificar as relações existentes entre as diferentes substâncias tóxicas e o desenvolvimento embrionário em dois organismos modelos, *Homo sapiens* e *Mus musculus*;

- Gerar redes de interação visando à visualização dessas redes para análises de topologia local e global;

- Aplicar programas de clusterização para avaliar os diferentes módulos que compõe os interatoma gerados e analisar as redes geradas e os módulos obtidos pelos bioprocessos associados a cada uma das redes;

- Verificar os principais nós das redes obtidas através de uma análise de centralidades utilizando os parâmetros de grau de nó, *closeness* e *betweenness*;

- Conduzir uma análise de transcritomas, referentes à exposição de organismos modelo às substâncias estudadas, previamente publicados no banco de dados do *Gene Expression Omnibus* (GEO) em busca de corroborar as hipóteses geradas. Da mesma forma, estudar as redes geradas somadas à análise transcritomica visando promover um melhor entendimento dos processos e mudanças na expressão gênica causada pelos compostos estudados em um contexto de desenvolvimento embrionário.

Capítulo I

Toxicological Effects of the Different Substances in Tobacco Smoke on Human Embryonic Development by a Systems Chemo-Biology Approach

Artigo publicado no periódico PLoS ONE

OPEN ACCESS Freely available online

PLOS ONE

Toxicological Effects of the Different Substances in Tobacco Smoke on Human Embryonic Development by a Systems Chemo-Biology Approach

Bruno César Feltes[1], Joice de Faria Poloni[2], Daniel Luis Notari[3], Diego Bonatto[1]*

1 Department of Molecular Biology and Biotechnology, Biotechnology Center of the Federal University of Rio Grande do Sul, Federal University of Rio Grande do Sul, Porto Alegre, RS – Brazil, **2** Institute of Biotechnology, University of Caxias do Sul, Caxias do Sul, RS – Brazil, **3** Computational and Information Technology Center, Universidade de Caxias do Sul, Caxias do Sul, RS – Brazil

Abstract

The physiological and molecular effects of tobacco smoke in adult humans and the development of cancer have been well described. In contrast, how tobacco smoke affects embryonic development remains poorly understood. Morphological studies of the fetuses of smoking pregnant women have shown various physical deformities induced by constant fetal exposure to tobacco components, especially nicotine. In addition, nicotine exposure decreases fetal body weight and bone/cartilage growth in addition to decreasing cranial diameter and tibia length. Unfortunately, the molecular pathways leading to these morphological anomalies are not completely understood. In this study, we applied interactome data mining tools and small compound interaction networks to elucidate possible molecular pathways associated with the effects of tobacco smoke components during embryonic development in pregnant female smokers. Our analysis showed a relationship between nicotine and 50 additional harmful substances involved in a variety of biological process that can cause abnormal proliferation, impaired cell differentiation, and increased oxidative stress. We also describe how nicotine can negatively affect retinoic acid signaling and cell differentiation through inhibition of retinoic acid receptors. In addition, nicotine causes a stress reaction and/or a pro-inflammatory response that inhibits the agonistic action of retinoic acid. Moreover, we show that the effect of cigarette smoke on the developing fetus could represent systemic and aggressive impacts in the short term, causing malformations during certain stages of development. Our work provides the first approach describing how different tobacco constituents affect a broad range of biological process in human embryonic development.

Citation: Feltes BC, Poloni JdF, Notari DL, Bonatto D (2013) Toxicological Effects of the Different Substances in Tobacco Smoke on Human Embryonic Development by a Systems Chemo-Biology Approach. PLoS ONE 8(4): e61743. doi:10.1371/journal.pone.0061743

Editor: Michael Schubert, Laboratoire de Biologie du Développement de Villefranche-sur-Mer, France

Received October 12, 2012; **Accepted** March 15, 2013; **Published** April 29, 2013

Funding: This work was supported by research grants from the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq; Grant Number 474117/2010-3), the Programa Institutos Nacionais de Ciência e Tecnologia (INCT de Processos Redox em Biomedicina-REDOXOMA; Grant Number 573530/2008-4; http://www.cnpq.br), Fundação de Amparo a Pesquisa do Rio Grande do Sul FAPERGS (PRONEM Grant Number 11/2072-2; http://www.fapergs.rs.gov.br) and CAPES (Cordenação de Aperfeiçoamento de Pessoal do Ensino Superior; http://www.capes.gov.br). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: diegobonatto@gmail.com

Introduction

There are more than 4,800 compounds present in the particulate and vapor phases of cigarette smoke [1], and many of these compounds are considered to represent a human health risk [2]. Known constituents of cigarette smoke include isoprene, butadiene, polycyclic aromatic hydrocarbons (PAHs), aldehydes, metals, *N*-nitrosamines, and aromatic amines, in addition to many others [1]. Although extensive anti-tobacco public advertisements promote smoking cessation in pregnant women, a considerable number of women still smoke during their pregnancies and/or are exposed to tobacco smoke via passive smoking [2], [3], [4].

We addressed two major issues in this work. Although prenatal smoke exposure has been previously associated with innumerable malformations during fetus growth and development and disruptions of reproductive physiology, there are gaps in the knowledge of how tobacco components (TCs) affect the developing embryo in pregnant women in a systemic way, [2], [3], [5], [6]. This knowledge gap is the first issue that we address. Interestingly, these abnormalities are not tissue specific or related to any unique pathway but, rather, are systemic and connected to a broad range of birth defects [2], [4]. The second issue that we address relates to the fact that nicotine is the principal psychoactive constituent of tobacco, understanding its biological effects on fetal and maternal health is critical, as it may affect distinct biochemical pathways when compared to other tobacco smoke constituents. Studies concerning the morphological effects of tobacco smoke constituents in fetuses from both active and passive smoking women have shown significant alterations in weight, fat mass and most anthropometric parameters as well as in the placenta with alterations in protein metabolism and enzyme activity [7]. These alterations are the results of a direct toxic effect on the fetal cells or an indirect effect through damage to, and/or functional disturbances of the placenta [7]. One possible explanation that could link nicotine and the negative regulation of development is retinoic acid (RA) signaling. RA is an indispensable molecule involved in the regulation of gene expression and cell-cell signaling during early development [8]. RA can cross the cell membrane and bind

to specific nuclear receptors, such as retinoic acid receptors (RARs) and retinoid × receptors (RXRs) [8]. Studies regarding the role of RA receptors during embryogenesis have shown that RARs are essential for the expression of HOX genes and skeletal development [8], [9]. Nicotine has been previously associated with inhibition of the RARβ gene in lung cancer, which suggests that nicotine affects RA signaling in human tissues [10]. Therefore, RA signaling is a plausible pathway through which nicotine could affect cell differentiation and cause human fetal morphological abnormalities. However, the molecular mechanisms underlying the progression or the cause of fetal abnormalities related to cigarette smoking remain unknown.

To understand these mechanisms, we performed systems chemo-biology analyses to elucidate the nature and number of proteins and modules that are associated with prenatal tobacco smoke exposure. Different protein-protein interaction (PPI) and chemical-protein interaction (CPI) networks derived from interactome projects were described. In a first analysis, we prospected and analyzed a network using a list of 95 commonly found harmful tobacco constituents [2], to elucidate how these substances could act together to influence embryonic and fetal development. In a second systems chemo-biology analysis, we prospected data on the interactome and small compounds for nicotine alone and examined how they could negatively affect cell differentiation and bone development and lead to morphological abnormalities. Furthermore, we conducted gene ontology (GO) analyses of the major biological processes derived from the PPI and CPI networks. Supporting the hypotheses gathered from systems chemo-biology analyses, a landscape network study was performed using available transcriptomic data of placenta and cord blood isolated from passive smoking women and non-smoking women [11].

A model of how selected TCs could influence embryonic development was generated. We also developed a separate model of how nicotine could affect cell differentiation and bone development. Taken together, our systems chemo-biology data are the first to show how tobacco smoke can affect fetal and embryonic development in a systemic matter at the molecular level.

Materials and Methods

Interactome Data Mining and Design of the Chemo-biology Network

To design chemo-biology interactome networks and to elucidate the interplay between development and TCs, the metasearch engines STITCH 3.1 [http://stitch.embl.de/] and STRING 9.0 [http://string-db.org/] [12], [13] were used. In this sense, a list of 51 commonly found TCs, many of them with known concentrations in the mainstream and sidestream tobacco smoke [2] were used as initial seed for network prospection in STITCH. STITCH software allows visualization of the physical connections among different proteins and chemical compounds, whereas STRING shows protein-protein interactions. Each protein-protein or protein-chemical connection (edge) shows a degree of confidence between 0 and 1.0 (with 1.0 indicating the highest confidence). The parameters used in STITCH software were as follows: all prediction methods enabled, excluding text mining; 20 to 50 interactions; degree of confidence, medium (0.400); and a network depth equal to 1. The results gathered using these search engines were analyzed with Cytoscape 2.8.2 [14]. In addition, the GeneCards [http://www.genecards.org/] [15], [16], KEGG [http://www.genome.jp/kegg/] [17], iHop [http://www.ihop-net.org/UniPub/iHOP/] [18], PubChem [http://pubchem.ncbi.nlm.nih.gov/], ALOGPS 2.1 [http://www.vcclab.org/lab/alogps/] [19], AmiGO 1.8 [http://amigo.geneontology.org/cgi-bin/amigo/go.cgi] [20], and Gene Expression Atlas [http://www.ebi.ac.uk/gxa/] [21] search engines were also employed using their default parameters.

To prospect protein-protein and chemical-protein interactions (PPI and CPI, respectively), we entered each TC into the STITCH program. TCs that were not present in the STITCH database (or those that did not shown any protein connections) and particularly well described components, such as nitric oxide, phenol and carbon monoxide, were excluded from the analysis.

Different small CPI and PPI networks were obtained (data not shown), and these networks were further analyzed using Cytoscape 2.8.2. Each network generated by STITCH and STRING was combined into a large network using the Advanced Merge Network function, which was fully implemented in Cytoscape software.

Gene Expression Data for the Main Associated Nodes of Tobacco Components

To determine whether mRNA sequences associated with specific proteins connected to each TC could be present during development, we searched the transcriptome data from the Gene Expression Atlas [22]. We used the protein name and expression data for *Homo sapiens* embryos and fetuses as the initial inputs. The expression data indicated overexpressed and underexpressed genes (Table S1 in Supporting Information S1). Gene Expression Atlas infers the expression data for a specified gene by providing a list of experimental studies [22]. We considered a gene overexpressed or underexpressed based on the number of studies that matched the expression state of our input. Proteins that are only present in embryonic tissue were colored green, whereas proteins that are only present in fetal tissue were colored pink (Table S1 in Supporting Information S1). The blue nodes indicate the presence of a protein in both embryonic and fetal tissue (Table S1 in Supporting Information S1). Uncolored nodes (default color white) connected to TCs were either not present in any of the selected tissues in the initial input or were not found in the Gene Expression Atlas database (**Fig. 1**).

Additionally, we evaluated the transcriptomic data gathered from placenta and cord blood of passive smoking women (termed group "a"), with cord blood cotinine levels >1.0 ng/mL, and from non-smoking women (group "b"), with cord blood cotinine levels <0.15 ng/mL [11]. For this purpose, the matrix file GSE30032 (available at Gene Expression Omnibus [http://www.ncbi.nlm.nih.gov/geo]) was used and a mean value of expression for each gene was generated for both groups "a" and "b". The mean value of expression was then overlaid in CPI-PPI-derived subnetworks with the software ViaComplex 1.0 [23]. By providing gene expression data and interactomic networks, the software ViaComplex generates a landscape view of gene expression in a specific network.

Solubility Predictions for Major Tobacco Component-associated CPI-PPI Networks

To predict the solubility of each TC in an aqueous environment, such as in blood and plasma, we used the program ALOGPS 2.1 [http://www.vcclab.org/lab/alogps/]. ALOGPS allows simulation of the probable solubility of a given compound determined based on its structural formula or CAS number. Compounds with a solubility of less than 35 g/L [values of ALOGpS and logS (exp)] were considered lipophilic. ALOGPS 2.1 was used with its default parameters.

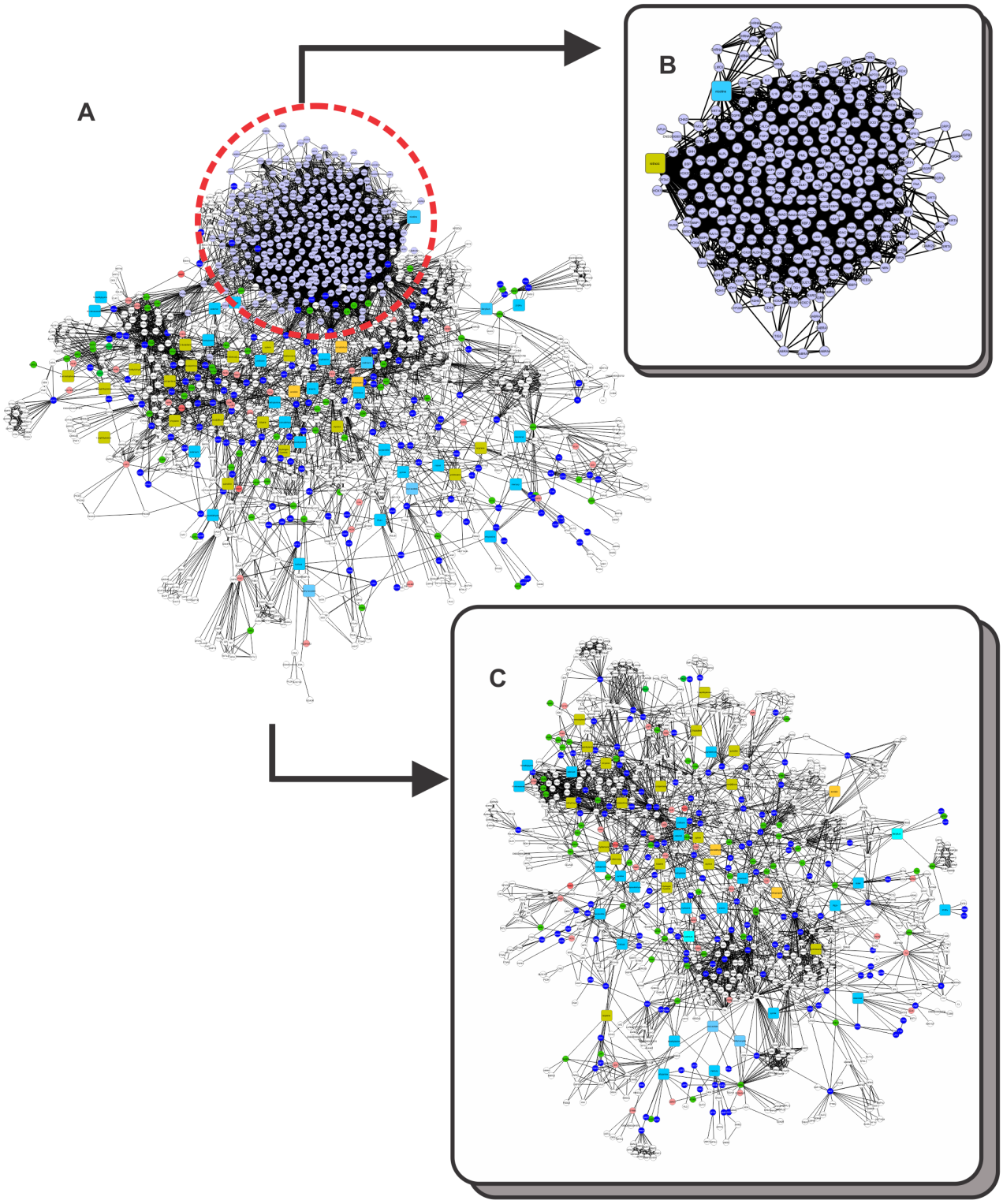


Figure 1. A binary network of chemical-protein and protein-protein interactions (CPI-PPI network) generated by the program Cytoscape 2.8.2. (A) The main network, showing 49 known substances present in tobacco, 1177 nodes (49 substances, 1128 proteins) and 7522 edges (connections). Proteins were colored to identify the tissue in which they were present: (i) pink indicates fetal tissue; (ii) green, embryonic tissue; and (iii) dark blue, both fetal and embryonic tissues. In addition, each substance was colored according to its solubility: (i) yellow indicates lipophilic and (ii) light blue, hydrophilic. We observed that nicotine resided in a module apart from the major network (A). Therefore, we separated it from the major CPI-PPI network and colored its module purple. (B) The nicotine subnetwork is shown separately from the major CPI-PPI network. It contained proteins related to retinoic acid signaling and retinoic acid (lipophilic molecule). (C) The final major CPI-PPI network after the nicotine module was extracted.

doi:10.1371/journal.pone.0061743.g001

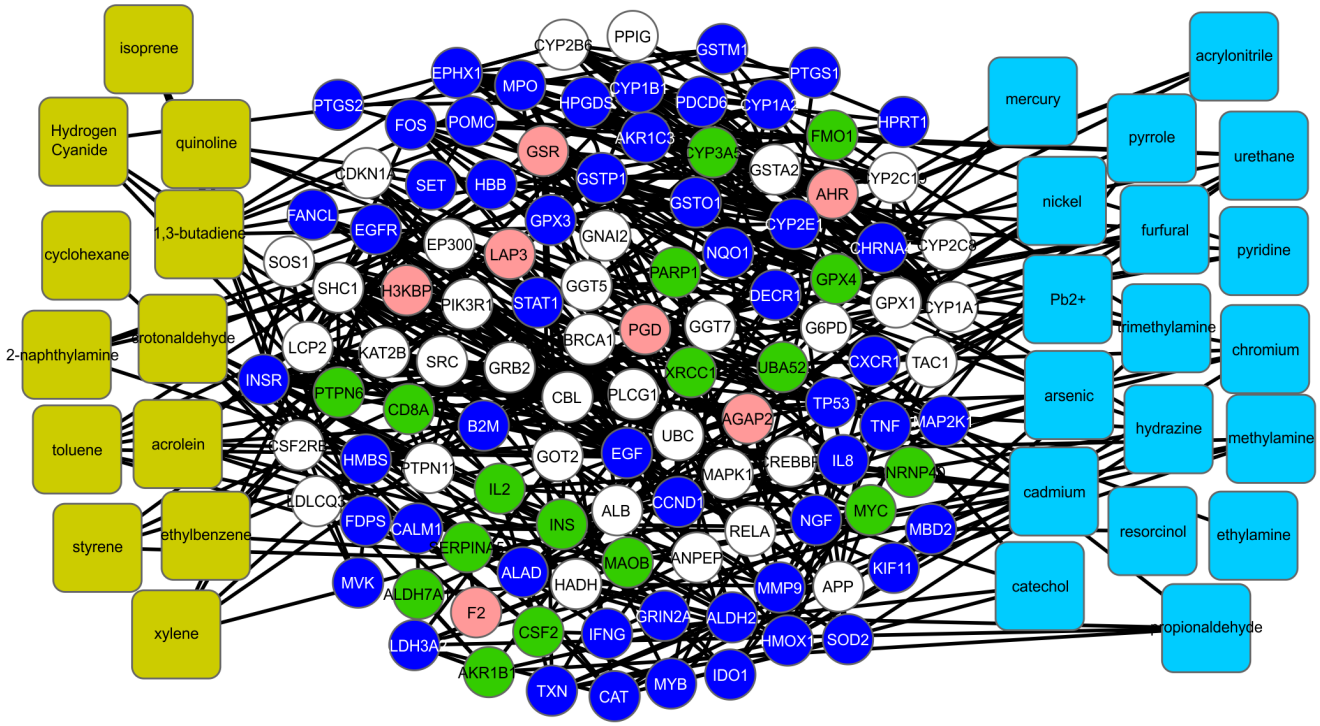


Figure 2. HBs found in the major CPI-PPI network. Betweenness and node degrees were assessed using the program CentiScaPe. Among the 143 HBs, 53 proteins are present in both tissues, reinforcing the idea of a prolonged effect of CS on embryonic development.
doi:10.1371/journal.pone.0061743.g002

Module Analysis of Major Tobacco Component-associated CPI-PPI Networks

The large CPI-PPI network obtained from the initial search (**Fig. 1**) was analyzed in terms of the major cluster or module composition using the program Molecular Complex Detection (MCODE) [24], which is available at http://baderlab.org/Software/MCODE. MCODE is based on vertex weighting by the local neighborhood density and outward traversal from a locally dense seed protein to isolate the dense regions according to given parameters stipulated by the researcher [24]. The parameters for cluster finding were as follows: loops included; degree cutoff, 2; expansion of a cluster by one neighbor shell allowed (fluff option enabled); deletion of a single connected node from clusters (haircut option enabled); node density cutoff, 0.1; node score cutoff, 0.2; kcore, 2; and maximum network depth, 100. Each cluster generates a value of "cliquishness" (C*i*), which is the degree of connection in a given group of proteins. Thus, the higher the C*i* value, the more connected the cluster [24].

Centrality Analysis of the Major Tobacco Component-associated CPI-PPI Networks

Centrality analysis was performed using the program CentiScaPe 1.2 [25]. In this analysis, the CentiScaPe algorithm evaluates each network node according to the node degree, betweenness and closeness to establish the most "central" nodes (proteins/chemicals) within the network. Thus, the most relevant node for a determined biochemical pathway or module can be obtained and further analyzed. In general terms, the closeness analysis (1) indicates the probability that any protein/chemical compound (node in our network) is relevant to another protein/chemical compound (node) in a signaling network or its associated network [25], as determined using Equation (1):

$$Clo(v) = \frac{1}{\sum_{w \in V} dist(v,w)} \quad (1)$$

where the closeness value of node v ($Clo(v)$) is determined by computing and totalizing the shortest paths among node v and all other nodes (w; $dist(v,w)$) found within a network (1). The average closeness (Clo) score was obtained by calculating the sum of different closeness scores (Clo_i) divided by the total number of nodes analyzed ($N(v)$) (Equation 2).

$$\langle Clo \rangle = \frac{\sum_i Clo_i}{N_{(v)}} \quad (2)$$

The higher the closeness value compared to the average closeness score, the higher the relevance of the protein/chemical compound to other protein nodes within the network/module. In turn, the betweenness indicates the number of the shortest paths that go through each node (Equation 3) [25], [26]:

$$Bet(v) = \sum_{s \neq v \neq t} \frac{\sigma_{sw}(v)}{\sigma_{sw}} \quad (3)$$

where σ_{stw} total number of the shortest paths from node s to node w, and σ_{sw} (v) is the number of those paths that pass through the node. The average betweenness score (Bet) of the network was calculated using equation (4), where the sum of different betweenness scores (Bet_i) is divided by the total number of nodes analyzed ($N(v)$):].

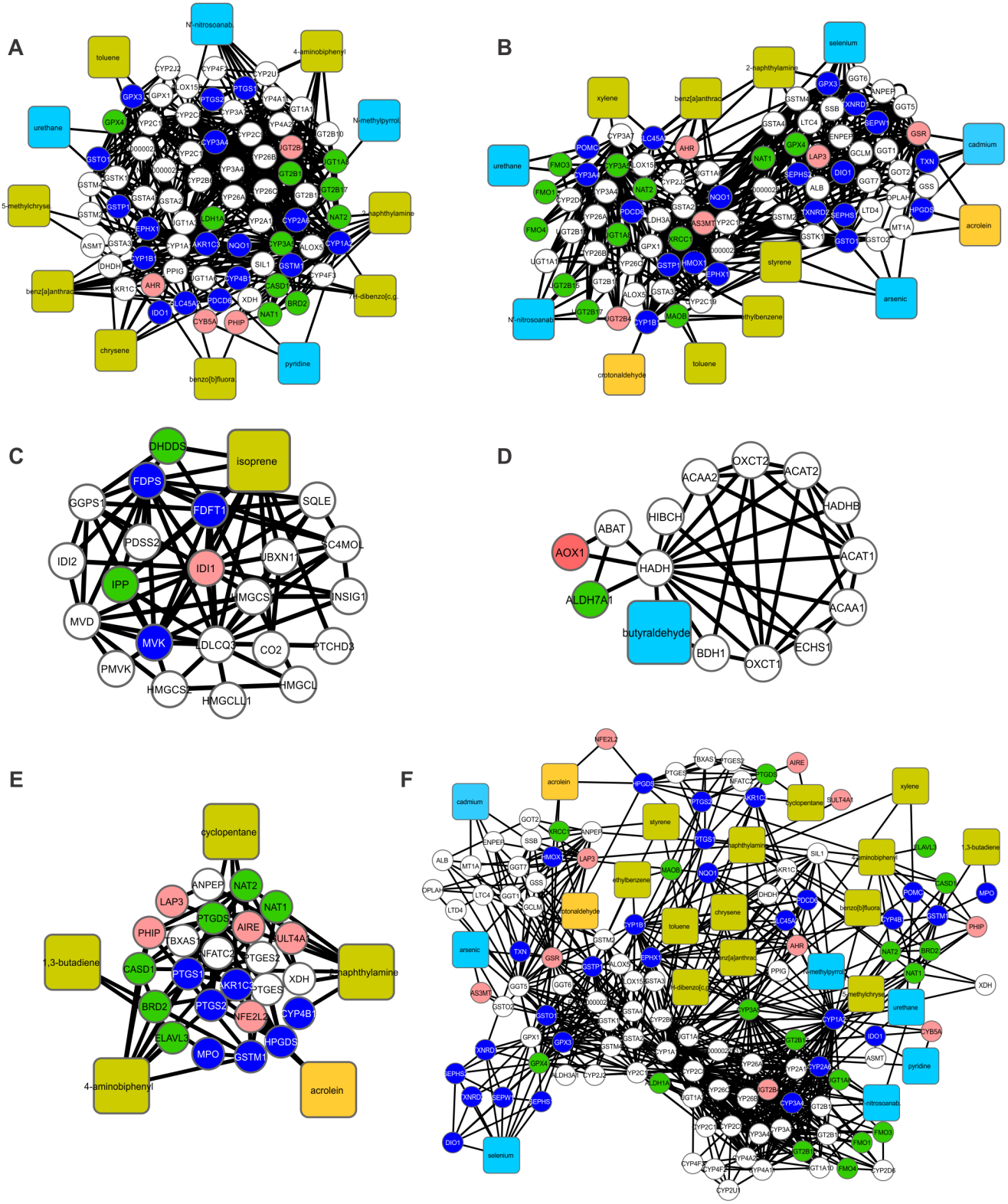


Figure 3. Cluster analysis of the major CPI-PPI network indicating clusters 1, 4, 11, 16 and 20. Cluster 1 (A) is composed of 83 nodes and 565 edges, with *Ci* = 6,843. The associated hydrophilic constituents are urethane, *N*-nitrosoanabasine, *N*-methylpyrrolidine and pyridine. The lipophilic constituents are toluene, 4-aminobiphenyl, 5-methylcrysene, benz[a]anthracene, chrysene, benzo[b]fluoranthene, 7H-dibenzo[c,g]carbazole and 2-naphthylamine. Related GO terms: oxidation reduction and unsaturated fatty acid metabolic processes. Cluster 4 (B) is composed of 90 nodes and 411 edges, with *Ci* = 4,567. The associated hydrophilic compounds are urethane, *N*-nitrosoanabasine, *N*-methylpyrrolidine, arsenic, selenium and cadmium, and the lipophilic compounds are acrolein, crotonaldehyde, toluene, xylene, ethylbenzene, benz[a]anthracene, styrene and 2-naphthylamine. Related GO term: oxidation reduction. Cluster 11 (C) is composed of 23 nodes and 74 edges, with *Ci* = 3,217. Only the lipophilic

compound isoprene is present in this cluster. Related GO term: steroid biosynthetic processes. Cluster 16 (D) is composed of 15 nodes and 36 edges, with $Ci = 2{,}400$. The associated hydrophilic compound is butyraldehyde. Related GO term: lipid modification. Cluster 20 (E) is composed of 29 nodes and 65 edges, with $Ci = 2{,}241$. The associated lipophilic compounds are acrolein, 2-naphthylamine, 1,3-butadiene, cyclopentane and 4-aminobiphenyl. Related GO terms: prostaglandin metabolic processes and unsaturated fatty acid metabolic processes. A merge of clusters 1, 4 and 20 (F). Clusters 11 and 16 did not show any proteins overlapping with any other cluster.
doi:10.1371/journal.pone.0061743.g003

$$\langle Bet \rangle = \frac{\sum_i Bet_i}{N_{(v)}} \quad (4)$$

Thus, nodes with high betweenness scores compared to the average betweenness score of the network are responsible for controlling the flow of information through the network topology. The higher a node's betweenness score, the higher the probability that the node connects different modules or biological processes, such nosed are called bottleneck nodes.

Finally, the node degree ($Deg(v)$) is a measure that indicates the number of connections (Ei) that involve a specific node (v) (Equation 5):

$$Deg(v) = \sum E_i \quad (5)$$

The average node degree of a network (Deg) is given by equation 6, where the sum of different node degree scores (Bet_i) is divided by the total number of nodes ($N(v)$) present in the network:

$$\langle Deg \rangle = \frac{\sum_i Deg_i}{N_{(v)}} \quad (6)$$

Nodes with a high node degree are called hubs [25] and have key regulatory functions in the cell.

Gene Ontology Analyses of Major Tobacco Component-associated CPI-PPI Networks

The CPI-PPI modules generated by MCODE were further studied by focusing on major biology-associated processes using the Biological Network Gene Ontology (BiNGO) 2.44 Cytoscape plugin [27], available at http://www.cytoscape.org/plugins2.php#IO_PLUGINS. The degree of functional enrichment for a given cluster and category was quantitatively assessed (p-value) using a hypergeometric distribution. Multiple test correction was also assessed by applying the false discovery rate (FDR) algorithm [28], which was fully implemented in BiNGO software at a significance level of $p<0.05$. The most statistically relevant processes were taken into account when developing the interaction model.

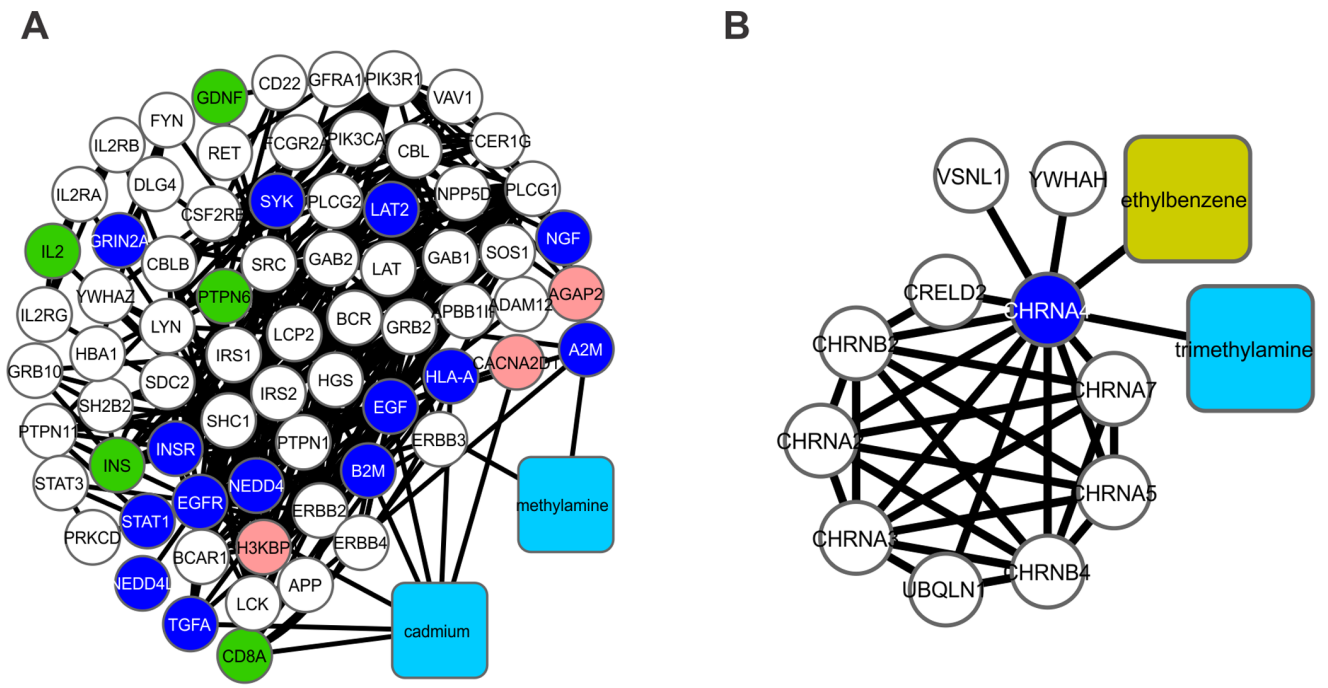


Figure 4. Cluster analysis of the major CPI-PPI network and the modules related to cell-cell signaling. Cluster 2 (A) is composed of 73 nodes and 354 edges, with $Ci = 4{,}849$. Cluster 2 contains the two hydrophilic substances cadmium and methylamine. Related GO term: regulation of cell communication. Cluster 18 (B) is composed of 13 nodes and 30 edges, with $Ci = 2, 304$. Cluster 18 contains one hydrophilic compound, trimethylamine, and one lipophilic compound, ethylbenzene. Related GO term: cell-cell signaling.
doi:10.1371/journal.pone.0061743.g004

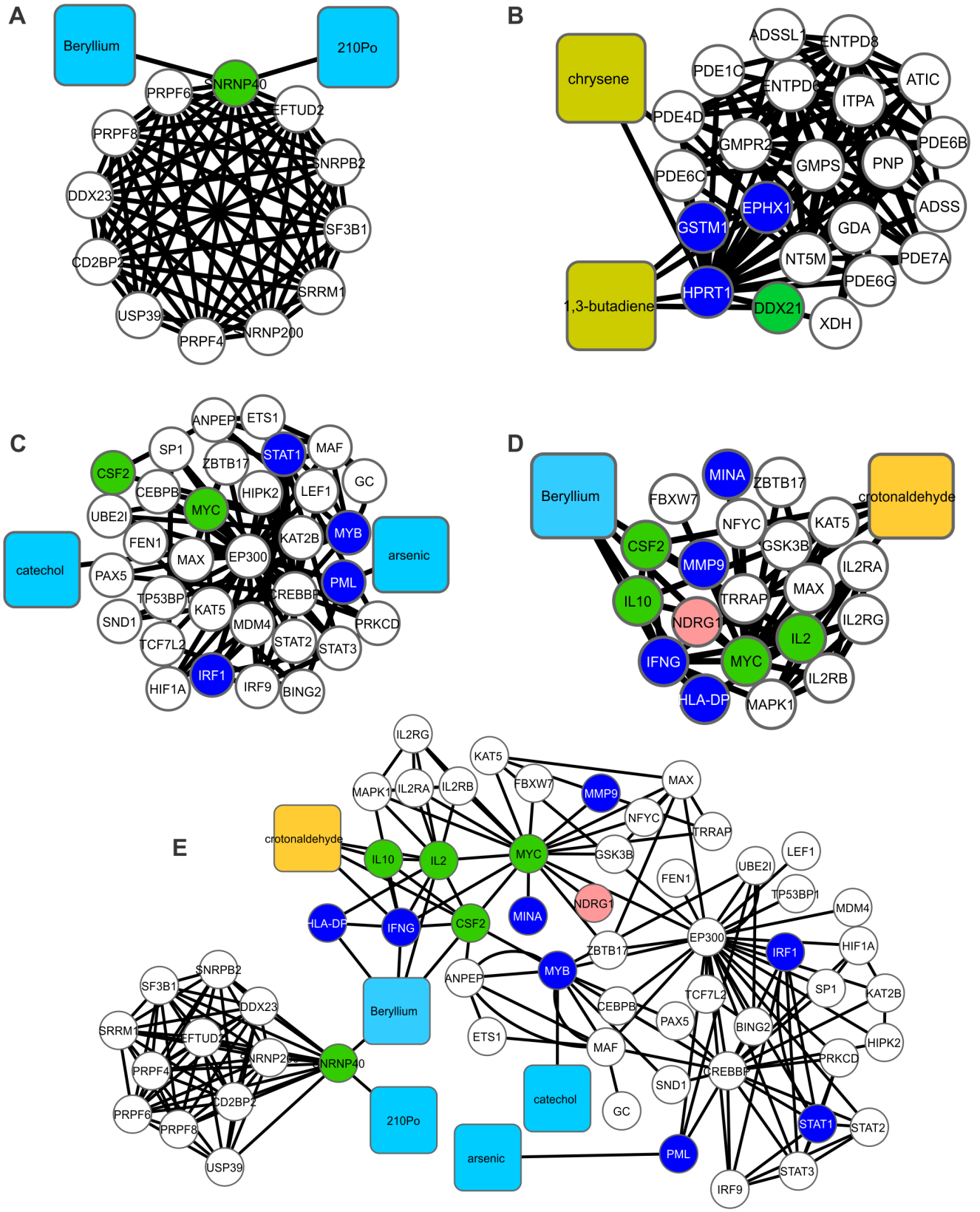


Figure 5. A merge of clusters 3, 6, 17 and 21. In (A), cluster 3 is composed of 14 nodes and 65 edges, with *Ci* = 4,643. The associated hydrophilic components are urethane, beryllium and polonium-210. Related GO terms: RNA-splicing and nucleobase, nucleoside, nucleotide and nucleic acid metabolic processes. Cluster 6 (B) is composed of 24 nodes and 102 edges, with *Ci* = 4,250. The associated lipophilic constituents are 1,3-butadiene and chrysene. Related GO term: nucleobase, nucleoside and nucleotide metabolic processes; cluster 17 (C) is composed of 35 nodes and 83 edges, with *Ci* = 2,371. The hydrophilic constituents present include catechol and arsenic. Related GO term: regulation of nucleobase, nucleoside nucleotide

and nucleic acid metabolic processes. Cluster 21 (D) is composed of 22 nodes and 49 edges, with $Ci = 2{,}227$. The associated hydrophilic constituent is beryllium, and the lipophilic constituent is crotonaldehyde. Related GO term: regulation of DNA metabolic processes. The union of clusters 3, 17 and 21 (E). Cluster 6 did not show any proteins overlapping with any other cluster.
doi:10.1371/journal.pone.0061743.g005

Results and Discussion

Data Prospecting and Topological Design of a Major CPI-PPI Network of Different Tobacco Constituents

Systems chemo-biology tools allow interactome networks of high-throughput data to be designed for CPI and PPI networks. In this sense, systems chemo-biology and systems pharmacology tools have been employed in different research areas, like prospection of new anticancer drugs [29], in order to evaluate the interaction of different small molecules with proteins and the main biological pathways potentially affected by these compounds.

Initially, our analysis was based on a list containing 95 TCs, extracted from [2]. From this initial list, we excluded compounds such as carbon monoxide, nitric oxide and phenol, which have different pleiotropic effects within a cell and could lead to the overrepresentation of many biological pathways not directly linked to development. In addition, we excluded all compounds without any protein target described, resulting in a final list containing 51 TCs commonly found in the mainstream and sidestream tobacco smoke (Table S2 in Supporting Information S1).

We have examined the relationship between 51 TCs and embryonic development pathways using systems chemo-biology tools. It should be noted that many of the thousands of substances in tobacco smoke are considered to represent public hazards, and some have carcinogenic potential [2]. Despite the growing interest in the elucidation of molecular pathways that can be affected by these compounds, many TCs do not have a known molecular target in the cell. However, our selected list of 51 TCs represents those substances with well described concentration in tobacco smoke, making them particularly attractive for experimental hypothesis testing. Moreover, these 51 TCs have some type of interaction with proteins already described, allowing systems chemo-biology studies. From this initial list of 51 TCs, we generated 51 small CPI-PPI networks (data not shown). Both STRING and STITCH add the nodes with the highest probability to be connected to a given node. Therefore, to create different CPI-PPI networks, we identified 20 to 50 additional proteins linked to each compound using only STITCH and STRING data and merged all of the networks using the Advanced Network Merge tool, which generates a single large network (referred to as the "main network", **Fig. 1A**). After creating the small networks, we found that RA receptors were present in the nicotine network. We decided to expand the nicotine network by adding a small network including RA and proteins related to RA signaling and embryonic development (**Fig. 1B**). The nicotine module was extracted from the first network to be studied independently because it showed a distinct module within the main network.

The resultant network after the nicotine module was extracted was referred to as the "major CPI-PPI network" and was composed of 898 nodes and 3,452 edges (**Fig. 1C**). It should be noted that, after merging each of the small CPI-PPI networks, two substances, 3-aminobiphenyl and dicyclohexyl, did not display any proteins in common with other compounds and were excluded from the analysis. Remarkably, the major CPI-PPI network did not show a wide overlap among the nodes, which indicates that TCs may have a broad influence and most likely affect different bioprocesses.

We next aimed to strengthen our understanding of our networks. We examined two types of data: (i) transcriptome data for each node directly associated with TCs to clarify whether the mRNA and, by inference, the proteins were present in the fetus (pink color), embryo (green color) or both (blue color) (Fig. 1, Table S1 in Supporting Information S1); and (ii) solubility predictions for the TCs and how this factor may influence the developing organism by characterizing each TC as hydrophilic or lipophilic (hydrophobic) (Fig. 1, Table S2 in Supporting Information S1). Nodes that did not show expression were left with uncolored (white) (Fig. 1, Table S1 in Supporting Information S1).

Interestingly, the majority of the nodes (145 of 234 total nodes; Fig. 1) have some role in human embryonic development, and thus, may affect the development of the organism. To predict TC solubility, we used the program ALOGPS 2.1. Among 48 TCs in our major CPI-PPI network (**Fig. 1**), we identified 21 lipophilic compounds and 27 hydrophilic components. Of the 27 hydrophilic components, 10 are inorganic, and 17 are organic (Table S2 in Supporting Information S1).

In addition, we used the program CentiScaPe 1.2 to examine the major CPI-PPI network for the most relevant proteins/compounds (**Figs. 1** and **2**). In a scale-free biological network, the most important nodes are the so called hub-bottlenecks (HBs) [30] because they combine the bottleneck function (nodes that controlling the information flow in a given network and displaying a betweenness score above the network average) and property hubs (nodes with a number of connections above the average node degree value of the network). Thus, HBs are critical nodes in a biological network [30]. In our analysis, we observed 143 HB nodes, of which 30 are TCs, and 53 were marked as present in both the fetus and embryo, 17 only in the embryo, 7 only in the fetus and 36 in neither the fetus nor embryo (white nodes) (**Fig. 2**). White nodes present in all of the networks are either not connected directly with the selected compound or do not show expression in any of the selected tissues. Because we only colored the direct nodes associated with a TC, it is clear that the TCs have a broad impact during development, acting in critical nodes that are necessary for development.

Furthermore, we sought to evaluate which TCs have the broadest effects on the major CPI-PPI network. Therefore, a closeness analysis was performed. Considering that the nodes showing the highest closeness are most relevant to the greatest number of nodes in a network [25], it can be assumed that the TCs exhibiting the highest closeness are those with the greatest systemic effects and impact the greatest number of proteins. A graph of closeness and betweenness was generated, showing that 33 TCs (from a total of 48) present closeness value above the average closeness of the network (Fig. S1 in Supporting Information S2). This finding is consistent with our interpretation that TCs have a systemic effect, impacting different proteins and physiological processes.

To understand how TCs interact with their targets, we analyzed the major CPI-PPI network for modules. From these analyses, we obtained the major TCs that affect different modules. After extraction of the nicotine subnetwork, MCODE found 22 significant modules (**Figs. 3–7**). Once the modules were obtained (**Figs. 3–7**), a gene ontology (GO) analysis was performed. Biological processes that are important for the development of organisms were listed (Table S3 in Supporting Information S1). Likewise, we performed additional GO analyses for the selected HBs (**Table 1**) and in each cluster (Tables S4–S25 in Supporting

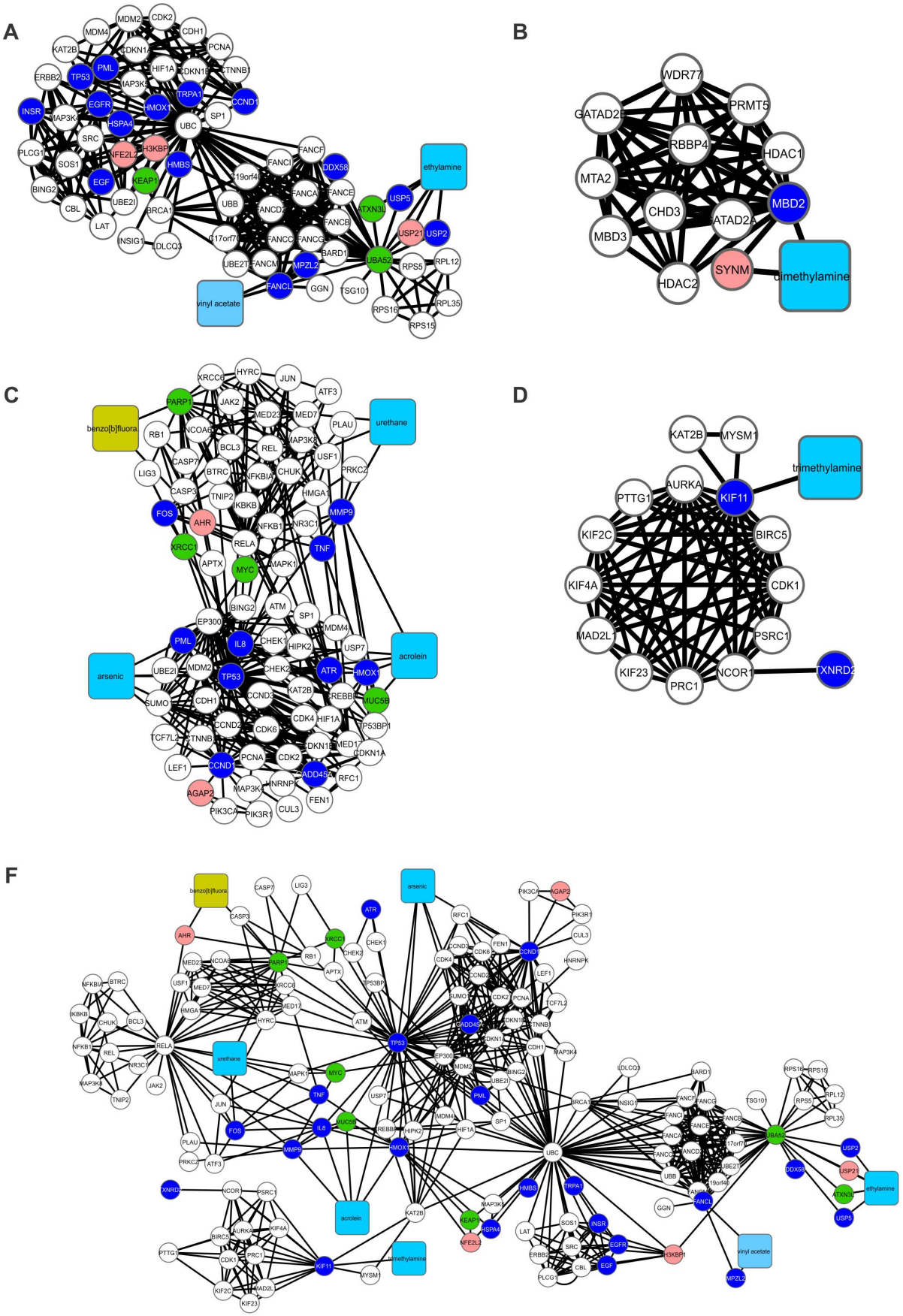

A
MDM2
CDK2
MDM4
CDH1
KAT2B
CDKN1A
PCNA
ERBB2
TP53
PML
HIF1A
MAP3K1
CDKN1B
CTNNB1
INSR
EGFR
HMOX1
TRPA1
CCND1
MAP3K4
HSPA4
SP1
UBC
PLCG1
SRC
SOS1
NFE2L2
H3KBP1
HMBS
FANCF
FANCI
DDX58
ethylamine
BING2
EGF
KEAP1
C19orf40
FANCA
FANCE
CBL
UBE2I
BRCA1
UBB
FANCD2
USP5
ATXN3L
LAT
FANCB
INSIG1
DLCQ3
C17orf70
FANCC
FANCG
USP21
USP2
UBE2T
FANCM
MPZL2
BARD1
UBA52
RPS5
RPL12
vinyl acetate
FANCL
GGN
TSG101
RPS16
RPL35
RPS15
B
WDR77
GATAD2B
PRMT5
RBBP4
HDAC1
MTA2
CHD3
MBD2
ATAD2A
MBD3
SYNM
HDAC2
dimethylamine
C
HYRC
KRCC6
JUN
ATF3
JAK2
MED23
MED7
PARP1
PLAU
urethane
benzo[b]fluora
NCOA6
RB1
MAP3K8
REL
BCL3
USF1
CASP7
LIG3
BTRC
PRKCZ
CHUK
NFKBIA
CASP3
TNIP2
HMGA1
KBKB
FOS
AHR
MMP9
NFKB1
NR3C1
XRCC1
RELA
TNF
APTX
MYC
MAPK1
ATM
BING2
SP1
EP300
MDM4
PML
IL8
CHEK1
HIPK2
USP7
acrolein
arsenic
UBE2I
MDM2
CHEK2
ATR
HMOX1
TP53
KAT2B
CCND3
TREBB2
MUC5B
SUMO
CDH1
CDK4
HIF1A
CCND2
CDK6
TP53BP1
TCF7L2
CTNNB1
CDKN1B
MED1
CDKN1A
PCNA
CDK2
LEF1
CCND1
ADD45A
RFC1
MAP3K4
HNRNPK
AGAP2
FEN1
PIK3CA
PIK3R1
CUL3
D
KAT2B
MYSM1
trimethylamine
AURKA
KIF11
PTTG1
KIF2C
BIRC5
KIF4A
CDK1
MAD2L1
PSRC1
KIF23
PRC1
NCOR1
TXNRD2
F
benzo[b]fluora
CASP7
LIG3
arsenic
PIK3CA
AGAP2
CASP3
ATR
RFC1
PIK3R1
AHR
XRCC6
CHEK1
CCND3
FEN1
CCND1
CUL3
RB1
CHEK2
CDK4
CDK6
HNRNPK
MED23
NCOA6
PARP1
APTX
CCND2
LEF1
NFKBIA
BTRC
USF1
MED7
XRCC6
TP53BP1
SUMO
CDK2
PCNA
TCF7L2
KBKB
HMGA1
MED1
HYRC
ATM
GADD45A
CDKN1B
CTNNB1
CHUK
BCL3
NFKB1
REL
RELA
TP53
CDH1
MAP3K4
urethane
EP300
NR3C1
MAPK1
MYC
MDM2
BING2
MAP3K8
JAK2
UBE2I
TNIP2
TNF
USP7
PML
IDLCQ3
BARD1
JUN
MDM4
SP1
BRCA1
INSIG1
FANCI
FANCG
TSG101
RPS16
RPS15
FOS
IL8
MUC5B
REBB2
HIPK2
HIF1A
RPS5
RPL12
PLAU
MMP9
HMOX1
UBC
FANCC
FANCD2
C17orf70
UBA52
USP2
PRKCZ
ATF3
UBB
UBE2T
C19orf40
DDX58
USP21
ethylamine
XNRD1
NCOR1
PSRC1
HMBS
TRPA1
FANCL
ATXN3L
GGN
acrolein
KAT2B
MAP3K1
KEAP1
HSPA4
NFE2L2
LAT
SOS1
INSR
H3KBP1
EGFR
SRC
ERBB2
CBL
EGF
PLCG1
vinyl acetate
MPZL2
AURKA
BIRC5
KIF4A
PTTG1
CDK1
PRC1
dimethylamine
KIF11
MYSM1
KIF2C
MAD2L1
KIF23

Figure 6. Subnetworks derived from the merge of clusters 5, 8 and 9. In (A), Cluster 5 is composed of 69 nodes and 315 edges, with $Ci = 4{,}565$. The associated hydrophilic components are vinyl acetate and ethylamine. Related GO terms: response to DNA-damage stimulus and cell cycle. cluster 7 (B) is composed of 13 nodes and 54 edges with $Ci = 4{,}154$. The associated hydrophilic compound is dimethylamine. Related GO term: chromatin organization. Cluster 8 (C) is composed of 85 nodes and 338 edges with $Ci = 3{,}976$. The associated lipophilic constituents are acrolein and benzo[b]fluoranthene, whereas the hydrophilic constituents are urethane and arsenic. Related GO terms: DNA-damage stimulus and regulation of cell cycle. Cluster 9 (D) is composed of 16 nodes and 58 edges, with $Ci = 3{,}625$. The hydrophilic constituent present is trimethylamine. Related GO term: cell cycle processes. The union of clusters 5, 8 and 9 (F). Clusters 7 did not show any proteins overlapping with any other cluster.
doi:10.1371/journal.pone.0061743.g006

Information S1). Clusters that were not associated with significant GO terms due to a lack of data or were highly speculative in our analysis were excluded (Tables S13, S15 to S18, S22 and S25 in Supporting Information S1, Fig. S2 in Supporting Information S2).

Systemic Effects of Tobacco Smoking in Human Embryogenesis: Redox and Prostaglandin Metabolic Processes

The modularity data gathered from the major PPI-CPI network (**Fig. 1C**) were subjected to GO analysis. The GO analysis of clusters 1, 4, 11, 16, and 20 (**Fig. 3A–E**) revealed five main process annotations: (i) oxidation-reduction (redox), (ii) prostaglandin metabolism, (iii) steroid biosynthesis, (iv) lipid modification, and (v) unsaturated fatty acid metabolism (Tables S4, S7, S14, S19 and S23 in Supporting Information S1). Given the overlap among the different processes, these subnetworks were merged into a single network (**Fig. 3F**). It was observed that lipophilic molecules (e.g., chrysene, toluene, benz[a]anthracene, benzo[b]fluoranthene, 7H-dibenzo(c,g)carbazole, 2-naphthylamine, 4-aminobiphenyl and 5-methylchrysene; Table S2 in Supporting Information S1) were observed to be most connected to the proteins annotated as being involved in redox processes (**Fig. 3A**). Tobacco consumption has been associated with altered redox mechanisms and the generation of oxidative stress, leading to an inflammatory response [31], [32], [33], [34]. In this sense, within the merged network (**Fig. 3F**), two prostaglandin synthases (PTGS1 and PTGS2), and two 5-lipooxygenases (ALOX5 and ALOX15B), which play a role in the synthesis of leukotrienes [35], were identified. PTGSs are not only related to inflammatory responses when they are present at high levels in tissues but are also associated with normal pregnancies due to promoting adequate circulatory adaptation and regular maternal-fetal blood flow [32], [36]. In addition to the results of our GO analyses, it is known that maternal smoke diminishes prostaglandin levels, which causes low birth weight [36]. In addition, arsenic (**Fig. 3B**), which is present in this module, is related to increased oxidative stress via redox mechanisms [37]. Considering the data amassed in this module, it is possible to speculate that pro-oxidative stimulation by TCs, such as those included in **Fig. 3,** can generate a pro-inflammatory cascade, followed by downregulation of PTGSs and increased availability leukotriene, which promotes a continuous pro-inflammatory process. To corroborate this information, we used the transcriptomic data available for placenta and cord blood of passive smoking and non-smoking women [11]. In fact, the transcriptomic data analysis of placenta and cord blood of passive smoking women using landscape evaluation of the clusters 1, 4, and 20 (Fig. S1 in Supporting Information S3) indicated that the PTGS and ALOX genes are underexpressed when compared to non-smoking women. Interestingly, almost all glutathione S-transferase genes (e.g., GSTM1, GSTA1), which catalyze the conjugation of reduced glutathione with toxic xenobiotic substrates and confer antioxidative stress protection [38], are also downregulated in the placenta and cord blood of passive smoking women (Fig. S1 in Supporting Information S3), supporting the idea that TCs induce a pro-oxidative condition in embryo.

Systemic Effects of Tobacco Smoking on Human Embryogenesis: Regulation of Cell Communication and Cell-cell Signaling

Cellular communication is of great importance for embryonic development, being essential to coordinate the different biochemical signals required to control cellular differentiation and migration. Interestingly, GO analysis of clusters 2 and 18 (**Fig. 4**) revealed two related processes: (i) regulation of cell communication and (ii) cell-cell signaling (Tables S5 and S21 in Supporting Information S1). Considering the different proteins found in cluster 2 (**Fig. 4A**), two nodes appear to be important TC targets: (i) signal transducer and activator of transcription 3 (STAT3), which is related to cell-cell signaling in stem cell cultures [39]; and (ii) colony stimulating factor receptor-β (CSF2RB), a CSF2 receptor molecule that is important for post-blastocyst embryonic development, embryo differentiation, and implantation [40]. Epidermal growth factor (EGF) and its receptor EGFR were also present in this subnetwork (**Fig. 4A**). EGFR is a plasma membrane glycoprotein that is necessary for implantation and epithelial differentiation as well as for cell signal transmission during embryogenesis [41], [42]. It should be noted that both EGF and EGFR were linked to cadmium and methylamine (**Fig. 4A**) in our systems chemo-biology data. Other growth factors, such as nerve growth factor (NGF) and transforming growth factor α (TGFA), are also present in cluster 2. It is possible that the selected constituents, cadmium and methylamine (**Fig. 4**), can play a negative role in cell-cell signaling via inhibition of growth factors and its receptors. Considering the transcriptomic data available for the placenta and cord blood of passive smoking women [11], we observed that EGFR gene and other cell-cell signaling-associated genes are downregulated when compared to non-smoking women (Figs. S2 and S3 in Supporting Information S3).

It should be noted that in the major CPI-PPI network, 1,3-butadiene is linked to HOXD13 (**Fig. 1C**), whose mutations are associated with abnormal limb length [43]. Considering that tobacco abuse can lead to limb aberrations in newborns [3], the HOXD cluster should be an interesting target with respect to understanding the effects of cigarette compounds during development. Moreover, cluster 2 (**Fig. 4A**) contains ERBB2, ERBB3 and ERBB4, which are all members of the tyrosine kinase family and show a similar structure to EGFR, which appears to be crucial for skeletal development [44], and are also downregulated in the placenta and cord blood of passive smoking women (Fig. S2 in Supporting Information S3).

Systemic Effects of Tobacco Smoking in Human Embryogenesis: Metabolism of DNA, DNA Damage Stimulus, the Cell Cycle and Chromatin Organization

In the GO analysis of clusters 3, 6, 17 and 21 (**Fig. 5**), we identified two related processes: (i) RNA-splicing and (ii) metabolism of nucleotides and DNA (Tables S6, S9, S20 and S24 in

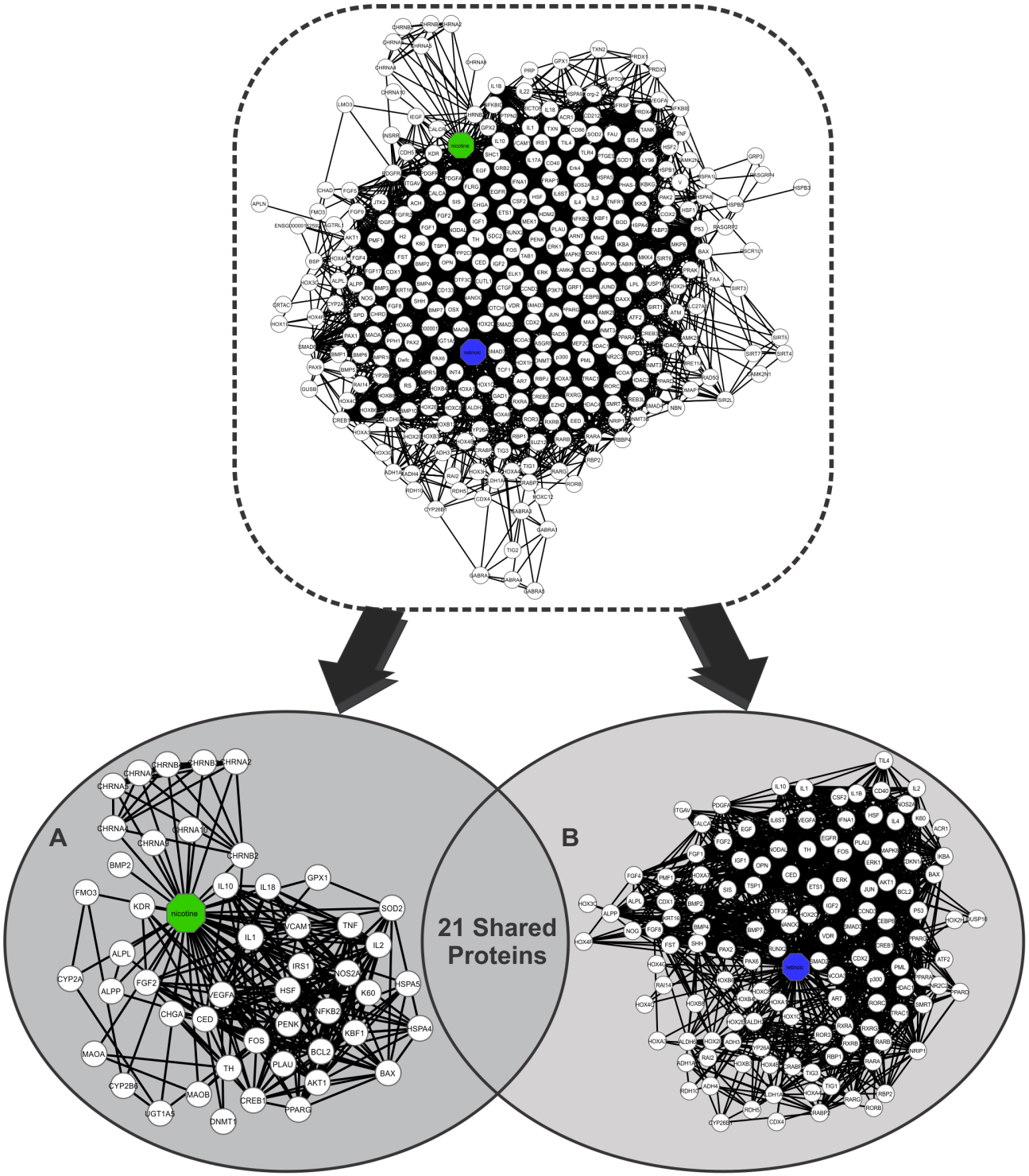


Figure 7. A binary network of the interactions between chemical compounds and proteins generated by the program Cytoscape 2.6.3, which contained 330 proteins and 4078 connections. Nicotine appears in the network as the green node, and RA appears as the blue node. White nodes are connected to both compounds are proteins. A) A subnetwork generated by the program Cytoscape containing 49 nodes and 281 edges and showing the proteins with direct connections with nicotine. B) A subnetwork generated by the program Cytoscape containing 130 nodes and 1,471 edges and showing the proteins that make direct connections with RA.
doi:10.1371/journal.pone.0061743.g007

Supporting Information S1). TCs were found associated with the metabolism of nucleotides in four different clusters, but each cluster contained different interacting compounds, including both hydrophilic (catechol) (**Fig. 5C to 5E**) and lipophilic (chrysene and 1,3-butadiene, crotonaldehyde) (**Fig. 5B**) as well as organic (chrysene and 1,3-butadiene) (**Fig. 5B**) and inorganic (beryllium,

Table 1. Major bioprocesses associated with the hub-bottleneck subnetwork.

GO-ID	GO	*p*-value	Corrected *p*-value	*k**	n#	Proteins
55114	Oxidation-reduction	4.4×10^{-16}	9.0×10^{-14}	26	645	CYP3A5;CYP1B1;PTGS2;CYP2C19;CYP2B6;PGD;PTGS1; ALDH3A2;AKR1C3;GSR;GPX1;FMO1;GPX4;HMOX1; GPX3;CAT;NQO1;HADH;CYP1A1;CYP2C8;MAOB;IDO1; CYP2E1;CYP1A2;DECR1;SOD2;LDLCQ3;ALDH7A1;G6PD; AKR1B1;TXN;ALDH2;MPO
48545	Regulation of steroid hormone stimuli	1.3×10^{-13}	1.8×10^{-11}	18	225	TNF;PTGS2;MAP2K1;RELA;PTGS1;MAOB;BRCA1; MAPK1;FOS;CCND1;CDKN1A;EP300;HMOX1;GPX4; GPX3;ALDH2;INSR;NGF
42127	Regulation of cell proliferation	1.3×10^{-12}	1.5×10^{-10}	30	848	CSF2;TNF;GNAI2;PTGS2;PTGS1;TAC1;GPX1;INS;HMOX1;IFNG;SHC1; EGF;INSR;MYC;EGFR;KAT2B;IL8;MAP2K1; RELA;TP53;IDO1;STAT1; MBD2;BRCA1;SOD2;MAPK1; CDKN1A;CCND1;IL2;NGF
42981	Regulation of apoptosis	9.3×10^{-12}	9.4×10^{-10}	29	282	CSF2;TNF;PTGS2;MMP9;GPX1;APP;INS;ALB;HMOX1; SOS1;IFNG;CAT;NQO1; EGFR;RELA;GRIN2A;TP53;IDO1; STAT1;BRCA1;SOD2;MAPK1;CDKN1A;F2;MPO;PDCD6; GSTP1;IL2;NGF
10646	Regulation of cell communication	6.0×10^{-10}	2.6×10^{-8}	31	1154	CSF2;TNF;GNAI2;PTGS2;CD8A;GRB2;TAC1;GPX1;APP; INS;SOS1;HMOX1;IFNG;CHRNA4;SHC1;CAT;EGF;INSR; AGAP2;EGFR;MAP2K1;RELA;MAOB;GRIN2A;TP53; MBD2;PTPN11;LAP3;CCND1;IL2;NGF

*Number of nodes for a given GO in the network;
#Total number of proteins for a given GO annotation.
doi:10.1371/journal.pone.0061743.t001

polonium-210 and arsenic) substances (**Fig. 5A, 5C, 5D, and 5E**). Remarkably, in cluster 6, these substances are linked to HPRT1 (**Fig. 5B**), a hypoxanthine phosphoribosyltransferase that is responsible for the metabolism of purines [45].

Moreover, 1,3-butadiene, has been found to be linked to increased genotoxic stress due to DNA damage through the formation of DNA-DNA cross-links at adenine and guanine nucleobases by its metabolites, 1,2,3,4-diepoxybutane and 3,4-epoxy-1,2-butanediol [46], [47]. The compound 1,3-butadiene has also been associated with epigenotoxic effects caused by the loss of global DNA methylation and trimethylation of histone H3 lysines 9 and 27 and H4 lysine 20, all of which are known for their roles in regulating gene expression patterns [47].

Next, in the GO analysis of clusters 5, 8 and 9 (**Fig. 6**), we identified two related processes: (i) DNA damage stimulation and (ii) the cell cycle (Tables S8, S11 and S12 in Supporting Information S1). In this cluster, arsenic binds directly to PLM (**Fig. 6B and 6D**), which is a protein with functions involved in chromatin organization, cell differentiation, DNA repair, protein sequestration and post-translational modifications [http://www.genecards.org]. PML is linked to significant proteins that regulate cell cycle such as p53, p300 and BING2 (**Fig. 6B and 6D**). Another TC, urethane, is directly connected to FOS (**Fig. 6B and 6D**), a central protein involved in proliferation, and TNF, a pro-inflammatory cytokine. Urethane is reported to alter placental morphology and down-regulates cell cycle genes as well as cytokines and other growth factors [48]. Interestingly, TCs downregulate the expression of genes associated with the metabolism of nucleotides and DNA, and cell cycle, as observed by transcriptomic analysis (Figs. S4 and S5 in Supporting Information S3).

In the GO analysis of cluster 7 (**Fig. 6B**), we only identified chromatin organization (Table S10 in Supporting Information S1) as a major biological process. Cluster 7 included dimethylamine, which is connected to MBD2 (**Fig. 6B**), a protein associated with regions of methylated DNA in CpG islands that can recruit histone deacetylases (HDACs) and DNA methyltransferases [http://www.genecards.org]. DNA methylation is also correlated with gene silencing through polycomb repression complexes (PRC) [49]. PRC is involved in the silencing of many HOX genes [49], which are critical for normal fetus development. Additionally, MBD2 is correlated with the inactivation of sexual chromosomes and is a candidate for recruiting DNA-methyltransferases (DNMTs) to the silenced promoters of long-term repressed genes [50]. Taking into account the effects of TCs in the expression of genes associated to chromatin remodeling, like HDACs, it can be observed that placenta and cord blood of passive smoking women showed a downregulation of those genes (Fig. S6 in Supporting Information S3), supporting the idea that TCs can affect chromatin remodeling during embryogenesis.

Effect of Nicotine on Retinoic Acid Signaling, Cell Proliferation and Differentiation

A second analysis using systems chemo-biology tools was developed to elucidate the relationships between nicotine, RA signaling and cell differentiation in the fetus during embryonic development in female smokers. The extracted subnetwork was examined separately due the distinct module involving nicotine and its interacting proteins. RA was added to the network because we observed that many proteins connected to nicotine are related to embryonic development and RA signaling.

Thus, the amassed data allowed the design of a major CPI network associated with nicotine and RA signaling (**Fig. 7**), which revealed several proteins that related to embryonic development, stress responses, and cell proliferation. Several of the proteins in the CPI network are directly connected to nicotine, including (i) VEGFA, a factor that induces blood vessel formation (angiogenesis) [51]; (ii) DNMT1, a DNA methyltransferase responsible for the methylation of 5′CpG islands in DNA (**Fig. 7**) [52]; (iii), FOS and JNK1 (MAPK8), which are both inducers of cell proliferation [53], [54]; and (iv) SOD2, which is responsible for mitochondrial superoxide dismutation. In addition, many proteins involved in cellular responses to stress, DNA damage and inflammation are interconnected with nicotine in the CPI network (**Fig. 7**).

We observed a connection between nicotine and JNK1 through their association with RARα in the CPI network (**Fig. 7**). JNK1 is

Table 2. The relationships between common proteins, nicotine and RA and their specific biochemical functions. These data were obtained from the GeneCards (http://www.genecards.org) and iHop (http://www.ihop-net.org/UniPub/iHOP/) databases.

Protein	Biological function	Role
VEGFA	Growth factor	Crucial role in angiogenesis, vasculogenesis and endothelial growth
TGFB1	Cytokine	Acts in differentiation, proliferation, adhesion and migration; also a potent stimulator of bone growth
ALPP	Alkaline phosphatase	Expressed in the placenta
HSF	Transcription factor	Activated under conditions of heat or other cellular stress
FGF2	Growth factor	Involved in tumor growth, development of the nervous system, cell differentiation and angiogenesis
TH	Hydroxylase	Hydroxylase that functions in the physiology of adrenergic neurons
FOS	Nuclear phosphoprotein	Nuclear phosphoprotein that participates in cell differentiation, proliferation and apoptosis
IL10	Cytokine	Involved in the immune response against pathogens and in the inflammatory response; also related to the intestinal immune system
DNMT1	Methyltransferase	DNA methylation and the establishment of methylation patterns
IL1	Cytokine	Involved in the immune response to pathogens and the inflammatory response
AKT1	Kinase	Involved in tumor formation, angiogenesis and insulin regulation
BAX	Transcription factor	Pro-apoptotic protein
ALPL	Alkaline phosphatase	Mineralization of bone matrix
BCL1 (IL5)	Cytokine	Involved in the immune response against pathogens and the inflammatory response
NOS2A	Nitric oxide synthase	Produces nitric oxide (NO)
PPARG	Proliferator peroxisome receptor	Regulator of adipocyte differentiation and glucose homeostasis
K60 (IL8)	Chemokine	Involved in the inflammatory response; angiogenesis inducer
CREB1	Transcription factor	Controls circadian rhythm, tumor suppressors and the expression of various genes involved in cell survival
PLAU	Protease	Involved in extracellular matrix degradation and possibly tumorigenesis
IL2	Cytokine	Essential in the proliferation of T-cells of the immune system. Stimulates the production of B-cells, monocytes and natural killer cells
KDR (VEGFR)	Growth factor	Plays a crucial role in vasculogenesis and angiogenesis
RARB	Retinoic Acid Receptor	Involved in cell differentiation, cell growth arrest, and signaling and transcription of target genes

doi:10.1371/journal.pone.0061743.t002

expressed when the cell undergoes cellular stresses, such as inflammation, oxidative stress, and heat [55]. In a murine model, nicotine was found to be related to the expression of JNK1 in respiratory system tissues through nicotinic receptors and receptor kinins B1 and B2, whose stimulation by bradykinin leads to increased levels of intracellular Ca^{2+} [53]. Cellular stress can activate JNK1, which phosphorylates RARα and causes its proteosomal degradation [55].

Supporting the idea that nicotine can induce the activation of pro-inflammatory cascades and different cellular stress pathways, the placenta and cord blood of passive smoking women showed an upregulation of interleukin receptors (e.g., IL2RA; IL2RB), VEGFA, FOS, JAK1, among others (Fig. S7 in Supporting Information S3). Moreover, genes associated with antioxidative stress, like SOD2, are underexpressed when compared to non-smoking women (Fig. S7 in Supporting Information S3).

The systems chemo-biology analysis performed in this study also showed that nicotine is directly connected to the protein CYP26A1 (**Fig. 7**), whose coding gene is downregulated in placenta and cord blood of passive smoking women (Fig. S7 in Supporting Information S3). This protein is responsible for regulating RA levels [56] and is expressed in a spatial-temporal manner during the development of mice, mainly in the anterior segment of the embryo and in the neural crest-derived mesenchyme [56]. However, inhibition of this protein generates an accumulation of RA and leads to deformities in the embryo, such as abnormalities in the cerebellum, urogenital tract, and spinal cord [56]. Moreover, nicotine exhibited 22 proteins in common with RA (**Table 2**). These proteins are mostly related to the immune system, stress, and cell proliferation (**Table 2**), indicating that nicotine affects RA signaling through cellular stress caused by constant tobacco use. Interestingly, we observed that nicotine was directly linked with VEGFA in our analysis (**Fig. 7**). Exposure to nicotine could results in an increase in pro-inflammatory signaling, leading to abnormal expression of VEGFA and other placental growth factors, reducing uroplacental blood flow and culminating in fetal growth restriction [57] (**Fig. 8**), an idea that is supported by transcriptomic data (Fig. S7 in Supporting Information S3).

Role of Nicotine in the Differentiation of Bone Tissue

An indirect association of nicotine with RA receptors was observed in the network via the influence of nicotine on the transcription factor JUN (**Fig. 7**). The JUN protein can be activated by the action of JNK1 during osteoblast differentiation [58]. In a smoking woman the blood concentration of nicotine are maintained at a stable level depending on the degree of tobacco use [59]. During embryogenesis, constant levels of nicotine can affect bone development, and morphological data have demon-

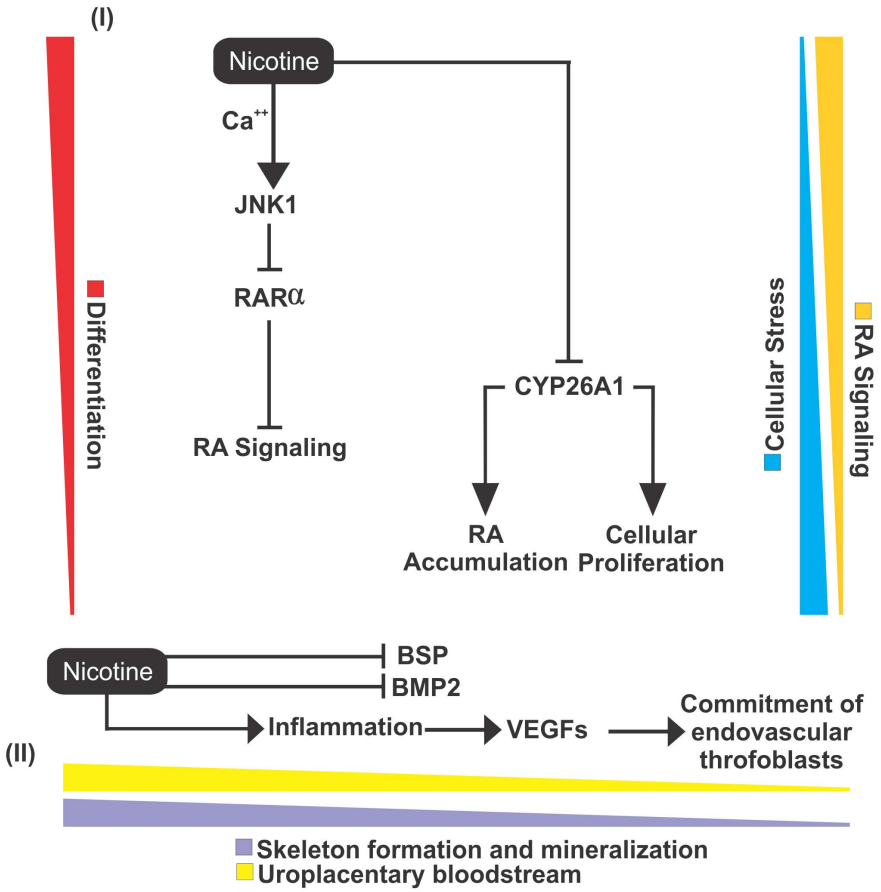


Figure 8. A molecular model illustrating how nicotine could potentially affects differentiation. In the first part of the model (I), it can be observed that by generating cellular stresses, nicotine promotes the recruitment of JNK1 through the influx of intracellular Ca^{2+}. JNK1, by itself, promotes the inhibition of RARα. Finally, nicotine promotes the inhibition of CYP26A1, which generates an accumulation of RA in the cell and an increase in cell proliferation. In the second part of the model (II), the inhibition of BMP2 and BSP is promoted by nicotine, which results in the negative regulation of bone mineralization and skeletal development. In addition, nicotine promotes a pro-inflammatory reaction that recruits VEGF and placental growth factors, which leads to an impairment of the endovascular trophoblast, resulting in a fetal growth restriction.
doi:10.1371/journal.pone.0061743.g008

strated a decrease in bone and cartilage growth [60]. An additional impact of nicotine on bone tissue differentiation involves the relationship with the BMP2 and BSP proteins. The BMP protein family includes the most potent osteogenic growth factors described to date [51] and is connected to nicotine (**Fig. 7**). A study in rabbits showed that treatment with nicotine affects BMP2 RNA levels and the activity of osteoblasts [51]. Similarly, the BSP protein is a glycoprotein that acts on bone mineralization, which has also been described as being inhibited by nicotine in rat osteoblast cells [61]. Corroborating these findings, the transcriptomic data of placenta and cord blood of passive smoking women support the fact that nicotine and other TCs inhibit the expression of BMP2 (Fig. S8 in Supporting Information S3).

Modularity and Centrality Analyses Linking Nicotine with Abnormal Embryonic Development

Once the CPI network was generated (**Fig. 7**), we aimed to understand which major protein clusters might be present. In this sense, the CPI network (**Fig. 7**) showed the presence of six modules with a coefficient of cohesion greater than or equal to 3.00 (Clusters 1–6, Fig. S3 in Supporting Information S2). It was observed that nicotine appeared in clusters 1–4 (Fig. S3A–D in Supporting Information S2), but not associated with RA (only in Fig. S3C in Supporting Information S2), which exhibits many connections other than nicotine in the network. Nicotine is connected to 49 proteins with 281 connections, and RA is connected to 130 proteins with 1,471 connections (**Fig. 7**). From the systems chemo-biology analysis, it was observed that nicotine more readily clustered in a network focused on proteins involved in development and cellular stress (**Fig. 7**). We also observed that certain clusters did not contain either nicotine or RA (Figs. S3E and F in Supporting Information S2). In cluster 5 (Fig. S3E in Supporting Information S2) there are a prevalence of proteins linked to (i) chromatin remodeling, such as EZH2, EED, SUZ12, DNMT1, DNMT3A, DNMT3B, HDAC2, HDAC4, HDAC5, and (ii) development and differentiation, including several HOX proteins [A1, 1C (A5), 4B (D4), 2I (B1), B13, B4 and 4F (A11)], PAX1, PAX6, NANOG, RAR, RXRβ, NOTCH1, CYP2B6, and CYP26A1.

To identify the major nodes within the CPI network (**Fig. 8**), we calculated betweenness, closeness and node degree centralities. From these analyses, two graphs were generated containing the proteins that showed the highest centrality values (Figs. S4 and S5 in Supporting Information S2). Interestingly, these nodes present a similar relevance order in both graphs. Thus, RA, nicotine, JNK1, p300, AKT1, p53 and ERK showed the highest betweenness, closeness and node degree values (Figs. S4 and S5 in Supporting

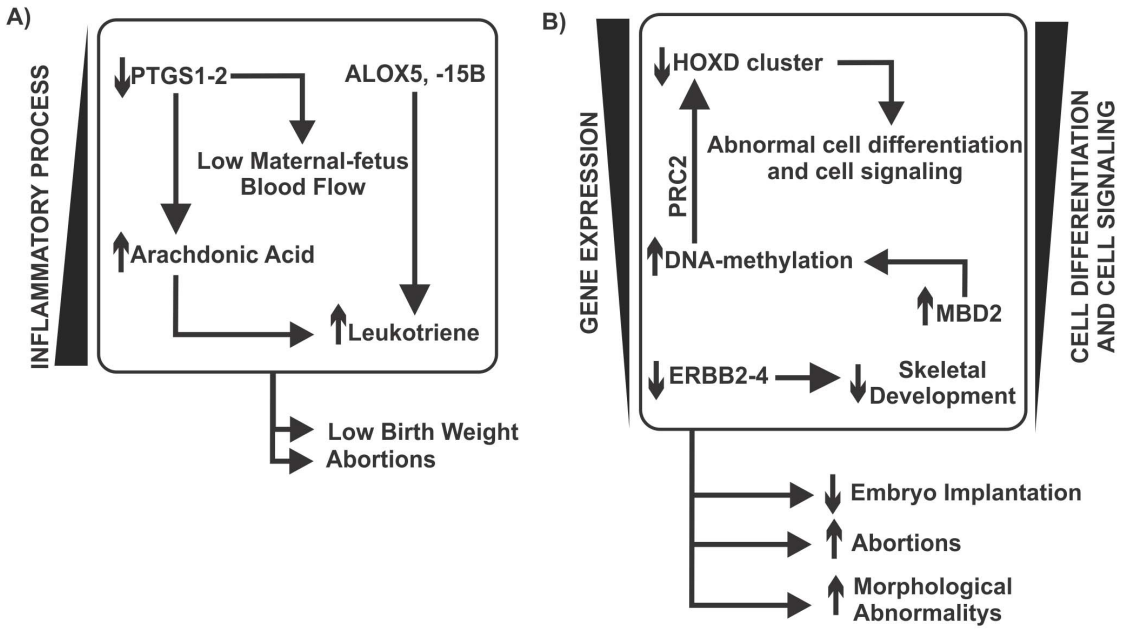


Figure 9. A model of the interactions from a systemic view showing how TCs affect development. In (A), we show that increasing TC levels generate a pro-inflammatory cascade by increasing the levels of PTGS1 and PTGS2. PTGSs are associated with inflammatory responses and are essential for normal pregnancy. Disturbances in PTGS expression could cause impairments in fetal development. TCs are connected to ALOX5 and ALOX15B, which are proteins involved in the synthesis of leukotriene, a molecule that plays pivotal roles in pro-inflammatory responses. The consequence of (A) is low birth weight in newborn infants, abortions and increased proliferation. Moreover, in (B), TCs are linked to BING2 and USP2, which are proteins related to increased activity of MDM2. This MDM2 mediated up-regulation can rapidly down-regulate p53 protein, leaving the cell more susceptible to DNA damage. TCs also down-regulate HPRT1, diminishing purine metabolism. This system exhibits a relationship with increased proliferation. The systems in (C) shows that TCs are associated with the generation of superoxides due to up-regulating NADH oxidase, which increases ROS levels and, consequently, oxidative stress. Increased oxidative stress is known to be related to birth defects. In addition, system (C) is associated with low birth weight and low neutrophil activity. Moreover, (D) shows the relationship between TCs and low hormone synthesis and signaling. Exposure to TCs could have a negative effect on androgen and estrogen solubility due to acting on the UGT cluster. TCs could also be associated with low levels of cholesterol synthesis due to increasing the levels of CYPs and diminishing the levels of FDFT1 and FDPS, which are two enzymes related to cholesterol synthesis. Low cholesterol availability would decrease general hormone synthesis. In addition, TCs affect the transport of cholesterol to the mitochondria by acting on the membrane protein StAR. Finally, in (E), the system shows the relationships between TCs and decreased global gene expression and cellular differentiation and signaling. The activities of the TCs would increase the levels of MBD2, a methylation enzyme. DNA methylation is related to gene silencing. We postulate that TCs could affect the PRC2 complex via its methylation and disturb gene expression, including that of HOX genes. TCs could also have a negative effect on gene expression by increasing YWHAH levels, which would decrease the levels of the master kinase PDPK1 and is linked to AKT activation and SMAD nuclear translocation. In addition, NOTCH signaling could be affected through the action of TCs on APP activation.
doi:10.1371/journal.pone.0061743.g009

Information S2). As these proteins play major roles in cellular physiology, it was expected that they would exhibit higher values for the three variables. The proteins with the highest values were taken into consideration in the design of a molecular model of the effect of nicotine on embryonic development (**Fig. 8**). In the centrality analysis, it was observed that p300 appeared as an important node, showing the highest values of betweenness, closeness, and the node degree (Figs. S4 and S5 in Supporting Information S2). This scenario demonstrates that there is a major influence of p300 on the network regarding the number of connections with other proteins (92 proteins), the implications of its importance for neighboring proteins (closeness) and its relationships to clusters and bioprocesses (betweenness). Therefore, the negative regulation of this protein induced by nicotine can also lead to fetal malformations and could be a potential study target for understanding the influence of nicotine in development. Noteworthy, the transcriptomic analysis of extraembryonic tissues extracted from pregnant passive smoking women showed a downregulation of p300-coding gene (Fig. S8 in Supporting Information S3).

An important issue that should be addressed in the future is the influence of the major nicotine metabolites on the activity of the enzymes and proteins observed in this work. It has been reported that 70–80% of nicotine is metabolized to cotinine by CYP2A6 to produce nicotine and a cytoplasmic aldehyde oxidase [62]. However, nicotine can generate an elevated number of different metabolites, whose mechanism of action is not clear [62]. Additionally, the mechanism of detoxification of nicotine and cotinine is based on the glucuronidation of both molecules, accounting for 40–60% of the nicotine found in urine [62]. Unfortunately, for the majority of compounds present in tobacco smoke observed in this work, the data about its metabolization or detoxification are virtually unknown. The use of metabolomic techniques associated with systems chemo-biology tools should improve our understanding of how nicotine and other TCs physiologically affect development.

Conclusions

In the present study, we showed, using systems chemo-biology tools, how the primary harmful constituents of tobacco interact with specific biological processes and affect them. Our cluster analysis results show that TCs act in many bioprocesses, including cell communication and signaling, hormone synthesis and

signaling, DNA metabolism, DNA repair, and inflammation, whose results were supported by landscape network analysis of transcriptomic data of extraembryonic tissues gathered from passive smoking women and non-smoking women. Although these processes have wide effects on cellular and embryonic physiology, they can be disturbed by the levels of the constituents of tobacco smoke. Because these effects are complex, we developed an interaction which comprises two main mechanisms associated with TCs: increased inflammatory processes **(Fig. 9A)**, and negative regulation of gene expression, cell differentiation and cell signaling (**Fig. 9B**). The systems model is related to low birth weight, an increased probability of abortion, morphological abnormalities (mainly in the skeletal system), low neutrophil activity and increased proliferation rates. Furthermore, our model can help improve knowledge and provide new insights regarding how the chemicals in tobacco cause the many morphological abnormalities observed in the newborn offspring of smoking pregnant women. The role of nicotine in embryonic development has also not been well studied. The analysis performed in this study demonstrates that nicotine has an aggressive effect on cell differentiation, affecting RA signaling in the embryo, inhibiting RA receptors due to intracellular calcium influx and stimulating cell proliferation proteins that antagonize RA activity. Osteoblast differentiation is also affected by nicotine via inhibiting proteins that stimulate bone tissue formation, which complements the TC model. Together, these data show that the birth defects observed in morphological studies could be caused by the negative action of nicotine on RA signaling. The networks also show that the pro-inflammatory pathway triggered by nicotine could be a factor leading to decreased body weight in the fetuses of smoking women. Finally, cluster analysis shows a systemic effect of nicotine, which could affect the network in a more aggressive and short-term way via cellular stress cascades.

Supporting Information

Supporting Information 1 **Table S1** Transcriptomic data of the proteins directly linked to the selected tobacco constituents (TCs). **Table S2** List of tobacco constituents (TCs) found in the major CPI-PPI network (Fig. 1). The solubility of each compound was accessed using the program ALOGPS 2.1. Those compounds with solubility less than 20 g/l were considered lipophilic. **Table S3** GO processes present in the main tobacco constituents (TCs)-associated CPI-PPI network (Fig. 1). **Table S4** GO processes present in the cluster 1 (Fig. 3A). **Table S5** GO processes present in the cluster 2 (Fig. 4A). **Table S6** GO processes present in the cluster 3 (Fig. 5A). **Table S7** GO processes present in the cluster 4 (Fig. 3B). **Table S8** GO processes present in the cluster 5 (Fig. 6A). **Table S9** GO processes present in the cluster 6 (Fig. 5B). **Table S10** GO processes present in the cluster 7 (Fig. 6B). **Table S11** GO processes present in the cluster 8 (Fig. 6C). **Table S12** GO processes present in the cluster 9 (Fig. 6D). **Table S13** GO processes present in the cluster 10 (S-Fig. 2A). **Table S14** GO processes present in the cluster 11 (Fig. 3C). **Table S15** GO processes present in the cluster 12 (S-Fig. 2B). **Table S16** GO processes present in the cluster 13 (S-Fig. 2C). **Table S17** GO processes present in the cluster 14 (S-Fig. 2D). **Table S18** GO processes present in the cluster 15 (S-Fig. 2E). **Table S19** GO processes present in the cluster 16 (Fig. 3D). **Table S20** GO processes present in the cluster 17 (Fig. 5C). **Table S21** GO processes present in the cluster 18 (Fig. 4B). **Table S22** GO processes present in the cluster 19 (S-Fig. 2F). **Table S23** GO processes present in the cluster 20 (Fig. 3E). **Table S24** GO processes present in the cluster 21 (Fig. 5D). **Table S25** GO processes present in the cluster 22 (S-Fig. 2G).
(XLSX)

Supporting Information 2 **Figure S1** Graph showing the relationship of closeness and betweenness of the TCs in the major CPI-PPI network. All nodes in the graph present a mean above average in both closeness and betweenness. The color represents the soluble property of the TCs (Light blue = hydrophilic and Yellow = lipophilic). Three nodes have distinct color/shape, since they shared a color with the adjacent node [Chromium = Large width node (black); Cadmium = Diamond shape/blue colored; and 7H-dibenzo[cg]carbazole = Orange node]. **Figure S2** Clusters excluded from the analysis due lack of literature data associated with TCs and their given GO, therefore, being highly speculative. In (A), Cluster 10 is composed by 12 nodes and 39 edges, with Ci = 3,250. The associated hydrophilic component is furfural. Related GO: Glucose Catabolic Process and Pentose-Phosphate Shunt. Cluster 12 (B) is composed by 16 nodes and 43 edges, with Ci = 2,750. The associated hydrophilic components are cadmium and acrynolitryle. Related GO: Antigen Processing and Presentation. Cluster 13 (C) is composed by 18 nodes and 48 edges, with Ci = 2,667. The associated hydrophilic component is urethane and the lipophilic is xylene. Related GO: G-Protein Coupled Receptor Protein Signaling Pathway. Cluster 14 (D) is composed by 42 nodes and 109 edges, with Ci = 2,595. The associated hydrophilic components are hydrazine, resorcinol, nickel and chromium. Related GO: Regulation of Insulin Signaling Pathway. Cluster 15 (E) is composed by 22 nodes and 55 edges, with Ci = 2,250. The associated hydrophilic components are chromium and acrynolitryle. Whereas the lipophic are xylene, chrysene, 5-methylcrysene, benz[a]anthracene and benzo[b]fluoracene. Related GO: Response to Chemical Stimuli. Cluster 19 (F) is composed by 12 nodes and 27 edges, with Ci = 2,250. The associated hydrophilic component is lead. Related GO: I-KappaB Kinase/NF-KappaB Cascade. Cluster 22 (G) is composed by 20 nodes and 43 edges, with Ci = 2,150. The associated hydrophilic components are cadmium, lead, pyrrole and arsenic. Related GO: Heme Biosynthetic Process. **Figure S3** Clusters 1 to 6, extracted from the nicotine CPI-PPI network by MCODE. The blue node is RA and the green node is nicotine. Cluster 1 (A) is composed by 159 nodes and 2373 edges, with Ci = 14, 925; Cluster 2 (B) is composed by 227 nodes and 2649 edges, with Ci = 11,670; Cluster 3 (C) is composed by 207 nodes and 1793 edges, with Ci = 8,662; Cluster 4 (D) is composed by 174 nodes and 1002 edges, with Ci = 5,759; Cluster 5 (E) is composed by 89 nodes and 300 edges, with Ci = 3,371; Cluster 6 (F) is composed by 12 nodes and 36 edges, with Ci = 3,000. Nicotine appears in four clusters (A to D), whereas RA only in C, showing that nicotine is more easily clustered. **Figure S4** Graph showing the relationship of node degree (ND) and betweenness (BT) using all proteins in the nicotine CPI-PPI network. The seven most significant proteins were selected (which are present near the value of 5.0×10^3). The dotted line shows the threshold of significance, and the values above the line are considered more relevant. **Figure S5** Graph showing the relationship of closeness (CL) and betweenness (BT) from all proteins in the nicotine CPPI-PPI network. The seven most significant proteins were selected (which are present near the value of 5.0×10^3). The dotted line shows the threshold of significance, and the values above the line are considered more relevant.
(DOCX)

Supporting Information 3 **Figure S1** Network representation of cluster 1,4, and 20 obtained from STRING metasearch engine

(A). This network was used for two-state landscape analysis of gene expression (B). Coordinates (X- and Y-axis) represent normalized values of the input network topology. Color gradient (Z-axis) represents the relative gene functional state mapped onto network according to the transcriptomic data input of GSE30032 series file [placenta plus cord blood transcriptomic data from passive smoking women (a) versus placenta plus cord blood from non-smoking women (b)]. In this sense, the mathematical equation z = a/(a+b) was used to calculated the relative gene functional state of condition (a) and condition (b). Thus, the gene expression in condition (a) is greater than condition (b) when z >0.55 (yellow to red colors), lower than (b) when z <0.45 (cyan to blue colors) and equivalent to (b) when 0.45< z <0.55 (green color). The landscape was generated by ViaComplex 1.0 software with the following options: plot as "3D-Graph", build on "node", resolution "level-50", contrast "level-50", smoothness "level-50" and zoom "level-50". **Figure S2** Network representation of cluster 2 obtained from STRING metasearch engine (A). This network was used for two-state landscape analysis of gene expression (B). Coordinates (X- and Y-axis) represent normalized values of the input network topology. Color gradient (Z-axis) represents the relative gene functional state mapped onto network according to the transcriptomic data input of GSE30032 series file [placenta plus cord blood transcriptomic data from passive smoking women (a) versus placenta plus cord blood from non-smoking women (b)]. In this sense, the mathematical equation z = a/(a+b) was used to calculated the relative gene functional state of condition (a) and condition (b). Thus, the gene expression in condition (a) is greater than condition (b) when z >0.55 (yellow to red colors), lower than (b) when z <0.45 (cyan to blue colors) and equivalent to (b) when 0.45< z <0.55 (green color). The landscape was generated by ViaComplex 1.0 software with the following options: plot as "3D-Graph", build on "node", resolution "level-50", contrast "level-50", smoothness "level-50" and zoom "level-50". **Figure S3** Network representation of cluster 18 obtained from STRING metasearch engine (A). This network was used for two-state landscape analysis of gene expression (B). Coordinates (X- and Y-axis) represent normalized values of the input network topology. Color gradient (Z-axis) represents the relative gene functional state mapped onto network according to the transcriptomic data input of GSE30032 series file [placenta plus cord blood transcriptomic data from passive smoking women (a) versus placenta plus cord blood from non-smoking women (b)]. In this sense, the mathematical equation z = a/(a+b) was used to calculated the relative gene functional state of condition (a) and condition (b). Thus, the gene expression in condition (a) is greater than condition (b) when z >0.55 (yellow to red colors), lower than (b) when z <0.45 (cyan to blue colors) and equivalent to (b) when 0.45< z <0.55 (green color). The landscape was generated by ViaComplex 1.0 software with the following options: plot as "3D-Graph", build on "node", resolution "level-50", contrast "level-50", smoothness "level-50" and zoom "level-50". **Figure S4** Network representation of cluster 3, 11 and 21 obtained from STRING metasearch engine (A). This network was used for two-state landscape analysis of gene expression (B). Coordinates (X- and Y-axis) represent normalized values of the input network topology. Color gradient (Z-axis) represents the relative gene functional state mapped onto network according to the transcriptomic data input of GSE30032 series file [placenta plus cord blood transcriptomic data from passive smoker women (a) versus placenta plus cord blood from non-smoker women (b)]. In this sense, the mathematical equation z = a/(a+b) was used to calculated the relative gene functional state of condition (a) and condition (b). Thus, the gene expression in condition (a) is greater than condition (b) when z >0.55 (yellow to red colors), lower than (b) when z <0.45 (cyan to blue colors) and equivalent to (b) when 0.45< z <0.55 (green color). The landscape was generated by ViaComplex 1.0 software with the following options: plot as "3D-Graph", build on "node", resolution "level-50", contrast "level-50", smoothness "level-50" and zoom "level-50". **Figure S5** Network representation of cluster 5, 8 and 9 obtained from STRING metasearch engine (A). This network was used for two-state landscape analysis of gene expression (B). Coordinates (X- and Y-axis) represent normalized values of the input network topology. Color gradient (Z-axis) represents the relative gene functional state mapped onto network according to the transcriptomic data input of GSE30032 series file [placenta plus cord blood transcriptomic data from passive smoker women (a) versus placenta plus cord blood from non-smoker women (b)]. In this sense, the mathematical equation z = a/(a+b) was used to calculated the relative gene functional state of condition (a) and condition (b). Thus, the gene expression in condition (a) is greater than condition (b) when z >0.55 (yellow to red colors), lower than (b) when z <0.45 (cyan to blue colors) and equivalent to (b) when 0.45< z <0.55 (green color). The landscape was generated by ViaComplex 1.0 software with the following options: plot as "3D-Graph", build on "node", resolution "level-50", contrast "level-50", smoothness "level-50" and zoom "level-50". **Figure S6** Network representation of cluster 7 obtained from STRING metasearch engine (A). This network was used for two-state landscape analysis of gene expression (B). Coordinates (X- and Y-axis) represent normalized values of the input network topology. Color gradient (Z-axis) represents the relative gene functional state mapped onto network according to the transcriptomic data input of GSE30032 series file [placenta plus cord blood transcriptomic data from passive smoker women (a) versus placenta plus cord blood from non-smoker women (b)]. In this sense, the mathematical equation z = a/(a+b) was used to calculated the relative gene functional state of condition (a) and condition (b). Thus, the gene expression in condition (a) is greater than condition (b) when z >0.55 (yellow to red colors), lower than (b) when z <0.45 (cyan to blue colors) and equivalent to (b) when 0.45< z <0.55 (green color). The landscape was generated by ViaComplex 1.0 software with the following options: plot as "3D-Graph", build on "node", resolution "level-50", contrast "level-50", smoothness "level-50" and zoom "level-50". **Figure S7** Nicotine-associated network obtained from STRING metasearch engine (A). This network was used for two-state landscape analysis of gene expression (B). Coordinates (X- and Y-axis) represent normalized values of the input network topology. Color gradient (Z-axis) represents the relative gene functional state mapped onto network according to the transcriptomic data input of GSE30032 series file [placenta plus cord blood transcriptomic data from passive smoker women (a) versus placenta plus cord blood from non-smoker women (b)]. In this sense, the mathematical equation z = a/(a+b) was used to calculated the relative gene functional state of condition (a) and condition (b). Thus, the gene expression in condition (a) is greater than condition (b) when z >0.55 (yellow to red colors), lower than (b) when z <0.45 (cyan to blue colors) and equivalent to (b) when 0.45< z <0.55 (green color). The landscape was generated by ViaComplex 1.0 software with the following options: plot as "3D-Graph", build on "node", resolution "level-50", contrast "level-50", smoothness "level-50" and zoom "level-50". **Figure S8** Retinoic acid-associated network obtained from STRING metasearch engine (A). This network was used for two-state landscape analysis of gene expression (B). Coordinates (X- and Y-axis) represent normalized values of the input network topology. Color gradient (Z-axis) represents the relative gene functional state mapped onto network according to the transcriptomic data input

of GSE30032 series file [placenta plus cord blood transcriptomic data from passive smoker women (a) versus placenta plus cord blood from non-smoker women (b)]. In this sense, the mathematical equation z = a/(a+b) was used to calculated the relative gene functional state of condition (a) and condition (b). Thus, the gene expression in condition (a) is greater than condition (b) when z >0.55 (yellow to red colors), lower than (b) when z <0.45 (cyan to blue colors) and equivalent to (b) when 0.45< z <0.55 (green color). The landscape was generated by ViaComplex 1.0 software with the following options: plot as "3D-Graph", build on "node", resolution "level-50", contrast "level-50", smoothness "level-50" and zoom "level-50".
(DOCX)

Author Contributions

Conceived and designed the experiments: BCF JFP DB. Performed the experiments: BCF JFP DB. Analyzed the data: BCF DB. Contributed reagents/materials/analysis tools: BCF JFP DB. Wrote the paper: BCF JFP DLN DB.

References

1. Pfeifer GP, Denissenko MF, Olivier M, Tretyakova N, Hecht SS, et al. (2002) Tobacco smoke carcinogens, DNA damage and p53 mutations in smoking-associated cancers. Oncogene 21(48): 7435–7451.
2. Fowles J, Bates M (2000). The chemical constituents in cigarette and cigarette smoke: priorities for harm reduction. Available: http://www.moh.govt.nz/moh.nsf/pagescm/1003/$File/chemicalconstituentscigarettespriorities.pdf. Accessed 2013 March 26.
3. Hackshaw A, Rodeck C, Boniface S (2011) Maternal smoking in pregnancy and birth defects: a systematic review based on 173 687 malformed cases and 11.7 million controls. Human Reprod Update 17(5): 589–604.
4. Florescu A, Ferrence R, Einarson T, Selby P, Soldin O, et al. (2009) Methods for quantification of exposure to cigarette smoking and environmental tobacco smoke: focus on developmental toxicology. Ther Drug Monit 31(1): 14–30.
5. Morris CV, DiNieri JA, Szutorisz H, Hurd YL (2011) Molecular mechanisms of maternal cannabis and cigarette use on human neurodevelopment. Eur J Neurosci 34(10): 1574–1583.
6. Sadeu JC, Foster WG (2011) Cigarette smoke condensate exposure delays follicular development and function in a stage-dependent manner. Fertil Steril 95(7): 2410–2417.
7. Jauniaux E, Burton GJ (2007) Morphological and biological effects of maternal exposure to tobacco smoke on the feto-placental unit. Early Hum Dev 83(11): 699–706.
8. Duester G (2008) Retinoic acid synthesis and signaling during early organogenesis. Cell 134(6): 921–931.
9. Daftary GS, Taylor HS (2006) Endocrine regulation of HOX genes. Endocr Rev 27(4): 331–355.
10. Cheng CG, Lin B, Dawson MI, Zhang XK (2002) Nicotine modulates the effects of retinoids on growth inhibition and RARβ expression in lung cancer cells. Int J Cancer 99 (2): 171–178.
11. Votavova H, Dostalova M, Krejcik Z, Fejglova K, Vasikova A, et al. (2012) Deregulation of gene expression induced by environmental tobacco smoke exposure in pregnancy. Nicotine Tob Res 14(9): 1073–1082.
12. Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, et al. (2009) STRING 8–a global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Res 37: D412–D416.
13. Snel B, Lehmann G, Bork P, Huynen MA (2000) STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. Nucleic Acid Res 28(18): 3442–3444.
14. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, et al. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13: 2498–2504.
15. Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D (1997) GeneCards: integrating information about genes, proteins and diseases. Trends Genet 13: 163.
16. Safran M, Dalah I, Alexander J, Rosen N, Iny Stein T, et al. (2010) GeneCards Version 3: the human gene integrator Database. Database (Oxford) 2010: baq020.
17. Kanehisa M, Goto S (2000) KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28: 27–30.
18. Hoffmann R, Valencia A (2004) A gene network for navigating the literature. Nature Genetics 36: 664.
19. Tetko IV, Tanchuk VY (2002) Application of associative neural networks for prediction of lipophilicity in ALOGPS 2.1 program. J Chem Inf Comput Sci 42(5): 1136–1145.
20. Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, et al. (2009) AmiGO: online access to ontology and annotation data. Bioinformatics 25(2): 288–289.
21. Kapushesky M, Emam I, Holloway E, Kurnosov P, Zorin A (2010) Gene expression atlas at the european bioinformatics institute. Nucleic Acids Res 38(Database issue): D690–D698.
22. Kapushesky M, Adamusiak T, Burdett T, Culhane A, Farne A, et al. (2012) Gene Expression Atlas update–a value-added database of microarray and sequencing-based functional genomics experiments. Nucleic Acids Res (Database issue): D1077–D1081.
23. Castro MA, Filho JL, Dalmolin RJ, Sinigaglia M, Moreira JC, et al. (2009) ViaComplex: software for landscape analysis of gene expression networks in genomic context. Bioinformatics 25(11): 1468–1469.
24. Bader GD, Hogue CW (2003) An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics 4: 2.
25. Scardoni G, Petrerlini M, Laudanna C (2009) Analyzing biological network parameters with CentiScaPe. Bioinformatics 25(21): 2857–2859.
26. Newman MEJ (2005) A measure of betweenness centrality based on random walks. Soc Networks 27: 39–54.
27. Maere S, Heymans K, Kuiper M (2005) BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. Bioinformatics 21(16): 3448–3449.
28. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B 57: 289–300.
29. Rosado JO, Henriques JP, Bonatto D (2011) A systems pharmacology analysis of major chemotherapy combination regimens used in gastric cancer treatment: predicting potential new protein targets and drugs. Curr Cancer Drug Targets 11(7): 849–869.
30. Yu H, Kim PM, Sprecher E, Trifonov V, Gerstein M (2007) The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. PLoS Comput Biol 3(4): e59.
31. van der Toorn M, Rezayat D, Kauffman HF, Bakker SJ, Gans RO, et al. (2009) Lipid-soluble components in cigarette smoke induce mitochondrial production of reactive oxygen species in lung epithelial cells. Am J Physiol Lung Cell Mol Physiol 297(1): L109–L114.
32. Menon R, Fortunato SJ, Yu J, Milne GL, Sanchez S, et al. (2011) Cigarette smoke induces oxidative stress and apoptosis in normal term fetal membranes. Placenta 32(4): 317–322.
33. Yao H, Yang SR, Kode A, Rajendrasozhan S, Caito S, et al. (2007) Redox regulation of lung inflammation: role of NADPH oxidase and NF-kappaB signaling. Biochem Soc Trans 35(Pt 5): 1151–1155.
34. Yao H, Edirisinghe I, Yang SR, Rajendrasozhan S, Kode A, et al. (2008) Genetic ablation of NADPH oxidase enhances susceptibility to cigarette smoke-induced lung inflammation and emphysema in mice. Am J Pathol 172(5): 1222–1237.
35. Haeggström JZ, Funk CD (2011) Lipoxygenase and leukotriene pathways: biochemistry, biology, and roles in disease. Chem Rev 111(10): 5866–5898.
36. Ylikorkala O, Viinikka L (1992) The role of prostaglandins in obstetrical disorders. Baillieres Clin Obstet Gynaecol 6(4): 809–827.
37. Lantz RC, Hays AM (2006) Role of oxidative stress in arsenic-induced toxicity. Drug Metab Rev 38(4): 791–804.
38. Board PG, Menon D (2012) Glutathione transferases, regulators of cellular metabolism and physiology. Biochim Biophys Acta.
39. Moledina F, Clarke G, Oskooei A, Onishi K, Günther A, et al. (2012) Predictive microfluidic control of regulatory ligand trajectories in individual pluripotent cells. Proc Natl Acad Sci U S A 109(9): 3264–3269.
40. Loureiro B, Oliveira LJ, Favoreto MG, Hansen PJ (2011) Colony-stimulating factor 2 inhibits induction of apoptosis in the bovine preimplantation embryo. Am J Reprod Immunol 65(6): 578–588.
41. Kim YJ, Lee GS, Hyun SH, Ka HH, Choi KC, et al. (2009) Uterine expression of epidermal growth factor family during the course of pregnancy in pigs. Reprod Domest Anim 44(5): 797–804.
42. Shilo BZ (2005) Regulating the dynamics of EGF receptor signaling in space and time. Development 132(18): 4017–4027.
43. Delpretti S, Zakany J, Duboule D (2012) A function for all posterior Hoxd genes during digit development? Dev Dyn 241(4): 792–802.
44. Singh AB, Harris RC (2005) Autocrine, paracrine and juxtacrine signaling by EGFR ligands. Cell Signal 17(10): 1183–1193.
45. Nyhan WL (2005) Disorders of purine and pyrimidine metabolism. Mol Genet Metab 86(1–2): 25–33.
46. Goggin M, Sangaraju D, Walker VE, Wickliffe J, Swenberg JA, et al. (2011) Persistence and repair of bifunctional DNA adducts in tissues of laboratory animals exposed to 1,3-butadiene by inhalation. Chem Res Toxicol 24(6): 809–817.
47. Koturbash I, Scherhag A, Sorrentino J, Sexton K, Bodnar W, et al. (2011) Epigenetic mechanisms of mouse interstrain variability in genotoxicity of the environmental toxicant 1,3-butadiene. Toxicol Sci 122(2): 448–456.
48. Kauffman SL (1969) Cell proliferation in embryonic mouse neural tube following urethane exposure. Dev Biol 20(2): 146–157.
49. Breiling A, Sessa L, Orlando V (2007) Biology of polycomb and trithorax group proteins. Int Rev Cytol 258: 83–136.

50. Matarazzo MR, De Bonis ML, Strazzullo M, Cerase A, Ferraro M, et al. (2007) Multiple binding of methyl-CpG and polycomb proteins in long-term gene silencing events. J Cell Physiol 210(3): 711–719.
51. Ma L, Zheng LW, Sham MH, Cheung LK (2010) Uncoupled angiogenesis and osteogenesis in nicotine-compromised bone healing. J Bone Miner Res 26(6): 1305–1313.
52. Lopatina N, Haskell JF, Andrews LG, Poole JC, Saldanha S, et al. (2002) Differential maintenance and de novo methylating activity by three DNA methyltransferases in aging and immortalized fibroblasts. J Cell Biochem 84(2): 324–334.
53. Xu Y, Zhang Y, Cardell LO (2010) Nicotine enhances murine airway contractile responses to kinin receptor agonists via activation of JNK- and PDE4-related intracellular pathways. Respir Res 11: 13.
54. León Y, Sanchez JA, Miner C, Ariza-McNaughton L, Represa JJ, et al. (1995) Developmental regulation of Fos-protein during proliferative growth of the otic vesicle and its relation to differentiation induced by retinoic acid. Dev Biol 167(1): 75–86.
55. Srinivas H, Juroske DM, Kalyankrishna S, Cody D, Price RE, et al. (2005) c-Jun N-terminal kinase contributes to aberrant retinoid signaling in lung cancer cells by phosphorylating and inducing proteasomal degradation of retinoic acid receptor alpha. Mol Cell Biol 25(3): 1054–1069.
56. Han BC, Xia HF, Sun J, Yang Y, Peng JP (2010) Retinoic acid-metabolizing enzyme cytochrome P450 26a1 (CYP26A1) is essential for implantation - functional study of its role in early pregnancy. J Cell Physiol 223(2): 471–479.
57. Feltes BC, de Faria Poloni J, Bonatto D (2011) The developmental aging and origins of health and disease hypotheses explained by different protein networks. Biogerontology 12(4): 293–308.
58. David P, Sabapathy K, Hoffmann O, Idarraga MH, Wagner EF (2002) JNK1 modulates osteoclastogenesis through both c-Jun phosphorylation-dependent and -independent mechanisms. J Cell Sci 115(22): 4317–4325.
59. Yildiz D (2004) Nicotine, its metabolism and an overview of its biological effects. Toxicon 43: 619–632.
60. Kawakita A, Sato K, Makino H, Ikegami H, Takayama S, et al. (2008) Nicotine Acts on Growth Plate Chondrocytes to Delay Skeletal Growth through the α7 Neuronal Nicotinic Acetylcholine Receptor. PLoS One 3(12): e3945.
61. Nakayama Y, Mezawa M, Araki S, Sasaki Y, Wang S, et al. (2009) Nicotine suppresses bone sialoprotein gene expression. J Periodont Res 44: 657–663.
62. Benowitz NL, Hukkanen J, Jacob P 3rd (2009) Nicotine chemistry, metabolism, kinetics and biomarkers. Handb Exp Pharmacol 192: 29–60.

Capítulo II

Evaluating the Effect of Ethanol on Vitamin Metabolism During Neurodevelopment Through a Systems Biology Analysis

Artigo submetido para a publicação no periódico Birth Defects Research Part A: Clinical and Molecular Teratology

Evaluating the effect of ethanol on vitamin metabolism during neurodevelopment through a systems biology analysis

Bruno César Feltes, Joice de Faria Poloni, Maurício Busatto, Itamar José Guimarães Nunes, and Diego Bonatto*

Centro de Biotecnologia da Universidade Federal do Rio Grande do Sul, Departamento de Biologia Molecular e Biotecnologia, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS – Brazil.

Short title: Ethanol and vitamins during neurodevelopment

* **To whom correspondence should be sent:**

Diego Bonatto

Centro de Biotecnologia da UFRGS - Sala 219

Departamento de Biologia Molecular e Biotecnologia

Universidade Federal do Rio Grande do Sul - UFRGS

Avenida Bento Gonçalves 9500 - Prédio 43421

Caixa Postal 15005

Porto Alegre – Rio Grande do Sul

BRAZIL

91509-900

Phone: (+55 51) 3308-6080

Fax: (+55 51) 3308-7309

Contract/grant sponsor: CNPq, CAPES, FAPERGS

Abstract

BACKGROUND: Fetal Alcohol Syndrome (FAS) is a prenatal disease characterized by fetal morphological abnormalities originating from exposure to alcohol. Thus, alcohol abuse during pregnancy leads to neural tissue malformations as the main outcome of FAS. Although FAS is a well-described pathology, the mechanism of FAS induction at the molecular level is virtually unknown. Moreover, alcohol abuse can affect vitamin metabolism and absorption, although how alcohol impairs such biochemical pathways remains to be elucidated.

METHODS: We employed a variety of systems chemo-biology tools to understand the interplay between ethanol metabolism and vitamins, during mouse neurodevelopment. For this purpose, we designed protein-protein and chemical-protein interaction networks and employed transcriptomic data analysis approaches to study the neural tissue of *Mus musculus* exposed to ethanol prenatally and postnatally, simulating conditions that could lead to FAS development at different life stages.

RESULTS: Our results showed that FAS can promote early changes in neurotransmitter release and glutamate equilibrium, as well as an abnormal calcium influx that can lead to impaired neurodifferentiation and neuroinflammation, which are all extensively connected with vitamin action and metabolism.

CONCLUSIONS: Ethanol is able to impair processes crucial for neural function and neurodevelopment and create an even more detrimental scenario that could lead to FAS.by altering the biosynthesis of multiple vitamins.

Key words: Ethanol, Embryonic Development, Systems Chemo-Biology, Neurodevelopment, Vitamins, Fetal Alcohol Syndrome

1. Introduction

The maternal consumption of alcohol, especially during the initial 3-6 weeks of brain development, can lead to abnormal fetal nervous system changes during pregnancy, resulting in Fetal Alcohol Syndrome (FAS) (Jaurena et al., 2011; O'Leary, 2004; Wentzel and Eriksson, 2009; Zhou et al., 2011). Although alcohol abstinence is recommended during pregnancy, more than 20% of pregnant women worldwide continue to abuse alcohol (van der Wulp et al., 2013). In these cases, a wide range of abnormal neurological outcomes can arise from FAS, including excessive neuron apoptosis (Genetta et al., 2007; Maffi et al., 2008), the risk of neuronal disorders (RNDs), and brain malformations during early embryonic development that affect neural crest and neural tube development (Wentzel and Eriksson, 2009; Zhou et al., 2011). In addition, alcohol abuse induces neuronal changes that affect both prenatal and postnatal life, including learning and cognitive impairments in young adults (O'Leary, 2004). Although FAS is an extensively studied pathology, the molecular pathways underlying its effects remain to be elucidated.

One of the many pathways affected by alcohol consumption is vitamin metabolism. Vitamin supplementation is necessary for fetal development, and specific vitamins play pivotal roles in the control of embryonic neurodevelopment (**Table 1**). For example, vitamins A and B_9 are related to neural tube closure and development (**Table 1**). In addition, reduced intake of vitamins has also been related to brain malformations or changes in neurodifferentiation patterns (**Table 1**). Moreover, alcohol consumption is already known to decrease the serum levels and absorption of the active forms of vitamin A (retinoic acid; RA), vitamin B_1 (thiamine; TM), vitamin B_9

(folic acid; FA), and vitamin E (α-tocopherol; α-TC) (Bjorneboe et al., 1987; Goez et al., 2011; Hewitt et al., 2011; Singleton and Martin, 2001). However, knowledge concerning the mechanisms through which alcohol affects the vitamin levels and metabolism at the molecular level is still scarce. Furthermore, because the effects of FAS are mainly on the nervous system and because all vitamins appear to play pivotal roles in brain formation (**Table 1**), it is crucial to understand the interplay between vitamins' biochemical pathways and alcohol during neurogenesis.

Using systems chemo-biology tools, we investigated different chemical-protein interaction (CPI) and protein-protein interaction (PPI) networks to elucidate the interplay between ethanol and different vitamins in the model organism *Mus musculus*. In addition, we compared transcriptomic data from available experimental studies that simulated maternal alcohol abuse and the effects of ethanol in the nervous system of the fetuses of *M. musculus*. Transcriptomic data originating from the adult *M. musculus* brain exposed to ethanol was also investigated to elucidate the main biological processes and changes in mRNA expression from ethanol over the short and long terms. Finally, the results gathered from systems chemo-biology analyses were used to developed interaction models for ethanol and vitamin metabolism as well as to identify the gene expression changes caused by ethanol exposure in the nervous systems at different developmental stages and adulthood.

2. Materials and Methods

2.1. Interactome data mining and the design of chemo-biology networks

To design chemo-biology interatomic networks and to elucidate the interplay among neurodevelopment, vitamins and ethanol, the metasearch engines STITCH 3.1

[http://stitch.embl.de] and STRING 9.05 [http://string-db.org] (Jensen et al., 2009; Snel et al., 2000) were used. All major active forms of vitamins (**Table 1**) commonly employed in commercial vitamin supplementation as well as ethanol were used as the initial seeds for network prospecting in STITCH. The STITCH software allows visualization of the physical connections between different proteins and chemical compounds, whereas STRING shows protein-protein interactions (Kuhn et al., 2012). Each protein-protein or protein-chemical connection (edge) possesses a degree of confidence between 0 and 1.0 (with 1.0 indicating the highest confidence). The parameters used to prospect the networks for *M. musculus* in STITCH and STRING software were as follows: all prediction methods enabled, excluding text mining; 95 to 100 interactions (for each vitamin subnetwork for the ethanol subnetwork), resulting in a 2,213-node network. In addition, a new network was developed to construct a PPI network for the microarray data for *M. musculus*, resulting in a network of 7,395 nodes (**Fig. 1**); degree of confidence, medium (0.400); and a network depth equal to 1. The results gathered using these search engines were analyzed with Cytoscape 2.8.2 (Shannon et al., 2003) and Cytoscape 3.0. In addition, the GeneCards [http://www.genecards.org/] (Rebhan et al., 1997; Safran et al., 2010), KEGG [http://www.genome.jp/kegg/] (Kanehisa and Goto, 2000) (Carbon et al., 2009), AmiGO 1.8 [http://amigo.geneontology.org/cgi-bin/amigo/go.cgi] (Carbon et al., 2009), Reactome [http://www.reactome.org] (Jupe et al., 2012), BioCyc [http://biocyc.org/] (Caspi et al., 2010) and QuickGO [http://www.ebi.ac.uk/QuickGO/] (Binns et al., 2009) search engines were also employed, using their default parameters.

Different small CPI and PPI networks were obtained (data not shown), and these networks were further analyzed using Cytoscape 2.8.2 and 3.0.

2.2. Gene expression data for the interatomic networks

We evaluated the transcriptomic data gathered from the matrix file GSE43324 (available at Gene Expression Omnibus (GEO) [http://www.ncbi.nlm.nih.gov/geo]) as follows: pregnant C57BL/6J mice treated with saline, where fetuses were euthanized at embryonic day 16 (E.16) and whole brains were removed (termed control group "b"), compared with pregnant C57BL/6J mice prenatally treated with intraperitoneal injections of ethanol (2.5 g/kg of ethanol in saline) during gestational days 14 and 16 (acute ethanol exposition; group "a"), followed by embryo euthanasia at E.16, as described by Janus and Singh (2013). A mean value of expression for each gene was generated for both groups "a" and "b". In addition, a transcriptional analysis, derived from matrix file GSE34469, was performed using pregnant *M. musculus* treated with saline, where the adult offspring were euthanized at postnatal day 70 (termed control group "b"), and compared to pregnant *M. musculus* treated with ethanol injections (2.5 g/kg of ethanol in saline) twice on gestational days 8 and 11, where the adult offspring were sacrificed at postnatal day 70 (ethanol group "a"), as described by Janus et al. (2012). The same transcriptomic study compared pregnant *M. musculus* treated with saline, where the adults were sacrificed at postnatal day 70 and the whole brains were removed (termed control group "b"), and pregnant *M. musculus* treated with ethanol injections (2.5 g/kg of ethanol in saline) twice on gestational days 14 and 16, where the adult offspring were sacrificed at postnatal day 70 (ethanol group "a"). Finally, data were gathered from another study derived from the matrix file GSE34549. Here, we compared *M. musculus* treated with 0.15 M saline alone as the control, where the adults were sacrificed at day postnatal 60 and the whole brains were

removed (termed control group “b”), with *M. musculus* treated with ethanol injections (2.5 g/kg of ethanol in 0.15 M saline) twice on days 4 and 7, the adult mouse was sacrificed at postnatal day 60 (ethanol group “a”), as described by Kleiber (2012).

Additionally to the average expression values (already calculated with log 2) of the datasets we applied the equation (1) (Castro et al., 2009) to value the relative expression, and the gathered data were overlaid in CPI-PPI-derived clusters.

$$Z = \frac{a}{(a+b)} \tag{1}$$

where *a* corresponds to the ethanol treated samples, and *b* indicates the control group.

Different Venn Diagrams were created using the online tool Data Overlapping and Area-Proportional Venn Diagram [http://apps.bioinforx.com/bxaf6/tools/app_overlap.php] to visualize the number of over- and underexpressed genes shared among the networks.

2.3. Modular analysis of the main CPI-PPI network

The MME-CPI-PPI network (**Fig. 1C**) was analyzed in terms of the major clusters or module composition using the program Molecular Complex Detection (MCODE) (Bader and Hogue, 2003). MCODE is based on vertex weighting by the local neighborhood density and outward traversal from a locally dense seed protein and isolates the dense regions according to parameters selected by the researcher (Bader and Hogue, 2003). The parameters for MCODE cluster finding were as follows: loops included; degree cutoff, 3; expansion of a cluster by one neighbor shell allowed (fluff

option enabled); deletion of a single connected node from clusters (haircut option enabled); node density cutoff, 0.1; node score cutoff, 0.2; kcore, 2; and maximum network depth, 100. Each cluster generates a value of "cliquishness" (C*i*), which is the degree of connection in a given group of proteins. Thus, the higher the C*i* value, the more connected the cluster (Bader and Hogue, 2003).

2.4. Centrality analysis of the major resulting network

Centrality analysis was performed for the "secondary network" (Fig. 1B) using the program CentiScaPe 1.2 (Scardoni et al., 2009). In this analysis, the CentiScaPe algorithm evaluates each network node according to the node degree, betweenness and closeness to establish the most "central" nodes (proteins/chemicals) within the network. Thus, the most topologically relevant node for a determined biochemical pathway or module can be obtained and further analyzed. In general terms, the closeness analysis (1) indicates the probability that any protein/chemical compound (node in our network) is relevant to another protein/chemical compound in a signaling network or its associated network (Scardoni et al., 2009), as determined using Equation (2):

$$Clo(v) = \frac{1}{\sum w \in v^{dist(v,w)}} \qquad (2)$$

where the closeness value of node *v* (*Clo(v)*) is determined by computing and totaling the shortest paths among node *v* and all other nodes (*w*; *dist(v,w)*) found within a network (1). The average closeness (*Clo*) score was obtained by calculating the

sum of different closeness scores (Clo_i) divided by the total number of nodes analyzed ($N(v)$) (Equation 3).

$$\langle Clo \rangle = \frac{\sum_i Clo_i}{N_{(v)}} \tag{3}$$

The higher the closeness value compared to the average closeness score, the higher the relevance of the protein/chemical compound to other protein nodes within the network/module. In turn, the betweenness indicates the number of the shortest paths that go through each node (Equation 4) (Newman, 2005; Scardoni et al., 2009):

$$Bet(v) = \sum_{s \neq v \neq w \in V} \frac{\sigma_{sw}(v)}{\sigma_{sw}} \tag{4}$$

where σ_{sw} total number of the shortest paths from node *s* to node *w*, and $\sigma_{sw}(v)$ is the number of those paths that pass through the node. The average betweenness score (*Bet*) of the network was calculated using equation (5), where the sum of different betweenness scores (Bet_i) is divided by the total number of nodes analyzed ($N(v)$):*)*].

$$\langle Bet \rangle = \frac{\sum_i Bet_i}{N_{(v)}} \tag{5}$$

Thus, nodes with high betweenness scores compared to the average betweenness score of the network are responsible for controlling the flow of information through the network topology. The higher a node's betweenness score,

the higher the probability that the node connects different modules or biological processes, such nodes are called bottleneck nodes.

Finally, the node degree (*Deg(v)*) is a measure that indicates the number of connections (*Ei*) that involve a specific node (*v*) (Equation 6):

$$Deg(v) = \sum E_i \tag{6}$$

The average node degree of a network (*Deg*) is given by equation 7, where the sum of different node degree scores (Bet_i) is divided by the total number of nodes (*N(v)*) present in the network:

$$\langle Deg \rangle = \frac{\sum_i Deg_i}{N_{(v)}} \tag{7}$$

Nodes with a high node degree are called hubs (Scardoni et al., 2009) and have key regulatory functions in the cell.

2.5. Gene ontology analyses of the major resulting network

The CPI-PPI modules generated by MCODE were further studied by focusing on major biology-associated processes using the Biological Network Gene Ontology (BiNGO) 2.44 Cytoscape 2.8.3 plugin (Maere et al., 2005), available at http://www.cytoscape.org/plugins2.php#IO_PLUGINS. The degree of functional enrichment for a given cluster and category was quantitatively assessed (*p*-value) using a hypergeometric distribution. BiNGO provides *p*-values assessed by functional themes that are overrepresented on a given set of genes (e.g., clusters) (Maere et al., 2005). Multiple test correction was also assessed by applying the false discovery rate (FDR)

algorithm (Benjamini and Hochberg, 1995), which was fully implemented in BiNGO software at a significance level of $p < 0.05$. The most statistically relevant processes were taken into account when developing the interaction model.

3. Results

3.1. Design of the CPI-PPI networks, topological analysis and transcriptomic data for Mus musculus

Systems chemo-biology tools allow prospecting of new drug targets and interaction between chemical compounds and biological networks (Chandra and Padiadpu, 2013; Csermely et al., 2013; Schneider and Klabunde, 2013). Our group has successfully employed systems chemo-biology to discover potential new anti-tumor drugs for gastric cancer (Rosado et al., 2011) and to understand the molecular pathways underlying fetal malformations associated with tobacco abuse during pregnancy (Feltes et al., 2013).

In this work, we first prospected small networks related to (i) the main active forms of each vitamin (**Table 1**), named the "primary network" (**Fig. 1A**), and (ii) metabolic-associated pathways for each vitamin and for ethanol in the STITCH and STRING databases for *M. musculus*, named the "secondary network" (**Fig. 1B**). Once gathered, these small networks were merged with the transcriptomic data in one large network named the *M. musculus*-ethanol network (MME-Network) (**Fig. 1C**).

The large MME-Network (**Fig. 1C**) was overlaid with four different transcriptomic datasets related to mouse offspring exposed to ethanol (**S-Table 1, see Supplementary Material 2**). For this purpose, we used the public transcriptomic data available in the GEO database regarding *M. musculus* females exposed to the same

concentration of ethanol (2.5 g ethanol/kg) during pregnancy and postnatal stage to simulate acute ethanol exposure. We also evaluated the late-life transcriptomic effects of ethanol exposure in the litters of pregnant females, which is necessary for understanding the cognitive and learning impairments observed in young adults with FAS (O'Leary, 2004). Thus, under- and over-upregulated genes were selected for transcriptome landscape analysis by overlaying these data on the following networks: (i) Prenatally-Exposed-Network (PE-Network; **Fig. 2A**), where the fetuses were exposed to ethanol during development (E.14 and E.16) and euthanized before birth (E.16), as described in the transcriptomics series GSE43324; (ii) Postnatal-Exposed MME-Network (PSE-Network; **Fig. 2B**), with pups exposed to ethanol (postnatal days 4 and 7) and euthanized at adult day 60 as indicated in GSE34549; (iii) Early Gestation-Exposed-Postnatal-Network (EGEP-Network; **Fig. 2C**), referent to transcriptomics series GSE34469 where the fetuses were exposed to ethanol during development (E.8 and E.11) and euthanized at adult day 70; and (iv) Late Gestation-Exposed-Postnatal-Network (LGEP-Network) (**Fig. 2D**), referent to the series GSE34469, in which the fetuses were exposed to ethanol during development (E.14 and E.16) and euthanized at adult day 70.

Once the networks were overlaid with transcriptomic data, we select all those genes whose expression were similar in all treatment conditions (**Fig. 3**; **S-Table 2**, **see Supplementary Material 2**), allowing us to further analyze what genes could be commonly associated with acute and chronic ethanol exposure. In this sense, 19 genes were identified that were underexpressed in the PE-, LGEP- and PSE-Networks (**Fig. 3A**; **S-Table 2**, **see Supplementary Material 2**). Interestingly, these same 19 genes were present in the EGEP-, LGEP- and PSE-Networks (**Fig. 3A**).

Next, we generated another set of diagrams for overexpressed genes (**Figs. 3B-E**). The data show only five overexpressed genes that are shared among all transcriptomic series (**Fig. 3B-E**; **S-Table 2, see Supplementary Material 2**). The relationship of these genes and their probable roles during pregnancy, neurogenesis and vitamin metabolism will be discussed further. Nevertheless, the fact they are present in different ethanol exposure experiments in individuals of different ages suggests that they are closely related with the long-term effects of ethanol in brain development and physiology. In addition, we evaluated the "secondary network" (Fig. 1B) for the most topologically relevant nodes.

In a scale-free biological network, the most topologically relevant nodes are the hub-bottlenecks (HBs) (Yu et al., 2007) because they combine the bottleneck function (nodes that connect different clusters within a network and, consequently, display a betweenness score above the network average) and the property of hubs (nodes with a number of connections above the average node degree value of the network). Thus, HBs are critical nodes in a biological network (Yu et al., 2007). In our analysis, we observed 349 HB nodes in the "secondary network" of *M. musculus* (**Fig. 1B**). Of the 349 HBs in the *M. musculus* secondary network, 174 (49.8%) were connected to the ethanol subnetwork (**Fig. 4A**).

To understand how ethanol interacts with vitamins and the different proteins studied; we evaluated each transcriptomic network for the presence of modules or clusters, which allowed us to discover major biochemical pathways related to ethanol-vitamin metabolism. We found 15 modules above our cutoff score (**S-Figs. 1-8 and Supplementary Material 1**). Once the modules were obtained, a gene ontology (GO) analysis was performed. Biological processes that are important for neurodevelopment

and neurobiological functions as well for vitamin metabolism that were present in each cluster were listed (**S-Tables 3-17**, **see Supporting Material 2**). The GOs relevant for vitamin, alcohol and neurological function in S-Tables 3 to 17 are green. In addition, processes that were related to inflammation are blue, and the GOs related to amino acid metabolism are purple. Likewise, we performed additional GO analyses for the selected over- and underexpressed genes of each transcriptomic set. The main observed GOs were inflammation, synapses and neurotransmitter release, glutamate synthesis and metabolism, calcium ion signaling and homeostasis and neurodifferentiation, and the summary of the GO information gathered in each network for over- and underexpressed genes is found in **Table 3 (Fig. 3**; for full data of the over- and underexpressed genes GO, see **S-Tables 18-25** in **Supplementary Material 2)**. Our analysis excluded GOs that were not associated with significant bioprocesses due to a lack of data or that were too general (e.g., the regulation of a biological process, regulation of transcription, or metabolism of organic substances). In addition, processes that were repeated among the GOs of over- and underexpressed genes were deleted. As expected, in the overexpressed GOs of different networks, the bioprocess of alcohol metabolism and processes associated with neuron physiology and function were highly expressed because the transcriptomic data were gathered from murine neural tissues (**S-Tables 26-32**, see **Supplementary Material 2**).

In addition, the modularity and GO combined analysis for the MME-Network reveled that all modules, with exception of clusters 2 and 9 (**S-Tables 3** and **11, S-Fig. 1B and 5A,** see **Supplementary Material 1 and 2**), were associated with neurodevelopment, alcohol metabolism and/or vitamin metabolism, indicating that those bioprocess are closely related. Because these clusters are defined by highly

dense, interconnected regions, the fact that they show close relationships with those processes may be useful for understanding how FAS affects neurodevelopment through vitamin metabolism.

3.2 Centralities analysis and overlaps among the ethanol-exposed groups

We also compared the under- and overexpressed genes in the centrality results of the PE-Network *versus* PSE-Network analyses (**Fig. 4B** and **4C**, respectively) to understand the main differences between alcohol abuse in the developing organism and in the adult individual. We also evaluated the results of the EGEP-Network *versus* the LGEP-Network analyses (**Fig. 4D** and **4E**) to observe the changes in HB status in adult individuals exposed to ethanol at different stages of embryonic development (**Fig. 4**).

The centrality analysis of the secondary network (**Fig. 1C**), the overlaps among the under- and overexpressed genes of the PE-, EGEP-, LGEP- and PSE-Networks, and the overlaps of the expression of the HB subnetworks (Fig. 4B-E) resulted in a list of 51 potential targets involved in FAS progression (**Table 2**). Our results for the under- and overexpressed genes are listed in **Table 2**.

Among the selected targets is AU-rich hydrolase (AUH) (**Table 2**), a protein that binds to AU-rich elements (ARE) in RNAs (Kurimoto et al., 2009). *AUH* mRNA lacks ARE and is upregulated in the mouse brain when mood stabilizers (e.g., lithium carbonate and valproic acid) are administered. Interestingly, these drugs upregulate the expression of ARE-containing mRNAs, such as the apoptotic inducer BCL2 (Kurimoto et al., 2009). This correlation indicates that AUH may promote neuron survival against apoptosis. Because AUH is downregulated in our networks, it becomes an important

target in understanding FAS-induced neuronal damage. Moreover, debrin 1 (Dbn1) was also found among the underexpressed genes in the prospected networks (**Table 2**). Dbn1 is related to the formation of neuronal gap junctions in the mesencephalic trigeminal nucleus, which is crucial for synapse function through receiving inputs from nerve terminals and neurotransmitters (Park et al., 2009). Synapse plasticity and protection against brain injury (Russo et al., 2012) is also related to mTor expression, which was reduced in the four networks (**Fig. 2A-D**). Changes in the mTor pathway have also been linked to neurological diseases such as Alzheimer's and Parkinson's diseases (Russo et al., 2012). mTor has been linked to synaptic plasticity, both in long term potentiation in the hippocampus and by coordinating protein synthesis (Russo et al., 2012). The co-activator-associated arginine methyltransferase (Carm1) was also among the overexpressed genes in the overlaps between the HB subnetworks of the PE- and PSE-Networks (**Fig. 4**; **Table 2**). Carm1 was found to be responsible for the inhibition of HuD, a protein that is related to synaptogenesis, learning, memory, and neurodifferentiation (Lim and Alkon, 2012). This finding indicated that some of the highly topologically relevant proteins affected by ethanol are associated with synaptic plasticity, consistent with FAS-associated learning and memory impairments.

Proteins that belong to the tubulin family, such as Tub11, or that affect tubulin mechanisms, such as Tcbc and Son, were among our potential targets. Both Son, a splicing cofactor linked to mitotic spindle assemble (Ahn et al., 2011), and Tub11, a tubulin Beta 1 class VI, were among our underexpressed genes. Both proteins appear to be present during neurogenesis in embryonic development and are related to the differentiation of different brain regions (Ahn et al., 2011; Oehlmann et al., 2004). We also identified Tbcb among the overexpressed genes in our networks (**Table 2**). Tbcb is

a tubulin-folding cofactor, primarily associated with axonogenesis, which can cause brain malformations when overexpressed (Lopez-Fanarraga et al., 2007). These results show that ethanol exposure induced the deregulation of tubulin and tubulin-associated proteins during neurodifferentiation.

The protein Gart was also among the overlaps of HB subnetworks in the underexpressed datasets for the PE- and PSE-Networks (**Fig. 4**). This protein appears to be expressed at higher levels in the prenatal cerebellum as compared to adults (Brodsky et al., 1997). More interestingly, human neuroblastoma cells treated with 6-hydroxy-dopamine, which mimic the effects of Parkinson's disease, showed Gart to be downregulated (Noelker et al., 2012). The authors also propose that downregulation of Gart could lead to neuron apoptosis. This is consistent with our data, which shows that in the PE- and PSE-Networks the process of negative regulation of apoptosis and neuron apoptosis to be overexpressed (**Table 3**). These observations indicate that Gart downregulation by ethanol during the early stages of development could lead to learning and memory impairments in adults.

The coactivator Ncoa3 (SCR3) is also present in the overlaps among the downregulated gene datasets in the HB subnetwork of the PE- and PSE-Networks (**Fig. 4**). Ncoa3 is expressed in the hippocampus and is related to retinoic acid (RA) signaling (Kashyap and Gudas, 2010). Retinoic acid (RA) is a vitamin A derivative involved in neural differentiation and neurogenesis (Chen et al., 2012a) (**Table 1**). The presence of RA mediated the release of coactivators, leading to the transcriptional activation of retinoic acid receptors (RARs) (Kashyap and Gudas, 2010). In addition, Ncoa3 is a downstream mediator of vitamin D (VD) signaling (Ahn et al., 2009). These results that ethanol is able to reduce VD and RA signaling through the downregulation of Ncoa3 in

the prenatal brain, affecting brain regions such as the hippocampus. Moreover, Shmt1is another gene underegulated in the same network overlaps that is also related to vitamin metabolism and is a serine hydroxymethyltransferase involved in folate metabolism (Beaudin et al., 2011). The authors note that *Shmt1*$^{(+/-)}$ mice showed impairment in neural tube closure and that Shmt1 expression is also coordinated by RA, indicating that this protein could have an important role in ethanol-mediated brain defects.

4. Discussion

4.1 Interpolation between ethanol and vitamin metabolism during neurodevelopment

4.1.1 Retinol signalization, folic acid metabolism, synapsis induction and circadian rhythm are affected by ethanol during embryogenesis

The GO analysis of cluster 3 (**S-Fig. 2A; S-Table 5**, see **Supplementary Material 2**) indicated the presence of proteins related to circadian rhythm and the folic acid, retinoid and neurotransmitter metabolic processes. It should be noted that RA synthesis is induced upon the loss of synaptic activity and decreased dendritic calcium levels (Chen et al., 2012a). In the prenatal ethanol exposure network (PE-network; **Fig. 2A**), we found that retinoid metabolic processes and Ca^{2+} ion homeostasis are underexpressed (**Table 3**). This is interesting because both calcium ion homeostasis and RA metabolism genes were underexpressed, showing that the induction of RA to overcome synaptic loss might not be possible in ethanol-exposed fetuses. Moreover,

RARs have also been found to be essential for improved learning and memory in the adult brain and have even alleviated memory deficits in a transgenic mice model for Alzheimer's disease (Nomoto et al., 2012). Cognitive and learning abilities have already been found to be affected in young adults displaying FAS (O'Leary, 2004), and the impairment of RA metabolism could be an explanation that has not yet been examined. Moreover, RARα is abundantly found in the cortex and hippocampus (Nomoto et al., 2012). Our data indicated that RARα is downregulated in the EGEP-Network (Chen et al., 2012a) (**S-Table 1**, see **Supplementary Material 2**), showing that the changes in synaptic plasticity and learning behaviors that depend on RA occur specifically during early development.

To corroborate the idea that ethanol affects RA, we observed that ALDH1A2 (RALDH2) gene, which codes for an aldehyde dehydrogenase and is responsible for the synthesis of RA from retinal (Strate et al., 2009), was found to be underexpressed in both the PE-Network and the PSE-Network (**Fig. 4**). This is interesting because ALDH1B1, another aldehyde dehydrogenase, is downregulated in ethanol-exposed embryos during nerulation (Zhou et al., 2011), a finding corroborated in our systems chemo-biology analysis, as ALDH1B1 was downregulated in both the EGEP-Network and the LGEP-Network (**Fig. 4**).

Folic acid (FA) metabolism was also associated with Cluster 3 (**Fig. 2**A; **S-Table 5**, see **Supplementary Material 2**). It is already known that ethanol affects folic acid absorption in guinea pigs (Hewitt et al., 2011) and that FA has multiple roles in neural tissue (**Table 1**). Remarkably, FA is able to differentiate neurospheres into multiple neural cell types and promote synaptic connections in Pax3-deficient mice (Ichi et al., 2012).

In the transcriptomic datasets analyzed, only the dihydrofolate reductase (DHFR) gene, which codes for a key enzyme in folate metabolism (**S-Table 5**, see **Supplementary Material 2**), was found to be underexpressed in fetuses exposed to ethanol in the final phase of development (LGEP-Network; **Fig. 2D**). One study shows that FA deficiency decreases neural progenitor cell proliferation in the mouse forebrain during late gestation (Craciunescu et al., 2004). Therefore, ethanol may also affect neuronal proliferation through FA metabolism, mainly through DHFR deregulation.

Another important process found within cluster 3 is circadian rhythm (**S-Table 5**, see **Supplementary Material 2**). Ethanol consumption and abuse affect are known to affect sleep cycles and melatonin secretion (Brager et al., 2010; Roehrs and Roth, 2001). Interestingly, our data analyses (PE-Network; **Fig. 2A**) indicate that calcium ion homeostasis is downregulated (**Table 3**). Circadian rhythm is controlled by melatonin secretion, which consequently has a central role in inducing neurodevelopment during embryogenesis by inducing calcium ion signaling (de Faria Poloni et al., 2011). In pregnant women, the abuse of ethanol could skew proper melatonin secretion and calcium ion signaling and subsequently change sleep induction and neurodevelopment. We found RA, thiamine (TM), α-tocopherol (α-TC) and phytonadione (phylloquinone; PQN) in cluster 3.

4.1.2 Ethanol negatively affects vitamin D metabolism and leads to its degradation

Another interesting cluster (cluster 5; **S-Fig. 3A,** see **Supplementary Material 1**) presented several GOs related to neurodevelopment and RA and VD metabolism (**S-**

Table 7, see **Supplementary Material 2**). Among all of the genes/proteins belonging to this cluster, CYP2R1 (**Fig. 2A** and **2C**) was found to be underexpressed in both murine fetuses exposed to ethanol (PE-Network; **Fig. 2A**) and also in murine adults that were exposed to ethanol during development (EGEP-Network; **Fig. 2C**).

CYP2R1 is a VD hydroxylase that converts vitamin D_3 into the first active ligand (25-hydroxy vitamin D_3 – $25OHD_3$) for the vitamin D receptor (VDR) (Eyles et al., 2013). VDR forms heterodimers with retinoid X receptors (RXR) to initiate transcription during the differentiation of different tissues (Eyles et al., 2013). It has been reported that VD deficiency is correlated with decreased intracellular calcium levels in rat cortex (Baksi and Hughes, 1982), which infers that ethanol affects calcium ion homeostasis in the PE-Network and is critical for neurodevelopment. VDR expression is also observed in differentiating fields in rodent brains and in proliferating cells in the lateral ventricle (Eyles et al., 2013).

Interestingly, VD deficiency is associated with low induction of neurogenesis and the loss of apoptosis, generating larger brains due to abnormal proliferation (Eyles et al., 2013). This statement corroborates our GO results that indicate an increase in the negative regulation of apoptosis in EGEP-Network (**Table 3**). Thus, by affecting VD metabolism, VDR-dependent transcription could also be affected by ethanol, not only through target-protein recruitment but also through the formation of heterodimers with RXR, culminating in the loss of RA signaling.

Another gene directly related to VD metabolism is CYP27B1. The gene product converts 25-OHD_3 into 24,25-hydroxy vitamin D_3 and 1,25-hydroxy vitamin D_3 (Eyles et al., 2013). CYP27B1 was found to be underexpressed in the murine pups exposed to ethanol (PSE-Network; **Fig. 2B**).

Another important gene found among the overexpressed genes in the LGEP-Network (**Fig. 2D**) is CDK11B (CDK11p58) (**S-Table 1**, see **Supplementary Material 1**). Remarkably, CDK11p58 promotes the inhibition of VDR through ubiquitin-proteasome-mediated degradation (Chi et al., 2009), indicating that ethanol could interfere with VD action by promoting the degradation of VDR. Consistent with that hypothesis, protein polyubiquitination was present in Cluster 5, ubiquitin-dependent catabolic processing was present in Cluster 15 (**S-Table 7 and 17**, see **Supplementary Material 2**), and the proteolytic genes were among those overexpressed in the transcriptomic sets of the PE- and PSE-Networks.

4.1.3 Interplay between ethanol exposure, vitamin deficiency, glutamate and neuroinflammation

A major result of this systems chemo-biology analysis is that ethanol has been directly connected to the positive induction of inflammation (clusters 3-6, 8, 10-11 and 14; **S-Tables 5-8, 10, 12-13** and **16**, respectively, see **Supplementary Materials 2**), especially in the overexpression of genes in murine adult individuals exposed to ethanol during embryogenesis (EGEP-Network; **Fig. 2C**; **S-Table 20**, see **Supplementary Material 2**). This result indicates that the induction of inflammation occurs during early gestation and extends through development into adult. Consistent with these data, VDR deletion was observed to reduce the activity of IκBα protein, a potent inhibitor of the inflammatory-associated transcriptional factor NFκB, (Wu et al., 2010). It is important to note that inflammatory insults during pregnancy have already been correlated with Alzheimer's and Parkinson's diseases (Miller and O'Callaghan, 2008).

Indeed, the activation of NFκB in adults leads to amyloid-β accumulation due to the synthesis of leukotriene D4 in cortical neurons (Wang et al., 2013).

In addition, IκBα promotes neurodifferentiation by blocking self-renewal and by indirectly reducing the levels of Repressor Element Silencing Transcription Factor (REST), an inhibitor of neurogenesis (Khoshnan and Patterson, 2012).

Corroborating the hypothesis that ethanol can promote neuroinflammation through VDR downregulation, IκBα was found to be underexpressed in the overlaps between the PE- and PSE-Networks (**Fig. 4**; **Table 2**). VD and AT, as well as RA, are fat-soluble vitamins and may protect neurons against inflammation.

4.1.4 Vitamin deficiency driven by ethanol exposure leads to altered glutamate uptake

Glutamate has a major role as an excitatory neurotransmitter in the mammalian brain and is responsible for multiple aspects of neural activity, such as cognition and memory, which are both affected by FAS (O'Leary, 2004; Ruediger and Bolz, 2007). However, the overstimulation of glutamate can be responsible for brain injury and neuron apoptosis (Lu et al., 2013). The data gathered in this study showed that fetuses exposed to ethanol (EGEP-Network; **Fig. 2C**) early in development display an overstimulation of glutamine metabolism, which is a normal component controlling of glutamate levels in the central nervous system through the glutamine-glutamate cycle (**S-Table 20**, see **Supplementary Material 2**).

Glutamate also activates G-protein-coupled metabotropic receptors that can exert their effects through the cyclic adenosine monophosphate (cAMP) pathway

(Ruediger and Bolz, 2007), which is related to neurodevelopment and melatonin regulation (de Faria Poloni et al., 2011). Interestingly, the cAMP pathway is downregulated (**Table 3**) in fetuses exposed to ethanol during development (PE-Network; **Fig. 2A**) and also in pups postnatally exposed to ethanol (PSE-Networks; **Fig. 2B**). This indicates that ethanol exposure in the brain might have a negative effect on the cAMP pathway, thereby resulting in defects in G-protein-coupled receptor activity.

Ascorbic acid (AC) is released into the extracellular space to protect neurons exposed to cytotoxic concentrations of glutamate (Lane and Lawen, 2013). AC is a water-soluble vitamin, and water-soluble vitamin metabolic processes were among the GOs observed in cluster 12 (**S-Table 14**, see **Supplementary Material 2**).

The results from fetuses exposed to ethanol during the late phase of development (LGEP-Network; **Fig. 2D**) indicated that the gene coding for nicotinamide-nucleotide adenylyltransferase (NMNAT2), an enzyme predominantly expressed in the brain and related to NADP biosynthesis, is underexpressed. This result is interesting, as it seems that nicotinamide [niacin (NC)] deficiency is already correlated to neuronal damage (**Table 1**). It should be pointed the NMNAT2 was among the overexpressed genes in the EGEP-Network, indicating that the ethanol-induced deficiency in NC may be more aggressive in later stages of development.

In summary, in our chemo-systems biology analysis, we prospected PPI-CPI networks and combined the topological, GO and transcriptomic analyses of four ethanol-exposed groups of mice at different ages. The results gathered from this work elucidated FAS development and its interaction with vitamin metabolism. Ethanol appears to impair biological processes such as (i) the circadian cycle; (ii) calcium ion homeostasis; (iii) the glutamine pathway; (iv) the cAMP pathway; (v) inflammation; (vi)

neuron differentiation; and (vii) synapse formation and plasticity. These processes appear to be closely related to vitamin metabolism, particularly for RA, VD, AC, AT, NC and FA. Because ethanol is already correlated with vitamin deficiency and vitamins are crucial for brain development, understanding the relationship between ethanol and vitamins appears to be essential for preventing the development of FAS and related outcomes. The targets selected in this work by data crossing and HB analysis among the different ethanol-exposed mice also generated important targets to be reviewed for FAS prevention and treatment because none of them had previously been correlated with FAS. Tubulin and tubulin-associated proteins, synapse plasticity proteins and the proteins related to neurodifferentiation are of particular interest.

References

Adam SA, Schnell O, Poschl J et al. 2012. ALDH1A1 is a marker of astrocytic differentiation during brain development and correlates with better survival in glioblastoma patients. *Brain Pathol* **22(6)**:788-797.

Adlard BP, De Souza SW, Moon S. 1974. Ascorbic acid in fetal human brain. *Arch Dis Child* **49(4)**:278-282.

Ahn EY, DeKelver RC, Lo MC et al. 2011. SON controls cell-cycle progression by coordinated regulation of RNA splicing. *Mol Cell* **42(2)**:185-198.

Ahn J, Albanes D, Berndt SI et al. 2009. Vitamin D-related genes, serum vitamin D concentrations and prostate cancer risk. *Carcinogenesis* **30(5)**:769-776.

Aikawa H, Suzuki K. 1986. Lesions in the skin, intestine, and central nervous system induced by an antimetabolite of niacin. *Am J Pathol* **122(2)**:335-342.

Alvarez-Dolado M, Gonzalez-Moreno M, Valencia A et al. 1999. Identification of a mammalian homologue of the fungal Tom70 mitochondrial precursor protein import receptor as a thyroid hormone-regulated gene in specific brain regions. *J Neurochem* **73(6)**:2240-2249.

Ba A. 2005. Functional vulnerability of developing central nervous system to maternal thiamine deficiencies in the rat. *Dev Psychobiol* **47(4)**:408-414.

Ba A, Seri BV, Han SH. 1996. Thiamine administration during chronic alcohol intake in pregnant and lactating rats: effects on the offspring neurobehavioural development. *Alcohol Alcohol* **31(1)**:27-40.

Bader GD, Hogue CW. 2003. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* **4**:2.

Baksi SN, Hughes MJ. 1982. Chronic vitamin D deficiency in the weanling rat alters catecholamine metabolism in the cortex. *Brain Res* **242(2)**:387-390.

Beaudin AE, Abarinov EV, Noden DM et al. 2011. Shmt1 and de novo thymidylate biosynthesis underlie folate-responsive neural tube defects in mice. *Am J Clin Nutr* **93(4)**:789-798.

Benjamini Y, Hochberg Y. 1995. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B-Methodological* **57(1)**:289-300.

Bhate V, Deshpande S, Bhat D et al. 2008. Vitamin B12 status of pregnant Indian women and cognitive function in their 9-year-old children. *Food Nutr Bull* **29(4)**:249-254.

Binns D, Dimmer E, Huntley R et al. 2009. QuickGO: a web-based tool for Gene Ontology searching. *Bioinformatics* **25(22)**:3045-3046.

Bjorneboe GE, Bjorneboe A, Hagen BF et al. 1987. Reduced hepatic alpha-tocopherol content after long-term administration of ethanol to rats. *Biochim Biophys Acta* **918(3)**:236-241.

Black MM. 2008. Effects of vitamin B12 and folate deficiency on brain development in children. *Food Nutr Bull* **29(2 Suppl)**:S126-131.

Bowers K, Li Q, Bressler J et al. 2011. Glutathione pathway gene variation and risk of autism spectrum disorders. *J Neurodev Disord* **3(2)**:132-143.

Brager AJ, Ruby CL, Prosser RA et al. 2010. Chronic ethanol disrupts circadian photic entrainment and daily locomotor activity in the mouse. *Alcohol Clin Exp Res* **34(7)**:1266-1273.

Brodsky G, Barnes T, Bleskan J et al. 1997. The human GARS-AIRS-GART gene encodes two proteins which are differentially expressed during human brain development and temporally overexpressed in cerebellum of individuals with Down syndrome. *Hum Mol Genet* **6(12)**:2043-2050.

Brosh S, Sperling O, Bromberg Y et al. 1990. Developmental changes in the activity of enzymes of purine metabolism in rat neuronal cells in culture and in whole brain. *J Neurochem* **54(5)**:1776-1781.

Cammer W, Downing M. 1991. Localization of the multifunctional protein CAD in astrocytes of rodent brain. *J Histochem Cytochem* **39(5)**:695-700.

Carbon S, Ireland A, Mungall CJ et al. 2009. AmiGO: online access to ontology and annotation data. *Bioinformatics* **25(2)**:288-289.

Caspi R, Altman T, Dale JM et al. 2010. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res* **38(Database issue)**:D473-479.

Castro MA, Filho JL, Dalmolin RJ et al. 2009. ViaComplex: software for landscape analysis of gene expression networks in genomic context. *Bioinformatics* **25(11)**:1468-1469.

Chandra N, Padiadpu J. 2013. Network approaches to drug discovery. *Expert Opin Drug Discov* **8(1)**:7-20.

Chen L, Lau AG, Sarti F. 2012a. Synaptic retinoic acid signaling and homeostatic synaptic plasticity. *Neuropharmacology* 10.1016/j.neuropharm.2012.12.004.

Chen X, Burdett TC, Desjardins CA et al. 2013. Disrupted and transgenic urate oxidase alter urate and dopaminergic neurodegeneration. *Proc Natl Acad Sci U S A* **110(1)**:300-305.

Chen Y, Wang Z, Xie Y et al. 2012b. Folic acid deficiency inhibits neural rosette formation and neuronal differentiation from rhesus monkey embryonic stem cells. *J Neurosci Res* **90(7)**:1382-1391.

Chi Y, Hong Y, Zong H et al. 2009. CDK11p58 represses vitamin D receptor-mediated transcriptional activation through promoting its ubiquitin-proteasome degradation. *Biochem Biophys Res Commun* **386(3)**:493-498.

Cipriani S, Desjardins CA, Burdett TC et al. 2012. Urate and its transgenic depletion modulate neuronal vulnerability in a cellular model of Parkinson's disease. *PLoS One* **7(5)**:e37331.

Craciunescu CN, Brown EC, Mar MH et al. 2004. Folic acid deficiency during late gestation decreases progenitor cell proliferation and increases apoptosis in fetal mouse brain. *J Nutr* **134(1)**:162-166.

Csermely P, Korcsmaros T, Kiss HJ et al. 2013. Structure and dynamics of molecular networks: A novel paradigm of drug discovery: A comprehensive review. *Pharmacol Ther* 10.1016/j.pharmthera.2013.01.016.

Danielyan KE, Abramyan RA, Galoyan AA et al. 2011. Vitamin B-complex initiates growth and development of human embryonic brain cells in vitro. *Bull Exp Biol Med* **151(5)**:579-583.

de Faria Poloni J, Feltes BC, Bonatto D. 2011. Melatonin as a central molecule connecting neural development and calcium signaling. *Funct Integr Genomics* **11(3)**:383-388.

Ebadi M. 1981. Regulation and function of pyridoxal phosphate in CNS. *Neurochem Int* **3(3-4)**:181-205.

Eyles D, Burne T, McGrath J. 2011. Vitamin D in fetal brain development. *Semin Cell Dev Biol* **22(6)**:629-636.

Eyles DW, Burne TH, McGrath JJ. 2013. Vitamin D, effects on brain development, adult brain function and the links between low levels of vitamin D and neuropsychiatric disease. *Front Neuroendocrinol* **34(1)**:47-64.

Feltes BC, Poloni Jde F, Notari DL et al. 2013. Toxicological effects of the different substances in tobacco smoke on human embryonic development by a systems chemo-biology approach. *PLoS One* **8(4)**:e61743.

Focher F, Mazzarello P, Verri A et al. 1990. Activity profiles of enzymes that control the uracil incorporation into DNA during neuronal development. *Mutat Res* **237(2)**:65-73.

Garcia M, Leonardi R, Zhang YM et al. 2012. Germline deletion of pantothenate kinases 1 and 2 reveals the key roles for CoA in postnatal metabolism. *PLoS One* **7(7)**:e40871.

Genetta T, Lee BH, Sola A. 2007. Low doses of ethanol and hypoxia administered together act synergistically to promote the death of cortical neurons. *J Neurosci Res* **85(1)**:131-138.

Goez HR, Scott O, Hasal S. 2011. Fetal exposure to alcohol, developmental brain anomaly, and vitamin a deficiency: a case report. *J Child Neurol* **26(2)**:231-234.

Goldshmit Y, Munro K, Leong SY et al. 2010. LPA receptor expression in the central nervous system in health and following injury. *Cell Tissue Res* **341(1)**:23-32.

Gueant JL, Caillerez-Fofou M, Battaglia-Hsu S et al. 2013. Molecular and cellular effects of vitamin B12 in brain, myocardium and liver through its role as co-factor of methionine synthase. *Biochimie* **95(5)**:1033-1040.

Harms LR, Burne TH, Eyles DW et al. 2011. Vitamin D and the brain. *Best Pract Res Clin Endocrinol Metab* **25(4)**:657-669.

Hewitt AJ, Knuff AL, Jefkins MJ et al. 2011. Chronic ethanol exposure and folic acid supplementation: fetal growth and folate status in the maternal and fetal guinea pig. *Reprod Toxicol* **31(4)**:500-506.

Howerton CL, Morgan CP, Fischer DB et al. 2013. O-GlcNAc transferase (OGT) as a placental biomarker of maternal stress and reprogramming of CNS gene transcription in development. *Proc Natl Acad Sci U S A* **110(13)**:5169-5174.

Iba MM, Storch A, Ghosal A et al. 2003. Constitutive and inducible levels of CYP1A1 and CYP1A2 in rat cerebral cortex and cerebellum. *Arch Toxicol* **77(10)**:547-554.

Ichi S, Nakazaki H, Boshnjaku V et al. 2012. Fetal neural tube stem cells from Pax3 mutant mice proliferate, differentiate, and form synaptic connections when stimulated with folic acid. *Stem Cells Dev* **21(2)**:321-330.

Jaurena MB, Carri NG, Battiato NL et al. 2011. Trophic and proliferative perturbations of in vivo/in vitro cephalic neural crest cells after ethanol exposure are prevented by neurotrophin 3. *Neurotoxicol Teratol* **33(3)**:422-430.

Jensen LJ, Kuhn M, Stark M et al. 2009. STRING 8--a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res* **37(Database issue)**:D412-416.

Jiang W, Yu Q, Gong M et al. 2012. Vitamin A deficiency impairs postnatal cognitive function via inhibition of neuronal calcium excitability in hippocampus. *J Neurochem* **121(6)**:932-943.

Josey BJ, Inks ES, Wen X et al. 2013. Structure-activity relationship study of vitamin k derivatives yields highly potent neuroprotective agents. *J Med Chem* **56(3)**:1007-1022.

Jupe S, Akkerman JW, Soranzo N et al. 2012. Reactome - a curated knowledgebase of biological pathways: megakaryocytes and platelets. *J Thromb Haemost* 10.1111/j.1538-7836.2012.04930.x.

Kanehisa M, Goto S. 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28(1)**:27-30.

Kapoor N, Pant AB, Dhawan A et al. 2006. Cytochrome P450 1A isoenzymes in brain cells: Expression and inducibility in cultured rat brain neuronal and glial cells. *Life Sci* **79(25)**:2387-2394.

Kashyap V, Gudas LJ. 2010. Epigenetic regulatory mechanisms distinguish retinoic acid-mediated transcriptional responses in stem cells and fibroblasts. *J Biol Chem* **285(19)**:14534-14548.

Khoshnan A, Patterson PH. 2012. Elevated IKKalpha accelerates the differentiation of human neuronal progenitor cells and induces MeCP2-dependent BDNF expression. *PLoS One* **7(7)**:e41794.

Kirsch SH, Herrmann W, Obeid R. 2013. Genetic defects in folate and cobalamin pathways affecting the brain. *Clin Chem Lab Med* **51(1)**:139-155.

Krishna AP, Ramakrishna T. 2004. Effect of pyridoxine deficiency on the structural and functional development of hippocampus. *Indian J Physiol Pharmacol* **48(3)**:304-310.

Kuhn M, Szklarczyk D, Franceschini A et al. 2012. STITCH 3: zooming in on protein-chemical interactions. *Nucleic Acids Res* **40(Database issue)**:D876-880.

Kurimoto K, Kuwasako K, Sandercock AM et al. 2009. AU-rich RNA-binding induces changes in the quaternary structure of AUH. *Proteins* **75(2)**:360-372.

Lane DJ, Lawen A. 2013. The glutamate aspartate transporter (GLAST) mediates L-glutamate-stimulated ascorbate-release via swelling-activated anion channels in cultured neonatal rodent astrocytes. *Cell Biochem Biophys* **65(2)**:107-119.

Lee JY, Chang MY, Park CH et al. 2003. Ascorbate-induced differentiation of embryonic cortical precursors into neurons and astrocytes. *J Neurosci Res* **73(2)**:156-165.

Leung KY, De Castro SC, Cabreiro F et al. 2013. Folate metabolite profiling of different cell types and embryos suggests variation in folate one-carbon metabolism, including developmental changes in human embryonic brain. *Mol Cell Biochem* 10.1007/s11010-013-1613-y.

Lim CS, Alkon DL. 2012. Protein kinase C stimulates HuD-mediated mRNA stability and protein expression of neurotrophic factors and enhances dendritic maturation of hippocampal neurons in culture. *Hippocampus* **22(12)**:2303-2319.

Lopez-Fanarraga M, Carranza G, Bellido J et al. 2007. Tubulin cofactor B plays a role in the neuronal growth cone. *J Neurochem* **100(6)**:1680-1687.

Lu XC, Dave JR, Chen Z et al. 2013. Nefiracetam attenuates post-ischemic nonconvulsive seizures in rats and protects neuronal cell death induced by veratridine and glutamate. *Life Sci* 10.1016/j.lfs.2013.04.004.

Macchi M, El Fissi N, Tufi R et al. 2013. The Drosophila inner-membrane protein PMI controls crista biogenesis and mitochondrial diameter. *J Cell Sci* **126(Pt 3)**:814-824.

Maekawa M, Ohnishi T, Hashimoto K et al. 2010. Analysis of strain-dependent prepulse inhibition points to a role for Shmt1 (SHMT1) in mice and in schizophrenia. *J Neurochem* **115(6)**:1374-1385.

Maere S, Heymans K, Kuiper M. 2005. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* **21(16)**:3448-3449.

Maffi SK, Rathinam ML, Cherian PP et al. 2008. Glutathione content as a potential mediator of the vulnerability of cultured fetal cortical neurons to ethanol-induced apoptosis. *J Neurosci Res* **86(5)**:1064-1076.

Marei HE, Ahmed AE, Michetti F et al. 2012. Gene expression profile of adult human olfactory bulb and embryonic neural stem cell suggests distinct signaling pathways and epigenetic control. *PLoS One* **7(4)**:e33542.

Meijer OC, Steenbergen PJ, De Kloet ER. 2000. Differential expression and regional distribution of steroid receptor coactivators SRC-1 and SRC-2 in brain and pituitary. *Endocrinology* **141(6)**:2192-2199.

Miller DB, O'Callaghan JP. 2008. Do early-life insults contribute to the late-life development of Parkinson and Alzheimer diseases? *Metabolism* **57 Suppl 2**:S44-49.

Miller GW, Ulatowski L, Labut EM et al. 2012. The alpha-tocopherol transfer protein is essential for vertebrate embryogenesis. *PLoS One* **7(10)**:e47402.

Morse NL. 2012. Benefits of docosahexaenoic acid, folic acid, vitamin D and iodine on foetal and infant brain development and function following maternal supplementation during pregnancy and lactation. *Nutrients* **4(7)**:799-840.

Nakajima M, Furukawa S, Hayashi K et al. 1993. Age-dependent survival-promoting activity of vitamin K on cultured CNS neurons. *Brain Res Dev Brain Res* **73(1)**:17-23.

Newman MEJ. 2005. A measure of betweenness centrality based on random walks. *Social Networks* **27(1)**:39-54.

Noelker C, Schwake M, Balzer-Geldsetzer M et al. 2012. Differentially expressed gene profile in the 6-hydroxy-dopamine-induced cell culture model of Parkinson's disease. *Neurosci Lett* **507(1)**:10-15.

Nomoto M, Takeda Y, Uchida S et al. 2012. Dysfunction of the RAR/RXR signaling pathway in the forebrain impairs hippocampal memory and synaptic plasticity. *Mol Brain* **5**:8.

O'Leary CM. 2004. Fetal alcohol syndrome: diagnosis, epidemiology, and developmental outcomes. *J Paediatr Child Health* **40(1-2)**:2-7.

Oehlmann VD, Berger S, Sterner C et al. 2004. Zebrafish beta tubulin 1 expression is limited to the nervous system throughout development, and in the adult brain is restricted to a subset of proliferative regions. *Gene Expr Patterns* **4(2)**:191-198.

Ogunleye AJ, Odutuga AA. 1989. The effect of riboflavin deficiency on cerebrum and cerebellum of developing rat brain. *J Nutr Sci Vitaminol (Tokyo)* **35(3)**:193-197.

Pangilinan F, Molloy AM, Mills JL et al. 2012. Evaluation of common genetic variants in 82 candidate genes as risk factors for neural tube defects. *BMC Med Genet* **13**:62.

Park H, Yamada K, Kojo A et al. 2009. Drebrin (developmentally regulated brain protein) is associated with axo-somatic synapses and neuronal gap junctions in rat mesencephalic trigeminal nucleus. *Neurosci Lett* **461(2)**:95-99.

Park JH, Lee SB, Lee KH et al. 2012. Nuclear Akt promotes neurite outgrowth in the early stage of neuritogenesis. *BMB Rep* **45(9)**:521-525.

Rebhan M, Chalifa-Caspi V, Prilusky J et al. 1997. GeneCards: integrating information about genes, proteins and diseases. *Trends Genet* **13(4)**:163.

Rhinn M, Dolle P. 2012. Retinoic acid signalling during development. *Development* **139(5)**:843-858.

Rodriguez-Rodriguez E, Mateo I, Infante J et al. 2009. Interaction between HMGCR and ABCA1 cholesterol-related genes modulates Alzheimer's disease risk. *Brain Res* **1280**:166-171.

Roehrs T, Roth T. 2001. Sleep, sleepiness, sleep disorders and alcohol use and abuse. *Sleep Med Rev* **5(4)**:287-297.

Rosado JO, Henriques JP, Bonatto D. 2011. A systems pharmacology analysis of major chemotherapy combination regimens used in gastric cancer treatment: predicting potential new protein targets and drugs. *Curr Cancer Drug Targets* **11(7)**:849-869.

Ross ME. 2010. Gene-environment interactions, folate metabolism and the embryonic nervous system. *Wiley Interdiscip Rev Syst Biol Med* **2(4)**:471-480.

Ruediger T, Bolz J. 2007. Neurotransmitters and the development of neuronal circuits. *Adv Exp Med Biol* **621**:104-115.

Russo E, Citraro R, Constanti A et al. 2012. The mTOR signaling pathway in the brain: focus on epilepsy and epileptogenesis. *Mol Neurobiol* **46(3)**:662-681.

Safran M, Dalah I, Alexander J et al. 2010. GeneCards Version 3: the human gene integrator. *Database (Oxford)* **2010**:baq020.

Scardoni G, Petterlini M, Laudanna C. 2009. Analyzing biological network parameters with CentiScaPe. *Bioinformatics* **25(21)**:2857-2859.

Schmidt MV, Oitzl M, Steenbergen P et al. 2007. Ontogeny of steroid receptor coactivators in the hippocampus and their role in regulating postnatal HPA axis function. *Brain Res* **1174**:1-6.

Schneider HC, Klabunde T. 2013. Understanding drugs and diseases by systems biology? *Bioorg Med Chem Lett* **23(5)**:1168-1176.

Shannon P, Markiel A, Ozier O et al. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13(11)**:2498-2504.

Shearer KD, Stoney PN, Morgan PJ et al. 2012. A vitamin for the brain. *Trends Neurosci* **35(12)**:733-741.

Singleton CK, Martin PR. 2001. Molecular mechanisms of thiamine utilization. *Curr Mol Med* **1(2)**:197-207.

Snel B, Lehmann G, Bork P et al. 2000. STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic Acids Res* **28(18)**:3442-3444.

Strate I, Min TH, Iliev D et al. 2009. Retinol dehydrogenase 10 is a feedback regulator of retinoic acid signalling during axis formation and patterning of the central nervous system. *Development* **136(3)**:461-472.

Tan XL, Zhai Y, Gao WX et al. 2009. p300 expression is induced by oxygen deficiency and protects neuron cells from damage. *Brain Res* **1254**:1-9.

Tetel MJ. 2009. Nuclear receptor coactivators: essential players for steroid hormone action in the brain and in behaviour. *J Neuroendocrinol* **21(4)**:229-237.

Tsaioun KI. 1999. Vitamin K-dependent proteins in the developing and aging nervous system. *Nutr Rev* **57(8)**:231-240.

Tveden-Nyborg P, Lykkesfeldt J. 2009. Does vitamin C deficiency result in impaired brain development in infants? *Redox Rep* **14(1)**:2-6.

Tveden-Nyborg P, Vogt L, Schjoldager JG et al. 2012. Maternal vitamin C deficiency during pregnancy persistently impairs hippocampal neurogenesis in offspring of guinea pigs. *PLoS One* **7(10)**:e48488.

van de Rest O, van Hooijdonk LW, Doets E et al. 2012. B vitamins and n-3 fatty acids for brain development and function: review of human studies. *Ann Nutr Metab* **60(4)**:272-292.

van der Wulp NY, Hoving C, de Vries H. 2013. A qualitative investigation of alcohol use advice during pregnancy: Experiences of Dutch midwives, pregnant women and their partners. *Midwifery* 10.1016/j.midw.2012.11.014.

Veena SR, Krishnaveni GV, Srinivasan K et al. 2010. Higher maternal plasma folate but not vitamin B-12 concentrations during pregnancy are associated with better cognitive function scores in 9- to 10- year-old children in South India. *J Nutr* **140(5)**:1014-1022.

Vitobello A, Ferretti E, Lampe X et al. 2011. Hox and Pbx factors control retinoic acid synthesis during hindbrain segmentation. *Dev Cell* **20(4)**:469-482.

Wang XY, Tang SS, Hu M et al. 2013. Leukotriene D4 induces amyloid-beta generation via CysLT(1)R-mediated NF-kappaB pathways in primary neurons. *Neurochem Int* **62(3)**:340-347.

Wentzel P, Eriksson UJ. 2009. Altered gene expression in neural crest cells exposed to ethanol in vitro. *Brain Res* **1305 Suppl**:S50-60.

Wu S, Xia Y, Liu X et al. 2010. Vitamin D receptor deletion leads to reduced level of IkappaBalpha protein through protein translation, protein-protein interaction, and post-translational modification. *Int J Biochem Cell Biol* **42(2)**:329-336.

Yokoi K, Ito T, Maeda Y et al. 2009. A case of holocarboxylase synthetase deficiency with insufficient response to prenatal biotin therapy. *Brain Dev* **31(10)**:775-778.

Yu H, Kim PM, Sprecher E et al. 2007. The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. *PLoS Comput Biol* **3(4)**:e59.

Zhou FC, Zhao Q, Liu Y et al. 2011. Alteration of gene expression by alcohol exposure at early neurulation. *BMC Genomics* **12**:124.

Tables

Table 1. Major vitamins present in our CPI-PPI networks for *M. musculus* and their role during neurodevelopment or proper neural tissue function throughout embryonic brain development or neuronal *in vitro* lineages.

Vitamin	Role in the neural tissue
Vitamin A (RA)	RA is related to the control of the hindbrain and forebrain development and neural tube differentiation (Jiang et al., 2012; Rhinn and Dolle, 2012). It also regulates neuronal patterning along the anterior-posterior axis (Shearer et al., 2012).
Vitamin C (AC)	Vitamin C is essential for hippocampal development and for hippocampal postnatal function in guinea pigs (Tveden-Nyborg et al., 2012). Vitamin C also exerts anti-oxidant effects, preventing neurotoxic insults in the brain, such as ROS production (Tveden-Nyborg and Lykkesfeldt, 2009). Another study showed that ascorbic acid is highly present throughout midbrain development during human pregnancy (Adlard et al., 1974). Moreover, ascorbic acid was able to induce differentiation of CNS precursor cells into neurons and astrocytes (Lee et al., 2003)
Vitamin D [25-(OH) D_3]/ [1,25-$(OH)_2D_3$]	Vitamin D induces neurite formation (Eyles et al., 2011). Additionally, 25-hydroxyvitamin D_3 upregulates nerve growth factor (NGF), which is essential for survival and growth of hippocampal and forebrain neurons (Eyles et al., 2011). Vitamin D deficiency is also correlated with decreased apoptosis and diminished cortex thickness (Eyles et al., 2013; Harms et al., 2011). The vitamin D receptor (VDR) is also present in the hippocampus (Eyles et al., 2013).
Vitamin K (PQN/MQN)	Promotes survival of cultured rat embryo CNS neurons (Nakajima et al., 1993). Pregnant women treated with vitamin K antagonist (warfarin) presented fetuses with abnormal dilatation of cerebral ventricle, microcephaly and mental retardation (Tsaioun, 1999). Showed neuroprotective role against oxidative stress (Josey et al., 2013).
Vitamin E (α-TC)	Described as playing an essential role in early brain formation, where tocopherol transporter protein (TTP) was present in the hindbrain and forebrain in zebrafish (Miller et al., 2012).
Vitamin B_{12} (CBL)	CBL is related to the process of myelination (Black, 2008), and deficiency in vitamin B_{12} is related to neural tube defects (Kirsch et al., 2013; van de Rest et al., 2012; Veena et al., 2010). Although not confirmed, a study shows that vitamin B_{12} deficiency might play a role in the developing brain and may change the normal cognitive status later in life (Bhate et al., 2008).
Vitamin B_9 (FA)	FA supplementation reduces the risk of neural tube defects in human embryos (Kirsch et al., 2013; Leung et al., 2013; Ross, 2010). It is also essential for fetal spine and cranial formation (Morse, 2012). Deficiency in FA is also related to inhibited neural rosette differentiation in monkey embryonic stem cells (Chen et al., 2012b).
Vitamin B_6 (PDX)	Pyridoxine was related to increased survival of neuronal cells *in vitro* by stimulation neurotransmitter release (Danielyan et al., 2011). Additionally, a study in rats shows that PDX deficiency caused diminished hippocampal weight and electrical activity, most likely due to poor myelination (Krishna and Ramakrishna, 2004). The catalytic form of vitamin B6 (pyridoxal phosphate) is also found in multiple parts of the brain and has its highest concentration in the olfactory tubercle (Ebadi, 1981).

Table 1 (*Continued*…)

Vitamin B_5 (PA)	Inactivation of panthotenate kinases, which phosphorylates PA, is related to neurodegeneration diseases during childhood . Nevertheless, no studies have been performed to elucidate the role of PA alone during brain formation throughout embryogenesis.
Vitamin B_3 (NC)	Newborn mice injected with an antagonist of niacin showed damage in the central nervous system (CNS), and motor neurons as well as dorsal horn cells in the spinal cord showed signs of chromatolysis (Aikawa and Suzuki, 1986). However, no studies have been performed to elucidate the role of NC alone during brain formation throughout embryogenesis.
Vitamin B_2 (RBF)	RBF deficiency reduced the levels of important components of the myelin membrane in adult rats (Ogunleye and Odutuga, 1989). However, no studies have been performed to elucidate the role of RBF alone during brain formation throughout embryogenesis.
Vitamin B_1 (TM)	TM deficiency in rats caused abnormal growth of the hippocampus (Ba et al., 1996), and appears to affect myelinogenesis, axonal growth and synapsis formation (Ba, 2005).
Vitamin H (BT)	Errors in biotin metabolism can cause enlargement of cerebral ventricles (Yokoi et al., 2009).

Legends: RA = Retinoic acid; **AC** = Ascorbic Acid; **25-(OH) D_3** = 25-hydroxyvitamin D_3; **1,25-$(OH)_2D_3$** = 1,25 hydroxyvitamin D_3; **PQN** = Phylloquinone; **MQN** = Menaquinone; **α-TC** = α-Tocopherol; **CBL** = Cobalamin; **FA** = Folic Acid (Folate); **PDX** = Pyridoxine; **PA** = Panthotenic Acid; **NC** = Niacin; **RBF** = Riboflavin; **TM** = Thiamine; **BT** = Biotin.

Table 2. List of overlapping nodes found in the under- and overexpressed genes of all four networks (**Figs. 3)** and among the HB networks (**Fig. 4)**. The full description of the most relevant targets and proteins are discussed along the study.

Protein	Identity	Role in neurodevelopment and/or neurological function	Expression
Adcy5	Adenylate cyclase	NDL	Underexpressed
Akr1d1	Aldo-Keto reductase	NDL	Underexpressed
Akt1	Kinase	Involved in neuronal differentiation (Park et al., 2012).	Underexpressed
Aldh1a2	Aldehyde dehydrogenase	Involved in the patterning of the CNS and neural tube (Marei et al., 2012; Strate et al., 2009). Could also be related to hindbrain defects in *Xenopus laevis* (Vitobello et al., 2011).	Underexpressed
Aldh1b1	Aldehyde dehydrogenase	Downregulated by ethanol during early nerulation (Zhou et al., 2011).	Underexpressed (EGEP-LGEP) Overexpressed (PE-PSE)
Aprt	Aphosphoribosyl-transferase	Aprt expression increases in course of neuron maturation in cell cultures (Brosh et al., 1990)	Underexpressed
Auh	Enoyl-CoA hydratase	Role in neural survival through its action on AU-rich elements (ARE) (Kurimoto et al., 2009).	Underexpressed
C1qtnf7	C1q and TNF related protein	NDL	Underexpressed
Cd3g	T-Cell surface glycoprotein	NDL	Underexpressed
Chuk (IκBα)	Serine/threonine kinase	Expression of this protein blocks self-renewal and induces neurodifferentiation (Khoshnan and Patterson, 2012).	Underexpressed
Coq6	Monooxygenase	NDL	Underexpressed
Cyp1a1	Cytochrome P450 family	Related to xenobiotic metabolism in the brain, where this protein was found with high activity in glial cells (Kapoor et al., 2006) and also abundant in the cerebral cortex and cerebellum (Iba et al., 2003).	Underexpressed
Cyp2c70	Cytochrome P450 family	NDL	Underexpressed
Cyp2d10	Cytochrome P450 family	NDL	Underexpressed
Dbn1	Actin-binding adapter protein	Plays a role in spine formation and synaptogenesis (Park et al., 2009)	Underexpressed
Ep300	Histone acetyltransferase	Expressed in multiple regions of the brain, including hippocampus, cerebral and cerebellar cortices and medulla oblongata (Tan et al., 2009)	Underexpressed

Table 2 (*Continued*...)

Gart	Phosphoribosyl-glycinamide Formyltransferase	Polymorphism in this gene was related to mouse neural tube defects (Pangilinan et al., 2012). This protein is also related to prenatal cerebellar development (Brodsky et al., 1997)	Underxpressed
Ggt1	Gamma-glutamyl transpeptidase	NDL	Underexpressed
Lpar2	Lysophosphatidic acid receptor	LPA has been implicated in neurogenesis of the CNS, targeting neural progenitors, neurons, astrocytes, microglia, oligodendrocytes and Schwann cells (Goldshmit et al., 2010)	Underexpressed
Mtor	Serine/Threonine kinase	Involved in synaptic plasticity, neuron survival and repair against brain injuries (Russo et al., 2012)	Underexpressed
Mvd	Mevalonate pyrophosphate decarboxylase	NDL	Underexpressed
Ncoa3 (Src3)	Histone acetyltransferase	Expressed at high levels in the hippocampus (Tetel, 2009).	Underexpressed
Nedd4	E3 ubiquitin-protein ligase	NDL	Underexpressed
Ogt	O-Linked N-acetylglucosamine transferase	Cellular nutrient sensor, which may play a role in placental protection and in neurodevelopment by protecting the brain from insults such as nutrient deficiency (Howerton et al., 2013)	Underexpressed
Olfr15	Olfactory receptor	NDL	Underexpressed
Olfr161	Olfactory receptor	NDL	Underexpressed
Pik3cd	Kinase	NDL	Underexpressed
Rpl32	Ribosomal protein	NDL	Underexpressed
Sap30	Sin3A-associated protein	NDL	Underexpressed
Shmt1	Serine hydroxymethyl-transferase	Related to prepulse inhibition in mice (Maekawa et al., 2010). Lack of Shmt1 also results in neural tube defects in mice (Beaudin et al., 2011)	Underexpressed
Tmen11 (PMI)	Transmembrane protein	Involved in *Drosophila melanogaster* synapse formation and lifespan (Macchi et al., 2013)	Underexpressed
Tubb1	Tubulin, Beta 1 Class VI	Protein restricted to regions of the peripheral and central nervous system during early-differentiating neurons in zebrafish (Oehlmann et al., 2004).	Underexpressed
Zfp622	Zinc finger protein	NDL	Underexpressed
Son	Splicing cofactor	Expression of Son was related to neurogenesis during embryogenesis and postnatal brain (Ahn et al., 2011).	Underexpressed

Table 2 (*Continued*…)

Aldh1a1	Aldehyde dehydrogenase	ALDH1A1 expression appears in more differentiated parts of the developing brain such as cerebellar vermis or fetal white matter (Adam et al., 2012)	Overexpressed
Cad	Trifunctional protein (carbamyl phosphate synthetase, aspartate transcarbamylase and, dihydroorotase).	Cad protein was observed to be elevated during rat and hamster prenatal brain formation and hamster early postnatal brain development (Cammer and Downing, 1991). The authors argue that Cad is related to pyrimidine synthesis in astrocytes and in the grey matter.	Overexpressed
Carm1	Methyltransferase	Expression of this protein, inhibits HuD, a protein that is important for neurodifferentiation, synaptogenesis and learning and memory (Lim and Alkon, 2012).	Overexpressed
Cdc25a	Phosphatase	NDL	Overexpressed
Cth	Broad substrate specificity (deaminase, dehydratase, lyase, desulfhydrase)	Polymorphisms in this gene were related to autism (Bowers et al., 2011).	Overexpressed
Dlat	Pyruvate dehydrogenase	NDL	Overexpressed
Dut (DUTPase)	Nucleotido-hydrolase	Expressed in the prenatal rat brain, DUTPase might is responsible to maintain the low frequency of dUMP incorporation into DNA (Focher et al., 1990).	Overexpressed
Hmgcr	Transmembrane glycoprotein	Overexpression of this protein, in combination to underexpression of ABCA1 (not present in our networks) is related to increased risk of Alzheimer's disease (Rodriguez-Rodriguez et al., 2009).	Overexpressed
Ikbkg	Kinase	NDL	Overexpressed
Med6	Transcription factor	NDL	Overexpressed
Mtr	Methyltransferase	Uses cobalamin (CBL) as co-factor, where deficiency of CBL causes dramatic degrease of Mtr (Gueant et al., 2013). In the same article is discussed that CBL deficiency during mice is associated with impaired memory.	Overexpressed
Ncoa2	Histone acetyltransferase	Overall, Ncoa2 expression is not detectable in the brain but is expressed in the anterior pituitary (Meijer et al., 2000) and in high levels in the dentate gyrus during adult stages but low on prenatal stages (Schmidt et al., 2007)	Overexpressed
Rasa1	GTPase-activating protein	NDL	Overexpressed
Tbcb	Tubulin folding cofactor B	Overexpression of TBCB results in abnormalities in the growth cone morphology, later causing neuronal degeneration (Lopez-Fanarraga et al., 2007)	Overexpressed
Tomm70a (KIAA0719)	Translocase	Regulated by thyroid hormone, which can lead to brain malformations when at abnormal levels (Alvarez-Dolado et al., 1999).	Overexpressed

Table 2 (*Continued*…)

Uox	Urate oxidase	Uox expression is correlated to diminished neuroprotective effects of urate in astrocytes and neurons (Cipriani et al., 2012). Its expression is also related to exacerbate the lesions caused by 6-hydroxydopmaine in dopaminergic neurons (Chen et al., 2013).	Overexpressed

Legend: NDL = No Direct Link

Table 3. Major GO terms referent to over- and underexpressed genes in the CPI-PPI networks for each transcriptomic sets.

Network (expression)	GO	Corr *p*-value	x	Proteins
PE-Network (overexpressed)	Negative regulation of apoptosis	4.6×10^{-13}	47	BMI1|SNCB|XIAP|PAFAH2|PRDX3|ITSN1|ADORA1|WT1|PCGF2|BDNF|CASP3|ATG5|LHX3|DNAJC5|MYC|HELLS|CD27|IHH|SPP1|FN1|ZC3HC1|MSH2|IL7|RXFP2|GRIN1|SPHK1|NR4A2|PROKR1|LIG4|DAPK1|ATF5|NME5|MNAT1|NOTCH1|EYA1|GNAQ|SFRP2|HIPK2|CX3CR1|SIX1|MTR|CFDP1|BMP7
	Vasculature development	4.1×10^{-7}	31	FGFR2|CAV1|NRP1|FGF9|LEPR|MMP2|WT1|CDH5|CTNNB1|SEMA5A|ATG5|APOE|TDGF1|GATAD2A|RHOB|NOS3|PLXND1|IHH|FN1|KLF5|SMAD5|SPHK1|EFNB2|GALT|FZD5|MNAT1|NOTCH1|PROK1|JUN|NTRK2|ZFPM2
	Aging	6.3×10^{-7}	14	MSH6|GNAO1|MSH2|POU1F1|GHRHR|NCAM1|CDKN2A|CYP27B1|APOE|MTR|MNT|SLC18A2|INPP5D|HAP1
	Negative regulation of cell communication	9.5×10^{-7}	28	HCRT|CAV1|FGFR3|GRIK1|FGF9|MBIP|ADORA1|GPC3|TDGF1|SKIL|INPP5D|AXIN1|IHH|PTPRC|AVP|PTPRF|PRKCD|CISH|SIGIRR|GRB10|CCND1|NOTCH1|BMPER|SFRP2|AVPR1A|BMP7|DRD1A|GRB14
	Negative regulation of neuron apoptosis	1.06×10^{-5}	12	BDNF|SNCB|XIAP|MSH2|HIPK2|SIX1|GRIN1|NR4A2|DNAJC5|PRDX3|LIG4|ITSN1
PE-Network (underexpressed)	Post-translational protein modification	9.8×10^{-10}	76	CDK19|CDC14B|STK35|PTPN22|RPS6KB2|LPAR2|LATS2|BTK|AKT1|GPX1|SIN3B|PRMT1|CRY2|PLOD1|SH2D1B1|PRKACA|FGF2|MAP2K7|EGFR|IRAK2|SRPK2|CAMK1G|PTPRM|PHKG2|CDK8|SOCS7|ARL6|PRKCQ|PPP1CA|EP300|PIAS4|PDGFRB|PIAS2|FBXO15|EIF2AK2|NSD1|UBE2T|MAP3K11|RAB3B|SRM|ERBB3|BRSK2|MAPKAPK3|TRIB3|KIT|EPHB3|CD74|GCKR|VRK1|MAP3K3|C1QTNF2|PKD1|PPP3CA|TCF3|PIK3R1|PTPN18|FLT4|TGFBR1|PTPRA|CS|PDE6G|EPHA1|RPS6KA1|GCK|PLK1|NEDD4|RNF2|NTRK1|PRKAR1A|GRK5|MTOR|MAPK8IP1|MERTK|IKBKB|BMPR1B|OPN4
	Calcium ion homeostasis	9.2×10^{-9}	25	GNA13|CCL2|PTGER3|PMCH|IL6ST|PIK3CB|HC|GRIK2|TRHR|PTH1R|NMB|PPOX|KCNA5|NPY1R|ITGB3|CSRP3|BAK1|HRH3|GCK|PLCG2|RYR1|TBXA2R|EPOR|BANK1|IL2
	Retinoid metabolic process	2.6×10^{-7}	15	EBP|CYP11A1|MVD|CYP11B1|AMACR|LSS|CPN2|SC4MOL|CYP17A1|INSIG2|AKR1C6|INSIG1|BMPR1B|FGF2|AKR1D1
	c-AMP-mediated signaling	1.9×10^{-6}	15	P2RY12|GNA13|ADRB3|NPB|ADRB1|PTGER3|ADCY8|S1PR4|ADCY5|PTH1R|LHCGR|HTR4|RAPGEF4|FSHR|OPRD1

Table 3 (*Continued*…)

	Cognition	5.9×10^{-4}	80	GLRA1|ADCY8|OLFR1254|OLFR703|OLFR295|UCHL1|RPE65|OLFR808|NR2E1|OLFR1016|GPX1|OLFR1469|OLFR692|OLFR1054|OLFR554|OLFR1427|OLFR228|PLCB2|GJE1|OLFR1058|OLFR1152|OLFR461|OLFR836|WNT10B|MYO6|OLFR460|OLFR1247|OLFR167|ESR2|OLFR1348|OLFR161|AAAS|OLFR399|OLFR59|OLFR11|OLFR96|OLFR15|OLFR90|OLFR392|OLFR1046|OLFR1045|OLFR16|OLFR1104|GJA10|OLFR1234|C3|OLFR1085|OLFR1442|OLFR829|OLFR1500|PPT1|KIT|OLFR137|OLFR729|OLFR437|OLFR821|OLFR578|OLFR381|ACE|HRH3|OLFR1176|OLFR720|OLFR1094|OLFR502|OLFR1226|OLFR282|OLFR1451|OLFR716|NPY1R|DBH|PDE6G|OLFR1458|OLFR2|OLFR449|OLFR993|OLFR1490|OLFR376|OLFR73|OPN4|OLFR1122
EGEP-Network (overexpressed)	Negative regulation of cell death	7.5×10^{-6}	30	RBP4|TSPO|CAV1|NR2E3|PDX1|IL15|PTTG1|SLFN3|GLI3|ADORA1|TGFB2|SLFN2|LIF|BDNF|GPC3|CDKN2B|HSF1|GATA3|RARA|ITCH|BMP2|WNT10B|JARID2|RALBP1|SMAD3|GJB6|CTH|PLA2G2A|GLMN|WNT11
EGEP-Network (underexpressed)	Positive regulation of axonogenesis	7.4×10^{-6}	9	NTRK3|APC2|TIAM1|PLXNB1|ADNP|PAFAH1B1|NEFL|DSCAM|NGF
	Positive regulation of cell communication	4.6×10^{-5}	25	IL6|FKBP8|UTS2|PPARD|CARD9|ERBB4|CD3E|TAC1|ITGA2|JAG1|DGKI|MBD2|FURIN|NCAM1|ACVR1B|CDKN2A|MYD88|CD36|AGT|EXOC4|ADAM17|IL1B|NMU|CHUK|GHR
	Response to axon injury	4.1×10^{-4}	5	LAMB2|BCL2|BAX|NEFL|MMP2
LGEP-Network (overexpressed)	Positive regulation of apoptosis	1.3×10^{-13}	42	USP7|CDK5R1|TLR4|RRM2B|NR3C1|ZBTB16|LPAR1|PMAIP1|MMP2|IL10|ALDH1A2|NOD1|ALDH1A3|TICAM1|PCSK9|DIABLO|INPP5D|FAS|TRAF6|CASP2|MAP2K7|MAP2K6|CCAR1|COL18A1|PRKCA|TXNIP|IL2RA|PTPRF|GRIN1|BRCA2|IDO1|ATM|CIDEC|NOTCH2|NOTCH1|ADRB2|PSEN1|EEF1E1|ENDOG|PDE5A|WNT11|LRP5
	TNF-mediated signaling pathway	1.9×10^{-4}	5	TRAF2|TNFRSF11A|TNFSF11|KRT18|FAS
LGEP-Network (underexpression)	Positive regulation of cell communication	4×10^{-9}	32	DCC|FGF18|FKBP8|CAV1|FGF9|CSF1|FGF10|LPAR2|EIF2A|ITGB3|TLR6|ITSN1|SRC|PHIP|ACVR1B|CD44|IFNG|GATA4|RBCK1|IL1A|CHUK|IL4|BMP4|DIXDC1|KL|CENPJ|KITL|WNT7B|CCR2|JAK2|MTOR|GHSR
	Lamellipodium assembly	7.2×10^{-4}	5	NCK1|SH2B1|CPB2|NCKAP1|FGD4
PSE-Network (overexpressed)	Negative regulation of apoptosis	2.3×10^{-10}	39	XRCC5|STIL|FGFR1|XIAP|SNCA|ELK1|NFKB1|BDKRB2|ADORA1|PHIP|BDNF|PTK2|CD44|ATG5|BCL2|AGT|PPP2CB|VNN1|NKX2-5|ERCC2|APC|BMP4|EEF1A2|SKP2|GIF|PROKR1|ESR2|TAX1BP1|DAPK1|RAD51|EYA1|TNFSF13B|IGBP1|MTR|CFDP1|TRP73|APIP|WNT7A|NGF
	Proteolysis	1.1×10^{-4}	41	C2|MASP1|CNDP2|UBE3A|MMP8|ENPEP|MMP2|PSMB4|CYLD|CUL7|PPP2CB|USP34|CUL1|CAPN7|SEC11C|UFD1L|FBXO2|SKP2|CAPN2|FURIN|AFG3L1|PSMB8|PSMB9|FOLH1|BLMH|CUL4A|TMPRSS11E|CLPP|PRCP|CTSC|ADAM12|TBL1X|CTSH|PMPCA|PMPCB|NCLN|PLAU
	Tachykinin receptor signaling pathway	2.1×10^{-4}	4	UQCRC2|METAP1|USP8|UQCRC1|APTACR2|TACR1|TAC1|TAC2

Table 3 (*Continued*...)

PSE-Network (underexpressed)	Axonogenesis	5.2×10^{-7}	24	FGFR2|ENAH|CDK5R1|WNT3A|UCHL1|KIF5C|DPYSL5| RTN4R|STXBP1|PIP5K1C|SLIT1|CXCL12|CTNNA2| NRCAM|ROBO1|CXCR4|MNX1|RIT1|SEMA3A|BMPR1B| BOC|APBB1|GAP43|KALRN
	Cytosolic ion calcium homeostasis	1.04×10^{-5}	16	CALCR|MCHR1|RXFP3|EDN1|PTH1R|NMB|CXCR3|ITPR3|EDN RA|GCK|AGTR1A|RYR1|TGM2|UTS2R|GLP1R| CACNA1A
	Regulation of c-AMP biosynthetic process	9×10^{-5}	13	CALCR|ADCY2|ADCYAP1R1|EDN1|PTH1R|TIMP2| EDNRA|S1PR3|HTR1B|S1PR4|HTR7|PTH|GLP1R
	Synaptic transmission	7.6×10^{-4}	21	GJD2|MYO5A|STX1A|HTT|GABRA6|MAOB|PPYR1| CLSTN1|STXBP1|SNAPIN|NTSR2|CTNNA2|CTNNB1| HTR1B|CAMK4|HTR7|HRG|VAMP2|TPR|SNAP25| CACNA1A

Legends: x = number of proteins associated with a given GO in the network.

Figure legends

Figure 1. Experimental systems chemo-biology workflow. In (A), the primary network is composed of 1,287 nodes (14 vitamins, ethanol and 1,262 proteins). The nodes linked to the ethanol subnetwork are shown with red borders. The vitamins were selected and used as initial inputs for searching different small subnetworks that were merged into a single large interactome. B) Secondary network composed of 2,213 nodes (14 vitamins, ethanol and 2198 proteins). The data for proteins associated with vitamin and ethanol metabolism were gathered from multiple databases and merged with the primary network. C) *Mus musculus*-Ethanol-Network (MME-Network), composed of 7,723 nodes (14 vitamins, ethanol and 7,708 proteins). The data gathered from the microarrays were collected from the GEO database and were then entered into STRING and merged with the Secondary network. The MME-Network was further analyzed with Cytoscape 2.8.2 and 3.0

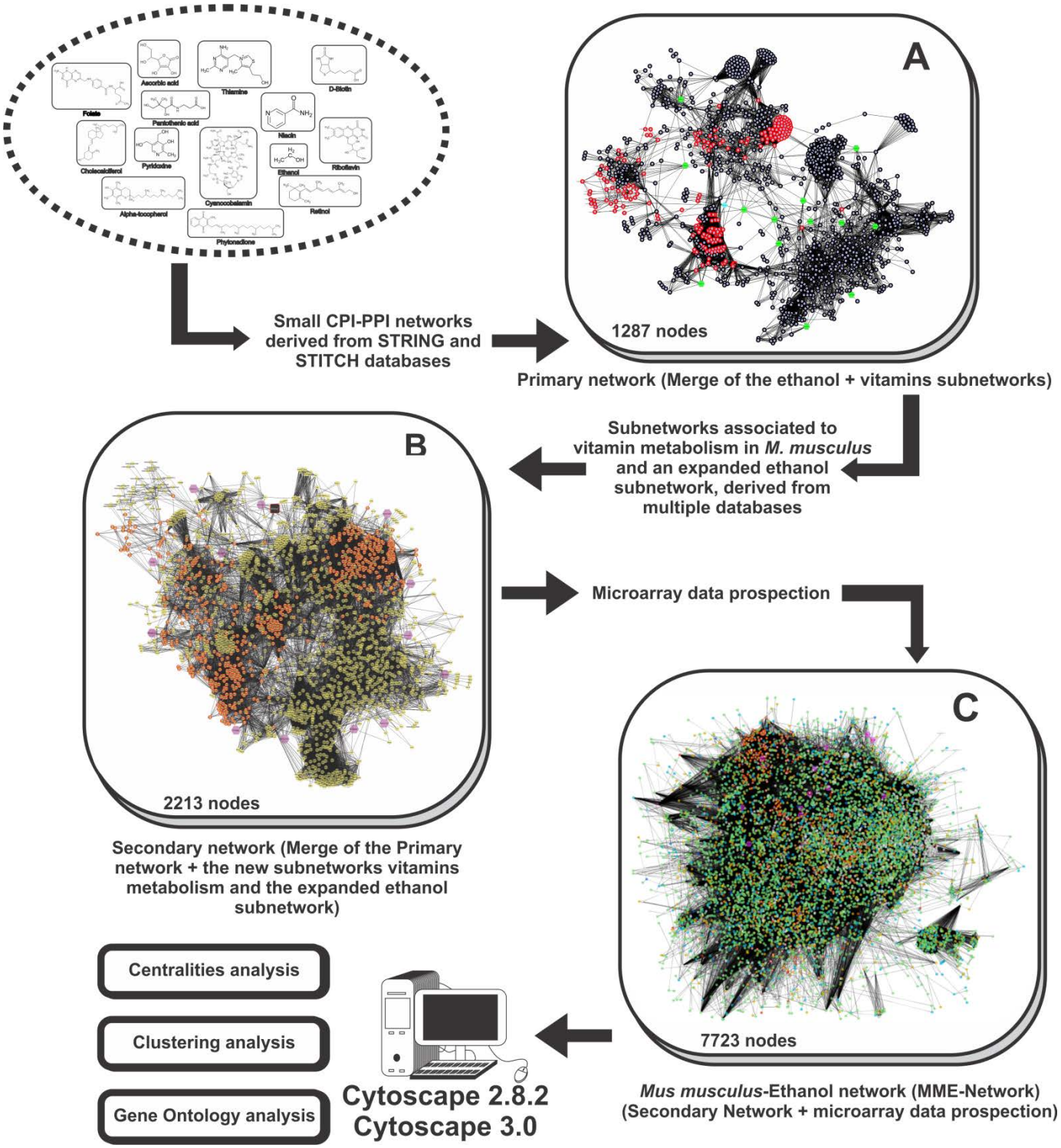
Figure 2. Network landscape analysis of microarray data from *Mus musculus* embryos exposed to ethanol (*M. musculus*-Ethanol Network – MME-Network). In (A), Prenatally Exposed (PE-) Network, derived from the transcriptomic analysis of prenatally ethanol-exposed embryonic brains from mice euthanized at E.16. B) Postnatal-Exposed Network (PSE-Network), derived from a transcriptomic analysis of brains of postnatal ethanol-exposed mice, euthanized at day 70. C) Early Gestation-Exposed-Postnatal-Network (EGEP-Network), derived from a transcriptomic analysis from the brains of prenatally ethanol-exposed embryos (E.4 and E.7), euthanized at postnatal day 60. D) Late Gestation-Exposed-Postnatal-Network (LGEP-Network), derived from a transcriptomic analysis from the brain of prenatally ethanol-exposed embryos (E.14 and E.16), euthanized at postnatal day 60.

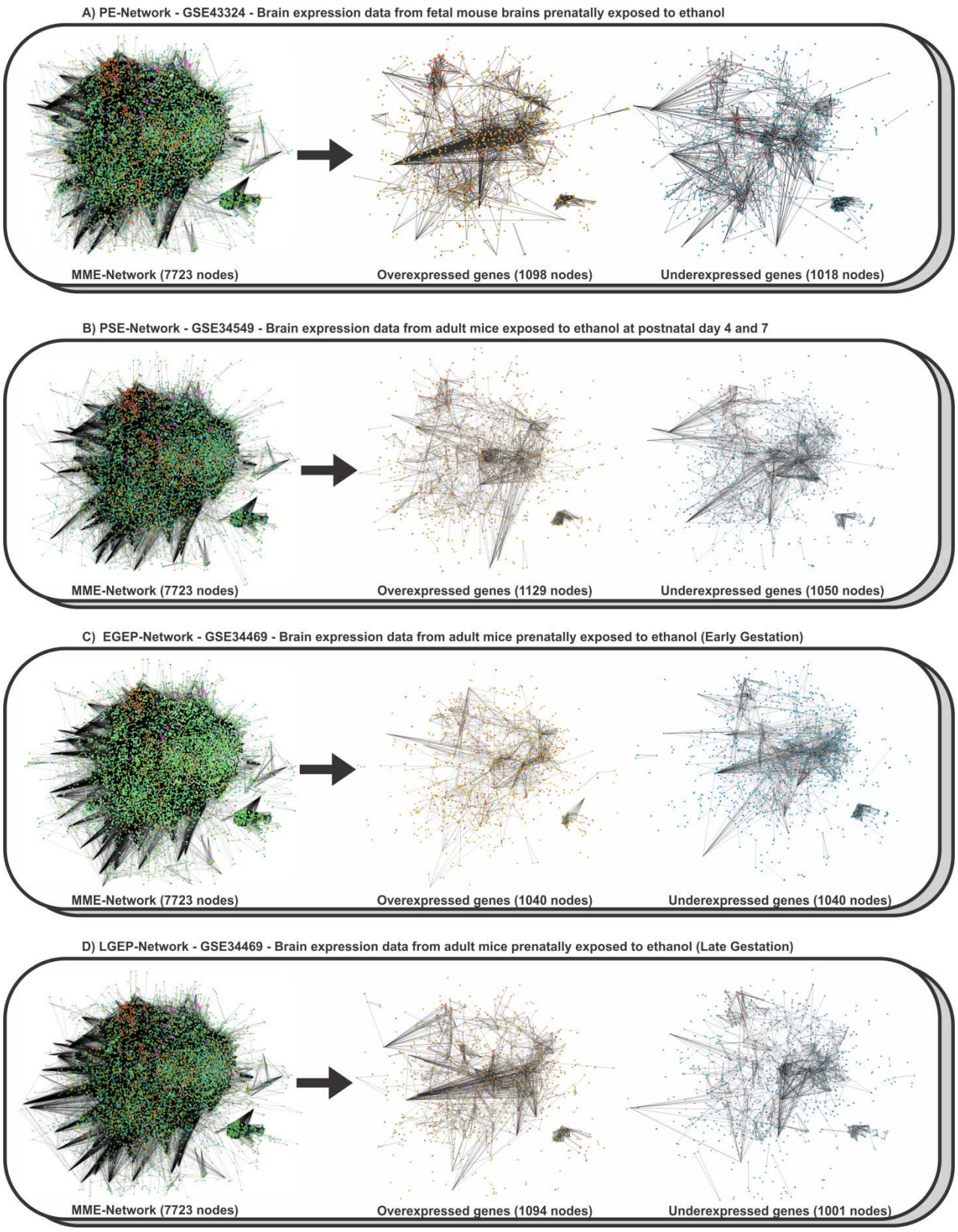
Figure 3. Venn diagrams created to observe the overlaps between the under- and overexpressed genes in all transcriptomic datasets. The green circle represents the Prenatally-Exposed (PE-Network, mice prenatally exposed to ethanol during E.14 and E.16 and euthanized at E.16). The yellow circle indicates the Postnatally-Exposed Network (PSE-Network, in which the dams were exposed to ethanol at postnatal days 4 and 7, and the offspring were euthanized at adult day 60). By its turns, the blue circle refers to the Late Gestation-Exposed-Postnatal-Network (LGEP-Network, in which the fetuses were exposed to ethanol during development at E.14 and E.16 and euthanized at adult day 70). The orange circle represents the Early Gestation-Exposed-Postnatal-Network (EGEP-Network, in which fetuses were exposed to ethanol at E.8 and E.11 and euthanized at adult day 70). Finally, the purple circle represents a fusion of the EGEP- and PE-Networks because they displayed the same underexpressed genes in the overlaps. A) Overlap of the underexpressed genes, which showed 19 genes in common (displayed in the table on the right side of the figure); B) Venn diagrams of the overexpressed genes of the PE-, EGEP- and LGEP-Networks, sharing 22 genes; C) Overexpressed genes overlapping among the PE-, PSE- and EGEP-Networks, which showed 34 shared nodes; D) Venn diagram showing the overlap between the overexpressed genes among the PE-, PSE- and LGEP-Networks, with 32 shared genes; E) Overlaps between the overexpressed genes in the LGEP-, PSE- and EGEP-Network, revealing 37 shared nodes. The common nodes among the Venn diagrams of B-E are also listed in the table on the right side of the figure.

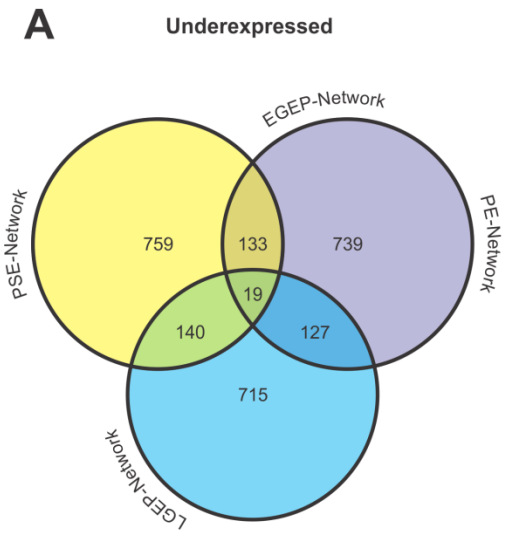
Figure 4. Subnetworks derived from hubs-bottleneck (HB) analysis. Nodes colored with red borders are those found in the *M. musculus*-Ethanol Network (MME-Network). In (A) *M. musculus* HBs; B) HB displaying the expression data from the Prenatally Exposed

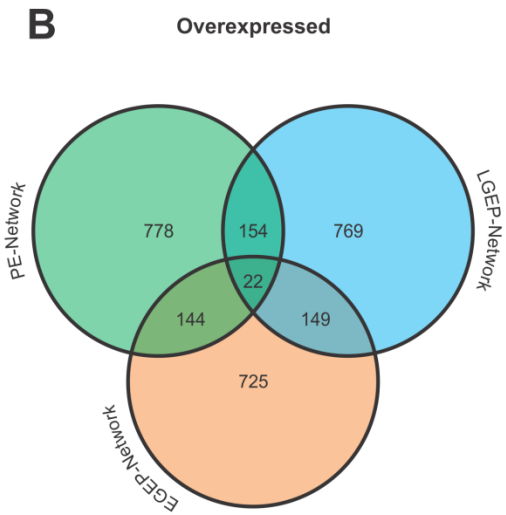
Network (PE-Network); C) HBs displaying the expression data from Postnatal-Exposed Network (PSE-Network); D) HBs displaying the expression data from the Early Gestation-Exposed-Postnatal-Network (EGEP-Network); E) HBs displaying the expression data from the Late Gestation-Exposed-Postnatal-Network (LGEP-Network). The Venn diagrams below each network display the overlaps between the indicated networks.

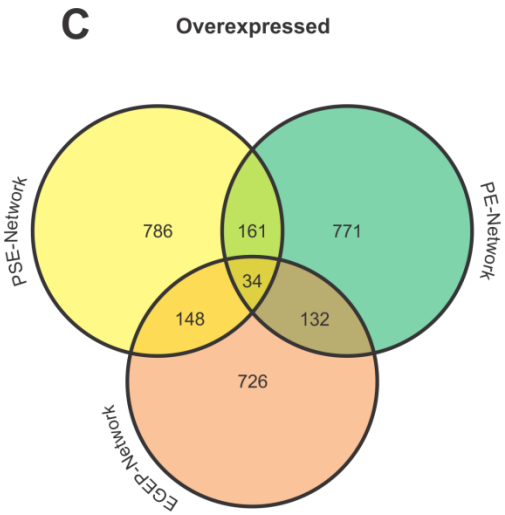
Figures

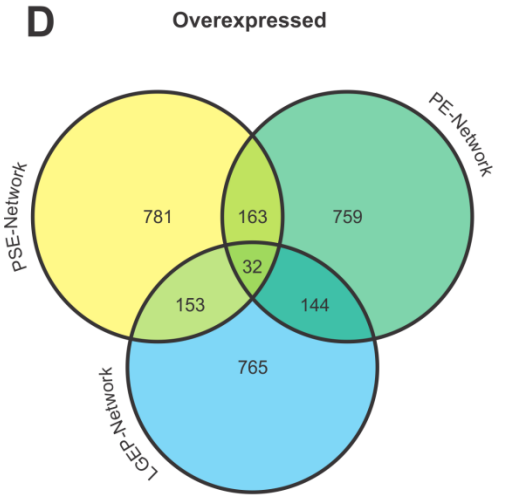

Ascorbic acid
Thiamine
D-Biotin
Folate
Pantothenic acid
Niacin
Cholecalciferol
Pyridoxine
Ethanol
Riboflavin
Alpha-tocopherol
Cyanocobalamin
Retinol
Phytonadione
Small CPI-PPI networks derived from STRING and STITCH databases
A
1287 nodes
Primary network (Merge of the ethanol + vitamins subnetworks)
Subnetworks associated to vitamin metabolism in M. musculus and an expanded ethanol subnetwork, derived from multiple databases
B
2213 nodes
Secondary network (Merge of the Primary network + the new subnetworks vitamins metabolism and the expanded ethanol subnetwork)
Microarray data prospection
C
7723 nodes
Mus musculus-Ethanol network (MME-Network)
(Secondary Network + microarray data prospection)
Centralities analysis
Clustering analysis
Gene Ontology analysis
Cytoscape 2.8.2
Cytoscape 3.0

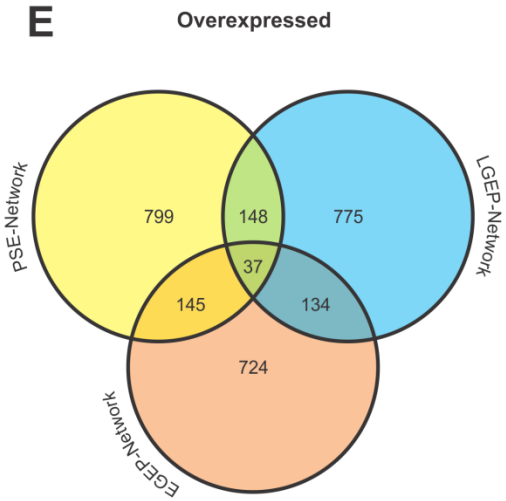

A) PE-Network - GSE43324 - Brain expression data from fetal mouse brains prenatally exposed to ethanol
MME-Network (7723 nodes)
Overexpressed genes (1098 nodes)
Underexpressed genes (1018 nodes)
B) PSE-Network - GSE34549 - Brain expression data from adult mice exposed to ethanol at postnatal day 4 and 7
MME-Network (7723 nodes)
Overexpressed genes (1129 nodes)
Underexpressed genes (1050 nodes)
C) EGEP-Network - GSE34469 - Brain expression data from adult mice prenatally exposed to ethanol (Early Gestation)
MME-Network (7723 nodes)
Overexpressed genes (1040 nodes)
Underexpressed genes (1040 nodes)
D) LGEP-Network - GSE34469 - Brain expression data from adult mice prenatally exposed to ethanol (Late Gestation)
MME-Network (7723 nodes)
Overexpressed genes (1094 nodes)
Underexpressed genes (1001 nodes)





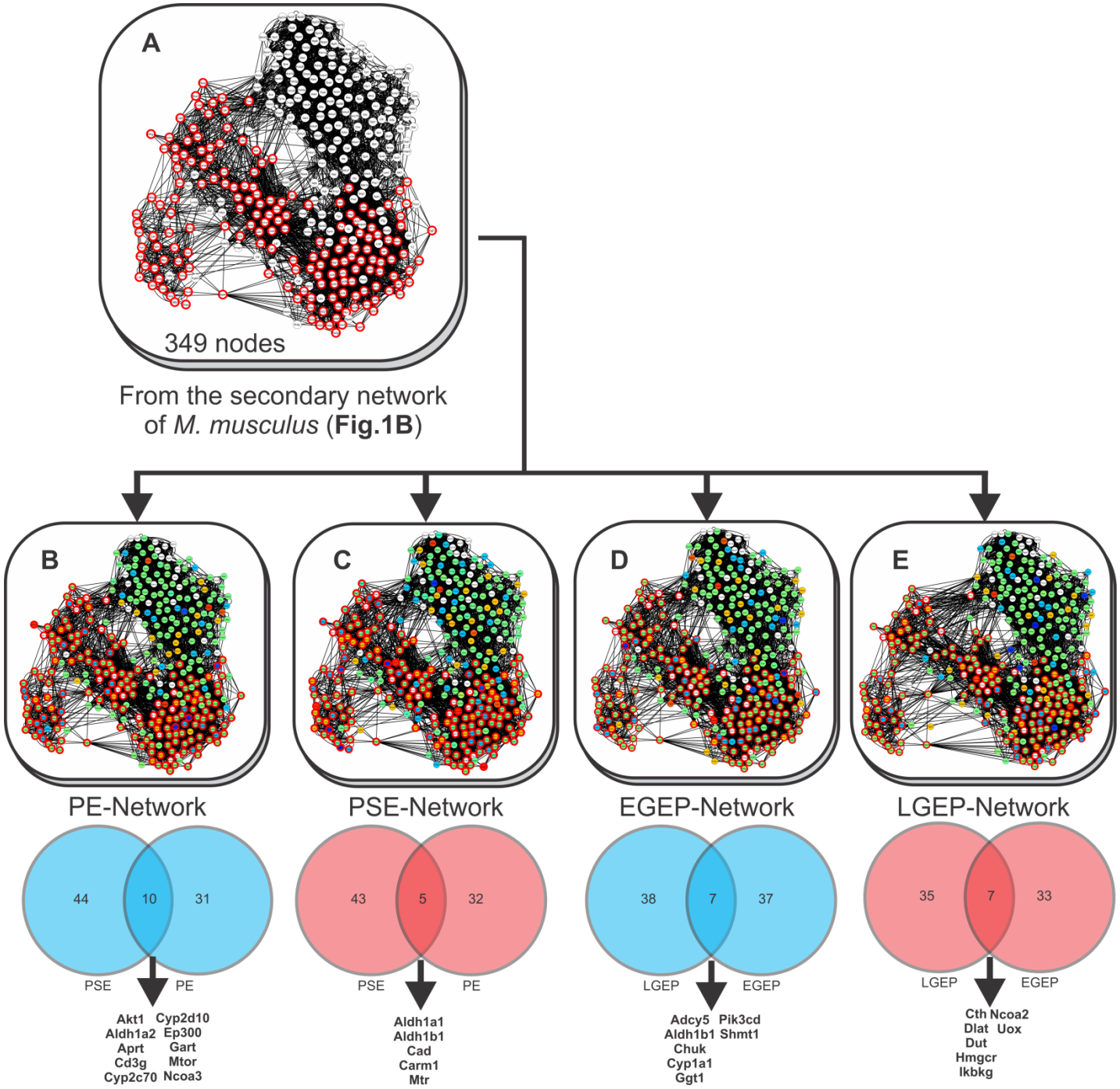






Overlaped

Underexpressed		Overexpressed
Akr1d1	Son	Tomm70a
Auh	Tmem11	Cdc25a
C1qtnf7	Tubb1	Rasa1
Coq6	Zfp622	Tbcb
Dbnl		Med6
Ep300		
Lpar2		
Mtor		
Mvd		
Nedd4		
Ogt		
Olfr15		
Olfr161		
Rpl32		
Sap30		


A
349 nodes
From the secondary network
of M. musculus (Fig.1B)
B
C
D
E
PE-Network
PSE-Network
EGEP-Network
LGEP-Network
44
10
31
PSE
PE
Akt1
Aldh1a2
Aprt
Cd3g
Cyp2c70
Cyp2d10
Ep300
Gart
Mtor
Ncoa3
43
5
32
PSE
PE
Aldh1a1
Aldh1b1
Cad
Carm1
Mtr
38
7
37
LGEP
EGEP
Adcy5
Aldh1b1
Chuk
Cyp1a1
Ggt1
Pik3cd
Shmt1
35
7
33
LGEP
EGEP
Cth
Dlat
Dut
Hmgcr
Ikbkg
Ncoa2
Uox

4. DISCUSSÃO GERAL

Entender a complexidade emergente das relações entre as diferentes biomoléculas e seu ambiente é um dos principais desafios das biociências. A criação de bancos de dados que armazenam informações sobre a identidade e funcionalidade de um gene e o seu produto, assim como o processo biológico em que estão associados, foi impulsionado pela era pós-genômica e permitiu que uma grande quantidade de dados biológicos fosse disponibilizada para os pesquisadores (Sobral, 1999).

Nesse caso, a análise de sistemas complexos e das interações entre biomoléculas por redes de interação e ferramentas de biologia de sistemas possibilitou o processamento desses grandes volumes de dados e posteriores análises utilizando algoritmos específicos para avaliar o que poderia estar ocorrendo a nível molecular. Assim, a busca de cenários que potencialmente se beneficiariam das análises de redes é de extrema importância para a pesquisa básica e aplicada. Neste sentido, foi prospectado, nos dois trabalhos desenvolvidos, como o álcool e as diferentes substâncias carcinogênicas presentes no tabaco poderiam afetar o desenvolvimento embrionário, pois o estudo do abuso dessas substâncias durante a gravidez não é claramente elucidado.

4.1. Outras considerações sobre a atuação dos compostos do tabaco no desenvolvimento embrionário

As redes de interação revelaram informações que são fundamentais para a compreensão dos mecanismos de atuação dos compostos do tabaco no desenvolvimento embrionário. Nesse sentido, é necessário aprofundar a discussão realizada no Capítulo 1 dessa dissertação de mestrado, onde se buscará analisar outros processos biológicos, tais como o metabolismo de lipídios e esteroides, bem como no metabolismo de DNA, que podem ser afetados por constituintes do tabaco (*Tobacco constituents* - TCs). As figuras listadas que são referentes ao artigo são devidamente apontadas no texto.

4.1.1 Biossíntese de esteroides e metabolismo de ácidos graxos insaturados

Um aspecto importante das ontologias gênicas observadas nos *clusters* presentes no **Capítulo I** (**Fig. 3A – 3F do artigo**) são aqueles relacionados à modificação de lipídios (isto é, modificações que resultam em alterações nas propriedades do lipídio, como modificações covalentes em um ou mais ácidos graxos) e esteroides (isto é, hormônios como androgênio, estrogênio e progesterona) (Feltes *et al*., 2013).

BOX3

StAR. Permite a clivagem do colesterol para pregnenolona através do seu transporte da membrana externa da mitocôndria para a membrana interna da mitocôndria (Bose *et al*., 2008). A ação de StAR esta ligada a um complexo de múltiplas proteínas, entre elas as proteínas *voltage-dependent anion chanel* (VDAC1), *translocator protein* (TSPO), *protein kinase A regulatory subunit 1*α (PKAR1A) e TSPO-*associated protein 7* (PAP7) Miller, 2013).

POMC. O gene de POMC codifica para um precursor de hormônio polipeptídico que é sintetizado na região anterior da pituitária (Dores & Baron, 2011). POMC pode sofrer uma série de modificações que resultam em diferentes produtos finais (Dores & Baron, 2011).

Foi mostrado que ratos expostos a um condensado líquido da fumaça do cigarro tiveram o desenvolvimento dos folículos ovarianos afetados, resultando numa diminuição da maturação dos ovócitos devido a deficiências nos níveis de estradiol (Sadeu & Foster, 2011). Como outros esteroides, o estradiol é sintetizado a partir do colesterol e um segundo hormônio, a pregnenolona, que é o precursor de todos outros esteroides, é produzindo na mitocôndria a partir do colesterol (Bose *et al*., 2008). O mesmo estudo indicou que o cigarro causa uma diminuição no transporte de colesterol para a matriz mitocondrial, pois interfere com a proteína transportadora de colesterol StAR *steroidogenic acute regulatory protein* - proteína reguladora aguda, esteroidogênica), que se localiza na membrana externa da mitocôndria e é essencial para o transporte de colesterol (Bose *et al*., 2008). A proteína StAR não estava presente na rede-CPI-PPI (*Chemical-Protein Interaction-Protein-Protein Interaction* – Interação Químico-Proteína-Interação Proteína-Proteína) (**Fig. 1C do artigo**), mas ampliando os dados de interatoma da rede nós verificamos que ela se conectava a um dos nossos *clusters*. De fato, StAR estava conectada a POMC (**Fig. 3B e 3F do artigo**), um receptor de melanocorticóide relacionado a uma ampla gama de processos, incluindo inflamação e esteroidogênese (Dores & Baron, 2011). POMC foi observada por estar expressa tanto no feto quanto no embrião (**S-*Table* 1** (Tabela S1), ver **Material Suplementar 1 do artigo**).

Outro TC associado com a esteroidogênese no nosso módulo (**Fig. 3C do artigo**) é o isopreno que, por sua vez, está conectado às proteínas FDPS e FDFT1 (**Fig. 3C, do artigo**). A enzima FDFT1 atua na primeira fase da síntese de colesterol, dimerizando farnesil-difosfato para formar esqualeno (Dhar *et al*., 2013), enquanto FDPS catalisa a produção de geranil-pirofosfato em farnesil-pirofosfato, um intermediário na biossíntese de esterol e colesterol (Trapani *et al*., 2011). Ambas as proteínas estão presentes no feto e no embrião (**S-*Table* 1** (Tabela S1), ver **Material Suplementar 1 do artigo**).

BOX4

FDPS. Essa proteína não apenas atua na biossíntese de colesterol, mas também uma ampla gama de isoprenóides como quinonas, ubiquinonas, menaquinonas e plastoquinonas (Dhar *et al*., 2013).

FDFT1. Essa enzima converte o produto formado pela FDPS em esqualeno, uma molécula que dará origem ao colesterol (Trapani *et al*., 2011).

UGT. Essas proteínas são responsáveis pela glucoronidação, ou seja, a adição de ácido glicorônico, em xenobióticos e outros esteroides, deixando-os mais solúveis em meio aquoso (King *et al*., 2000). As proteínas UGT também destoxificam substâncias lipofílicas tóxicas (Sugatani, 2013).

A exposição ao tabaco durante a gravidez esta relacionada a defeitos no desenvolvimento dos testículos (Fowler *et al*., 2008). Neste sentido, o composto *N*-nitrosoanabasina esta conectado ao cluster de proteínas UGT (**Fig. 3A** e **3F do artigo**). As proteínas UGT são responsáveis por catalisar a glicuronidação de estrogênios e androgênios, tornando-os mais solúveis e facilitando seu transporte intracelular. Esse processo pode ser crítico durante o desenvolvimento para a exposição adequada do feto aos hormônios sexuais maternos, especialmente durante a determinação do sexo e da formação de padrões corporais.

Em resumo, os dados de biologia de sistemas mostram que os TCs afetam negativamente a síntese de colesterol, atuando em proteínas necessárias para síntese do mesmo, como FDPS e FDFT1. A ação de moléculas lipofílicas na mitocôndria foram observadas em estudos prévios (van der Toorn *et al*., 2009) e compostos hidrofílicos também poderiam afetar a função mitocondrial. De fato em nossos módulos é possível observar a interação de compostos hidrofílicos com as proteínas relacionadas ao metabolismo e biossíntese de lipídios e esteroides (**Fig. 11A**).

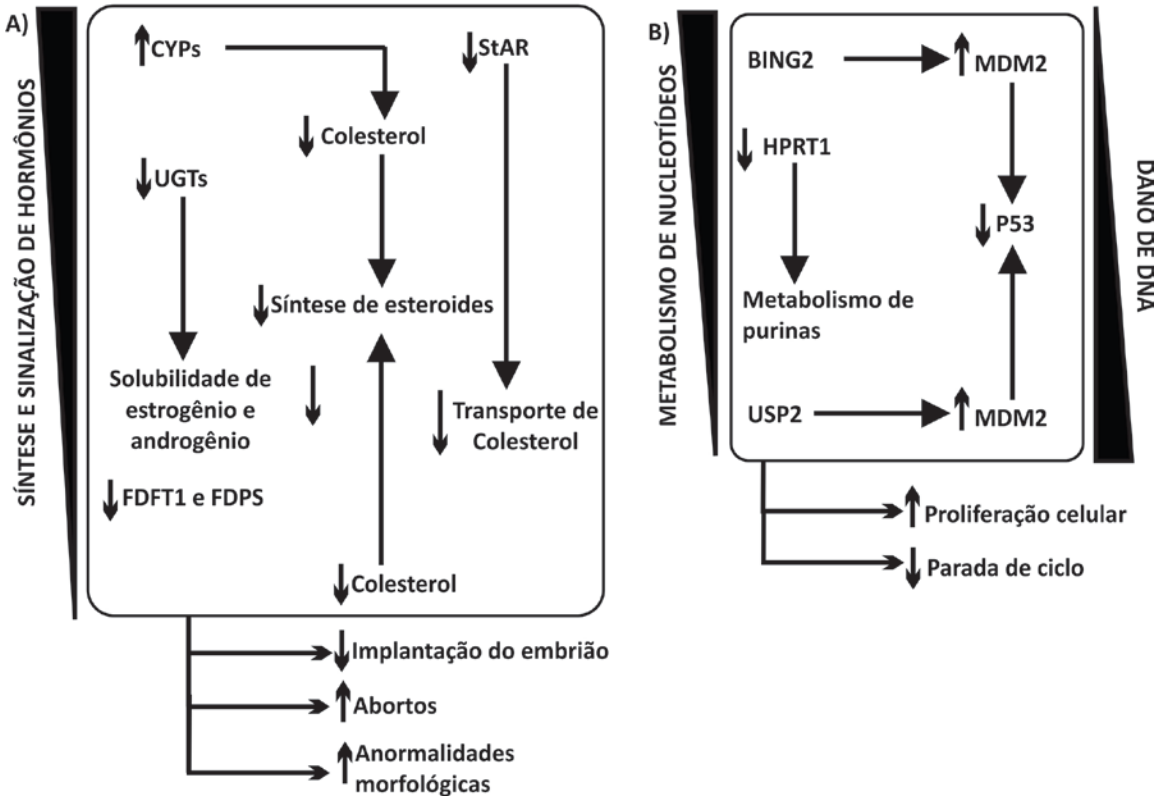


Figura 11. Modelo de interação mostrando as consequências da exposição cigarro baseado nas nossas redes de interação. Em **(A)** pode ser observado que os TCs provocam uma baixa síntese e sinalização de hormônios. Para tanto, os TCS podem afetar negativamente a solubilidade de estrogênio e androgênio, atuando no *cluster* de proteínas UGT. Eles também estão associados com uma baixa síntese de colesterol, aumentando os níveis de CYP e diminuindo os níveis de FDFT1 e FDPS, duas enzimas responsáveis pela síntese do colesterol. Baixa disponibilidade de colesterol diminuiria a síntese global de hormônios na mitocôndria. Ademais, os TCs afetariam o transporte de colesterol para mitocôndria, pois afetam a proteína de membrana StAR. Esses fatores culminam em problemas na implantação do embrião, maior probabilidade de abortos e anormalidades morfológicas. Em **(B)** TCs também estão ligados a BING2 e USP2, proteínas que são responsáveis pelo aumento da atividade de MDM2. Esse aumento de atividade de MDM2, rapidamente pode regular negativamente a proteína p53, levando à uma maior susceptibilidade a dano de DNA. Em adição eles podem diminuir a síntese de purinas através da sua ação em HPRT1. Isso levaria a um amento da proliferação celular e diminuição da eficiência na parada de ciclo.

4.1.2. Metabolismo e reparo de DNA

Na analise de ontologias gênicas dos *clusters* 3, 6, 17 e 21 (**Fig. 5 do artigo**), nós identificamos dois processos relacionados: (i) *splicing* de RNA e (ii) metabolismo de DNA (**S-*Tables* 6, 9, 20** e **24** (Tabelas S6, 9, 20 e 24), ver **Material Suplementar 1 do artigo**).

As proteínas do citocromo-P450 (CYP) estão presentes nos módulos (**Fig. 3A do artigo**) e são responsáveis pelo metabolismo de xenobióticos e pela ativação de compostos carcinogênicos, como PAHs e aminas aromáticas (Pliarchopoulou *et al*., 2012). Enquanto as CYPs convertem estes compostos em carcinógenos reativos, as glutaminas S-transferases (GSTs) (**Fig. 5B do artigo**) destoxificam PAHs carcinógenos (Pliarchopoulou *et al*., 2012).

BOX5

CYPs. As proteínas dessa família estão envolvidas na biossíntese e metabolismo de diversas biomoléculas, como uma ampla gama de produtos derivados de colesterol, vitamina D_3, pregnenolona e ácidos biliais (Pikuleva & Waterman, 2013).

GSTs. A maior parte das GSTs adicionam glutationa reduzida (GDH) em eletrófilos (moléculas que aceitam elétrons) (Strange *et al*., 2001). GSTs também estão relacionadas na destoxificação de produtos resultantes da alteração do estado redox (por exemplo, ânion superóxido) (Hayes *et al*., 2005).

HPRT1. Converte hipoxantina, um composto derivado do ácido úrico, e guanina em inosina monofosfato e guanosina monofosfato (Chang *et al*., 2005).

Ademais, a proteína GSTM1 foi observada nas nossas análises. GSTM1 é uma proteína importante na destoxificação de PAHs e um estudo mostrou que uma substituição de adenina por uma guanina no gene de GSTP1 que gera uma valina ao invés de uma isoleucina ($Ile^{105}Val$), leva a perda da capacidade de detoxificação da proteína resultante (Pliarchopoulou *et al*., 2012).

Essa perda da capacidade de destoxificação de GSTs pode ser causada pela ação combinada dos TCs. Na nossa análise de ontologias gênicas, não apenas os TCs estão associados com o metabolismo de nucleotídeos em quatro *clusters* diferentes, mas em cada cluster continha compostos exclusivos, incluindo hidrofílicos (catecol) (**Fig. 5C** e **5E do artigo**) e lipofílicos (criseno, 1,3-butadieno, crotonaldeído) (**Fig. 5B do artigo**), assim como orgânicos (criseno e 1,3-butadieno) (**Fig. 5B do artigo**) e inorgânicos (berílio, polônio-210 e arsênio) (**Fig. 5A**, **5C**, **5D** e **5E do artigo**). Notavelmente, no *cluster* 6, ambas substâncias (criseno e 1,3-butadieno) estão ligados a HPRT1 (**Fig. 5B do artigo**), uma hipoxantina fosforibosiltransferase que é responsável

pelo metabolismo de purinas. Corroborando com nossas hipóteses, um estudo mostra que fumantes possuem menores níveis de HPRT1 (Chang *et al*., 2005).

As ferramentas de biologia de sistemas aplicadas também identificaram ITPA no *cluster* 6 (**Fig. 5B do artigo**). ITPA é importante para a remoção de purinas desaminadas em mamíferos (Sakumi *et al*., 2010). A mutação ITPA$^{-/-}$ mostrou ser letal durante o desenvolvimento perinatal e fibroblastos embrionários de camundongo ITPA$^{-/-}$ exibiram aumento na taxa de geração de anormalidades cromossomais e acúmulo de danos de DNA do tipo quebra simples (Abolhassani *et al*., 2010). Portanto, dado o papel de ITPA na letalidade embrionária e na geração de aberrações cromossômicas, ITPA é passível de ser um bom alvo para entender a ação dos TCs no metabolismo e reparo de DNA.

Ademais, o 1,3-butadieno foi observado por estar ligado ao aumento de estresse genotóxico devido ao dano de DNA através da formação de *pontes intercadeias* entre adenina e guanina (Goggin *et al*., 2011; Koturbash *et al*., 2011). Este composto também já foi associado com efeitos epigenotóxicos causados pela perda global da metilação de DNA e trimetilação da histona H3 nos resíduos de lisina 9 e 272 e histona H4 lisina 20; todas conhecidas por regular os padrões da expressão gênica (Koturbash *et al*., 2011).

BOX6

ITPA. É responsável pela hidrolise ITP em IMP, uma molécula essencial para a biossíntese de purinas e precursor do AMP e GMP (Lin *et al*., 2001).

PML. também possui um importante papel na supressão tumoral, atuando como ativador da proteína de parada de ciclo CHK2, a proteína cinase inibidora de ciclina p21 e da proteína p53 (Martin-Martin *et al*., 2013). PML é alvo de diferentes modificações pós-traducionais, tais como fosforilação, sumoilação e ubiquitinação (Martin-Martin *et al*., 2013).

Essas associações revelam que as substâncias relacionadas ao tabaco podem afetar caminhos distintos do metabolismo de DNA, como metabolismo de purinas e mutações nos genes que causam detoxificação de compostos tóxicos ao organismo.

Na análise de ontologias do cluster 3, 8 e 9 (**Fig. 6 do artigo**), nós identificamos os processos de: (i) estímulo ao dano de DNA e (ii) ciclo celular (**S-*Tables*** 8, 11 e 12 (Tabelas S8, 11 e 12) no **Material Suplementar 1 do artigo**) .

Nesses *clusters* foi observado que o arsênio se liga diretamente à proteína PLM (**Fig. 6B** e **6D do artigo**), que é uma proteína relacionada a organização da cromatina, diferenciação, reparo de DNA e modificações pós-traducionais

[http://www.genecards.org]. Em nossos *clusters*, PLM é ligada a proteínas como p53 e BING2 (DAXX) (**Fig. 6B** e **6D do artigo**). BING2 é uma proteína que se transloca entre o núcleo e o citoplasma e é descrita como ser ativadora da atividade ubquitinadora de MDM2. Levando em consideração que MDM2 é fortemente ligada a degradação de p53 (Nag *et al*., 2013), os TCs podem possuir um papel na degradação mediada por proteassomas de p53, levando a uma parada de ciclo ineficiente e, consequentemente, dano ao DNA. MDM2 se encontra nos *clusters* 5 e 8 (**Fig. 6A** e **6B do artigo**). Essas relações são suportadas pela alta incidência de cânceres em pessoas fumantes.

Essa proteína também é ligada a ativação de JNK1 (Salomoni, 2013) que, como discutida previamente no Capítulo I, pode estar envolvida na regulação negativa de RARα.

BOX7

BING2/DAXX. A perda de DAXX é relacionada à letalidade embrionária (Salomoni, 2013). Essa proteína também funciona como uma chaperona para a histona H3, mostrando sua ligação direta com organização da cromatina (Salomoni, 2013).

USP2. Promove a desubiquitinação de MDM2, estabilizando-a e promovendo a degradação de p53 (Stevenson *et al*., 2007). A inibição dessa proteína leva à ativação de p53 *in vivo* (Stevenson *et al*., 2007).

Da mesma forma, verificamos que etilamina estava conectada à USP2 (**Fig. 6A** e **6D do artigo**), uma proteína relacionada a diferenciação miogênica durante a embriogênese, e é indiretamente associada com a degradação de p53 por MDM2 [http://www.genecards.org]. Um aumento na degradação de p53 afetaria a parada de ciclo em G_1/S e aumentaria a probabilidade de danos ao DNA no feto. Neste sentido, em tecido pulmonar, ratos expostos à fumaça do cigarro, e que carregavam mutações no gene de p53, mostraram ineficiência na indução de genes pró-apoptóticos, e um aumento na expressão de genes relacionados à proliferação celular, resposta imune e informação (Izzotti *et al*., 2004). Evidência também sugerem que inflamação crônica causada por intoxicantes ambientais é ligada a diferentes aspectos a tumorigênese e ao dano de DNA (Kamp *et al*., 2011), que se torna consistente com as conexões observadas entre TCs e p53 (**Fig. 11B**).

4.2. Outras considerações sobre a atuação do etanol no neurodesenvolvimento e na progressão da SAF

No artigo do capítulo II nos focamos nos genes alterados em comum entre todos os grupos de dados transcriptômicos, visando estudar quais genes seriam os melhores alvos para o entendimento da SAF e como o etanol poderia afetar o metabolismo de vitaminas, afetando o neurodesenvolvimento (**Figura 12**). Contudo, algumas alterações na expressão em diferentes dados transcriptômicos se sobressaem e torna-se possível observar outras considerações em relação ao etanol e sua atuação no neurodesenvolvimento, basicamente em alterações pontuais encontradas em cada rede.

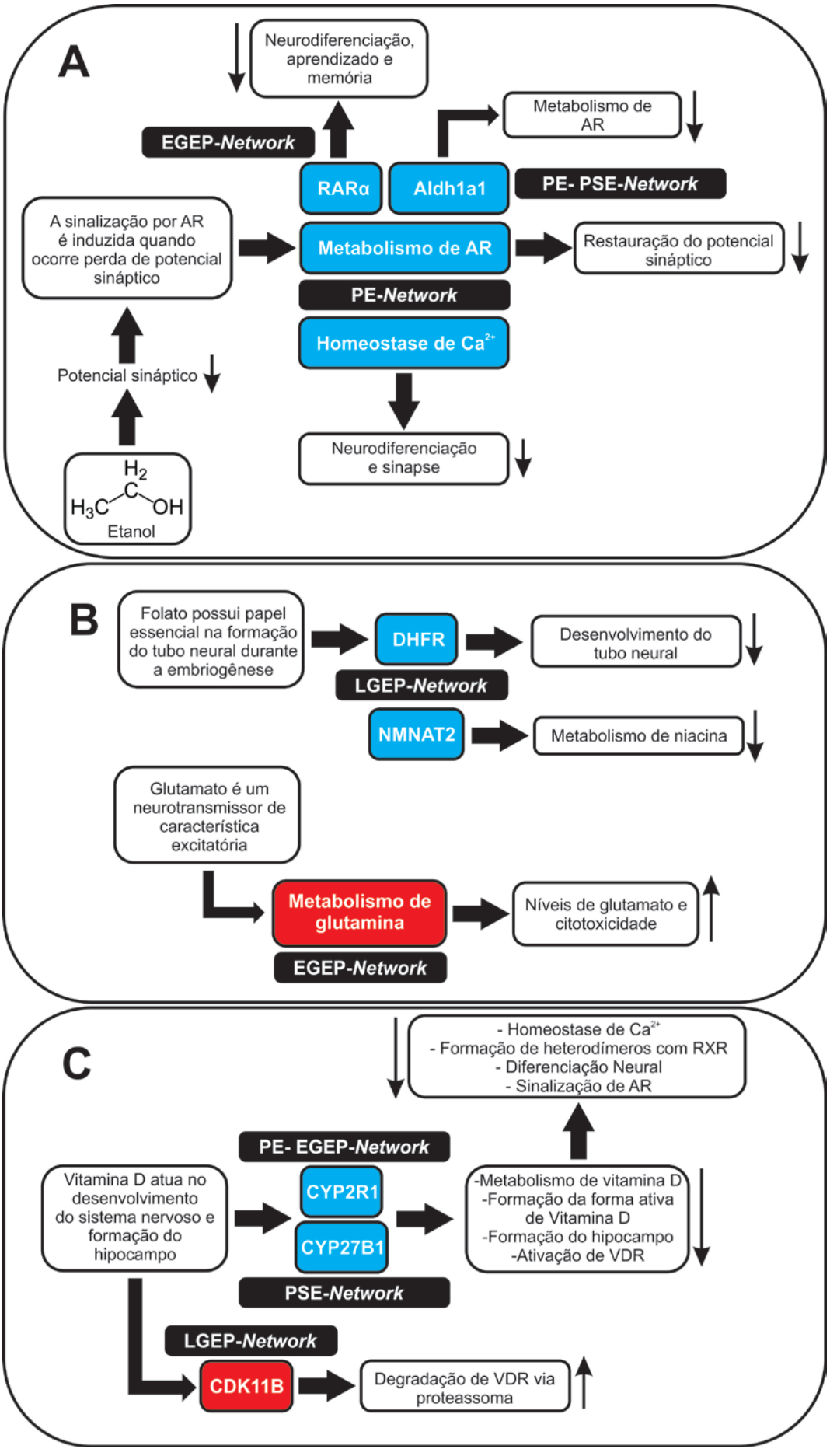

A
Neurodiferenciação, aprendizado e memória
EGEP-Network
Metabolismo de AR
RARα
Aldh1a1
PE- PSE-Network
A sinalização por AR é induzida quando ocorre perda de potencial sináptico
Metabolismo de AR
Restauração do potencial sináptico
PE-Network
Homeostase de Ca2+
Potencial sináptico
Neurodiferenciação e sinapse
H2
H3C
C
OH
Etanol
B
Folato possui papel essencial na formação do tubo neural durante a embriogênese
DHFR
Desenvolvimento do tubo neural
LGEP-Network
NMNAT2
Metabolismo de niacina
Glutamato é um neurotransmissor de característica excitatória
Metabolismo de glutamina
Níveis de glutamato e citotoxicidade
EGEP-Network
C
- Homeostase de Ca2+
- Formação de heterodímeros com RXR
- Diferenciação Neural
- Sinalização de AR
PE- EGEP-Network
Vitamina D atua no desenvolvimento do sistema nervoso e formação do hipocampo
CYP2R1
CYP27B1
-Metabolismo de vitamina D
-Formação da forma ativa de Vitamina D
-Formação do hipocampo
-Ativação de VDR
PSE-Network
LGEP-Network
CDK11B
Degradação de VDR via proteassoma

Figura 12. Resumo da atuação do etanol no neurodesenvolvimento de acordo com os dados do Capítulo II. Os retângulos azuis indicam bioprocessos cujos genes associados estavam com baixa expressão nos dados analisados. Já os quadrados vermelhos indicam bioprocessos cujos genes associados estavam com superexpressão. Os retângulos pretos indicam as redes em que esses bioprocessos foram encontrados. Em **(A)** está o resumo da atuação do etanol no metabolismo de ácido retinóico (AR) e na homeostase de cálcio. Em **(B)** se encontra o resumo da atuação do etanol no metabolismo de folato e niacina. Por fim, em **(C)** está o resumo da atuação do etanol na degradação de vitamina D.

4.2.1. Gas7

Na PE-*Network* (onde ratas grávidas foram expostas ao etanol nos dias embrionários 14 (E.14) e 16 (E.16), e o feto foi eutanasiado no dia E.16) foi observado Gas7 entre os genes com baixa expressão (**S-*Table* 2** (Tabela S2), ver **Material Suplementar 2**). Gas7 e suas variantes (hGas7-a que é homóloga de Gas7-cb de ratos, e hGas7-b) é uma proteína expressa abundantemente em tecido cerebral, fundamental para o crescimento dos lamelipodios e filopodios em cultura celular de cerebelo (Lazakovitch *et al*., 1999; Chao *et al*., 2005) e vistos porém ser necessária para o crescimentos de neuritos em neurônios de hipocampo (You & Lin-Chao, 2010). Da mesma forma, ela já foi discutida por potencializar a diferenciação neural e ser um potencial alvo para a reparação de danos cerebrais (Lortie *et al*., 2005). É interessante ressaltar que no LGEP-*Network* (onde ratas grávidas foram expostas ao etanol nos dias E.14 e E.16, e o indivíduo adulto foi eutanasiado pós-natal no dia 60) Gas7 ainda é vista com baixa expressão (**Fig. 2D, ver artigo**; **S-*Table* 2** (Tabela S2), ver **Material Suplementar 2**) ao contrário do EGEP-*Network* (onde ratas grávidas foram expostas ao etanol nos dias E.8 e E.11, e o indivíduo adulto foi eutanasiado pós-natal no dia 60) (**Fig. 2C, ver artigo**; **S-*Table* 2** (Tabela S2), ver **Material Suplementar 2**). Essa observação indica que a baixa expressão de Gas7 causada pelo etanol só causa danos à longo prazo no tecido

BOX8

Gas7. Ratos envelhecidos Gas7-deficientes mostraram perda de habilidades motoras devido à uma diminuição na quantidade de neurônios na coluna vertebral, e mudanças na composição das fibras musculares, causando perda de força (Huang *et al*., 2012). Inibição desta proteína também está ligada a uma diminuição da osteogênese e mineralização óssea (Chao *et al*., 2013).

cerebral caso o feto seja exposto ao etanol durante os estágios mais avançados da gestação. O fato de que Gas7 é realmente afetada durante o desenvolvimento é que no PSE-*Network* (onde ratos adultos foram expostos ao etanol nos dias 4 e 7, e o indivíduo foi eutanasiado no dia 70) (**Fig. 2B, ver artigo**; **S-*Table* 2** (Tabela S2), ver **Material Suplementar 2**) Gas7 se encontra não-diferencialmente expressa.

Como discutido previamente, SAF é uma doença amplamente conhecida por gerar problemas cognitivos sociais, motores e de aprendizado que podem se amplificar progressivamente (O'Leary, 2004) e a atuação de Gas7 pode ser um importante alvo para os estudos da ação do etanol na progressão da patologia.

BOX9

FGF2. Um estudo reporta que a expressão anormal de Fgf2 em E.11.5 (antes da neurogênese) induz a deformações no tamanho córtex e neocortex (Rash *et al*., 2013).

FGF3. Em galinhas, FGF3 esta envolvida na expressão de marcadores de desenvolvimento do rombencéfalo (Weisinger *et al*., 2012). Em paulistinha e galinha, essa proteína também está envolvida com o enervamento e vascularização da neurohipófise (Liu *et al*., 2013).

FGF4. Sua expressão é relacionada à diferenciação de células-tronco embrionárias de ratos em linhagens neurais (Huang *et al*., 2010).

FGF8. Fgf8 foi demonstrada por ser essencial para a orientação da formação da cápsula óptica e da estrutura frontonasal (Griffin *et al*., 2013)

FGF9. Envolvida na osteogênese e angiogênse do tecido ósseo (Kizhner *et al*., 2011; Behr *et al*., 2010)

FGF10. Assim como FGF3, em paulistinha e galinha, essa proteína está envolvida com o enervamento e vascularização da neurohipófise (Liu *et al*., 2013).

4.2.2. Família Fgf

Dentre os genes observados por estarem com baixa expressão em todos os conjuntos de dados destaca-se a família fatores de crescimento de fibroblastos (*Fibroblast growth factor* – Fgf) (**S-*Table* 2** (Tabela S2), ver **Material Suplementar 2**). Consistente com nossos dados, as proteínas Fgf já foram vistas por serem afetadas pelo etanol em *Danio rerio* (isto é, paulistinha ou *zebrafish*), onde Fgf2 e Fgf8 já foram vistos por terem seus mRNAs alterados e Fgf19 (Fgf15 em *M. musculus*)/8 foram relacionados a defeitos na sinalização GABAérgicas nas regiões prosencefálicas e no cerebelo (Zhang *et al*., 2013). Interessantemente, genes da família Fgf, como Fgf2 e Fgf14, foram estudados por estarem com alta expressão durante o aprendizado de ratos (Cheung *et al*., 2013), sendo que deleção do gene Fgf14 já foi vista por afetar a transmissão sináptica nos neurônios de Purkinje do cerebelo (Xiao *et al*., 2013). Isso se torna um aspecto importante para ser observado, uma vez que pacientes com SAF mostram problemas na coordenação motora e no aprendizado ao longo da vida (O'Leary, 2004). Ademais, a expressão de Fgf8 é vista durante o

desenvolvimento do SNC e (Kataoka & Shimogori, 2008) defeitos no gene de *Fgf8* causam diversas anomalias faciais em ratos (Macatee *et al*., 2003). Desta forma, qualquer disfunção incidente na sinalização de Fgf pode se tornar crucial para o desenvolvimento de SAF.

BOX10

FGF14. Fgf14 está relacionada a diminuição da excitabilidade em neurônios do hipocampo, onde uma mutação que altera o gene de Fgf14 causa uma diminuição a atividade de canais de voltagem-dependentes de Na^{+} (Laezza et al., 2007). Da mesma forma, essa proteína regula membros da família de canais de Ca^{++} (CaV2), influenciando impulsos pré-sinápticos (Yan *et al*., 2013).

FGF17. Envolvida na formação do mesencéfalo (Zanni, *et al*., 2011). Expressão reduzida de FGF17 em ratos é relacionada anormalidades na vermis cerebelar (Zanni, *et al*., 2011). Interessantemente, ratos $FGF17^{-/-}$ mostram diminuição no volume cerebelo e do córtex dorso-frontal, assim como déficits de aprendizado social (Terwisscha van Scheltinga *et al*., 2013).

FGF18. Envolvida com diferenciação osteogênica e na condrogênese (Hague *et al*., 2007; Nagayama *et al*., 2013). Em adição FGF18 é expressa no telencéfalo de ratos (Borello *et al*., 2008).

FGF19 (FGF15 em *M. musculus***).** Essa proteína também está envolvida no controle da homeostase de ácido biliar hepático, inibindo sua síntese (Kir *et al*., 2011). Envolvida na síntese de glicogênio (Kir *et al*., 2011).

Nossas redes de interação também revelaram outros genes com baixa expressão da família Fgf. Neste sentido, na rede onde ratas grávidas foram expostas ao etanol nos dias E.14 e E.16 e os fetos foram eutanasiados em E.16 (PE-*Network*; **Fig.2A, ver artigo**) e na rede em que ratas grávidas foram expostas ao etanol nos dias E.14 e E.16 e o indivíduo adulto foi eutanasiado no dia 60 (LGEP-*Network*; **Fig.2D**, **ver artigo**), foram as que mostraram o maior número de membros da família Fgf com baixa expressão, como Fgf4, 9, 10, 15, 17 e 18. Já na EGEP-*Network*, apenas Fgf3 e Fgfr1op foram observados (**Fig.2C, ver artigo**), enquanto na PSE-*Network*, apenas os receptores de Fgf, Fgfr2, 3 e 18, estavam presentes (**Fig.2B**, **ver artigo**).

Esses resultados sugerem um modelo onde os membros da família Fgf são criticamente afetados nos estágios mais tardios do desenvolvimento e os efeitos negativos são mais brandos nos estágios iniciais. Contudo, vale ressaltar que os resultados da PE- e LGEP-*Networks* foram os mesmos, mostrando que os efeitos do etanol nas proteínas Fgf em questão começam no desenvolvimento e se estendem para a idade adulta. Da mesma forma, as redes permitiram a identificação de novos membros de Fgf para serem levados em consideração no entendimento de SAF.

4.2.3. *Família Hox*

Outro importante resultado obtido pela análise de biologia de sistemas foi a observação de proteínas da família Hox ente os genes com baixa expressão nas PE-, EGEP- e LGEP-*Networks* (**S-*Table* 2** (Tabela S2), ver **Material Suplementar 2**).

BOX11

HOXD8. Um estudo indica que esta proteína esta envolvida com o desenvolvimento das vértebras (van der Akker *et al*., 2001).

HOXD9. Participa na formação dos membros anteriores (Xu & Wellik, 2011), e do sistema esquelético (Gersch *et al*., 2005).

HOXD10. Expressa na região lombar da medula espinhal e envolvida na formação de enervações motoras (Choe *et al*., 2006).

HOXD11. Envolvida na diferenciação de condrócitos (Gross *et al*., 2012). Mutações no gene de HOXD11 causa infertilidade em machos e anomalias no sistema esquelético (Boulet & Capecchi, 2002).

HOXD12. Foi estudado que HOXD12 está envolvida com ossificação, formação do sistema esquelético e dígitos em ratos (Cho *et al.*, 2008; Villavicencio-Lorini *et al*., 2010).

Os genes Hox, em mamíferos, são uma família especial de genes agrupados em quatro *clusters* distintos (HoxA-D), localizados nos cromossomos 7, 17, 12, 2, respectivamente em humanos e 6, 11, 15 e 2, respectivamente em ratos (Daftary & Taylor, 2006). Eles codificam para proteínas que controlam, de uma maneira espaço-temporal, o estabelecimento do eixo anteroposterior durante o desenvolvimento embrionário, sendo responsáveis pela formação da estrutura bilateral e regionalização dos membros e órgãos em formação (Daftary & Taylor, 2006; Favier & Dollé, 1997; Martinez & Amemyia, 2002). Da mesma forma eles são responsáveis pela transcrição de diversos genes, incluindo genes relacionados a diferenciação celular (Daftary & Taylor, 2006; Ladam & Sagerström, 2013; Martinez & Amemyia, 2002). Neste sentido, genes Hox são fundamentais para o desenvolvimento da crista neural, assim como a região rombencefálica (Favier & Dollé, 1997).

O interatoma da PE-*Network* mostrou quatro genes do *cluster* HoxD com baixa expressão, Hoxd8, Hoxd9, Hoxd11 e Hoxd12 (**S-*Table* 2** (Tabela S2), ver **Material Suplementar 2**), enquanto a na LGEP-*Network* foi observado dois do *cluster* HoxD (Hoxd10 e Hoxd12) e dois do *cluster* HoxA (Hoxa10 e Hoxa11) (**S-*Table* 2** (Tabela S2), ver **Material Suplementar 2**). Por fim, a EGEP-*Network* mostrou apenas um gene Hox com baixa expressão, Hoxa11 (**S-*Table* 2** (Tabela S2), ver **Material Suplementar 2**).

Interessantemente, o *cluster* de HoxD, que é mais comumente ligado ao desenvolvimento do sistema esquelético (Delpretti *et al*., 2012; Favier & Dollé, 1997; Zákány *et al*., 2004), apareceu em nossas redes. Da mesma forma, Hoxd10 já foi

observada por estar presente na região caudal do tubo neural e na medula espinhal (Kelly *et al*., 2009). Em nossa análise foi possível detectar diversos genes do *cluster* HoxD, incluindo Hoxd10 (na EGEP-*Network*). Essas relações implicam que o abuso do álcool durante a gravidez pode gerar baixa expressão de diversos genes HoxD no tecido cerebral dependendo do estágio do desenvolvimento em que foi exposto. Neste sentido, a ação do etanol durante o período pré-natal é mais agressiva visto que não foi detectado nenhum gene Hox na PSE-*Network*, e alguns genes HoxD pode continuar afetados caso a exposição seja dada nos estágios mais avançados do desenvolvimento como visto na LGEP-*Network*.

BOX12

HOXA2. Essencial também para o desenvolvimento da orelha (Brown *et al*., 2013)**.** Da mesma forma está envolvida com o desenvolvimento do telencéfalo (Wolf *et al*., 2001).

HOXA3. Importante para o desenvolvimento da glândula paratireoide (Kameda *et al*., 2004) e do sistema arterial da carótida (Kameda, 2009).

HOXA10. Expressa no endométrio e ligada a gravidez de ratos (Godbole *et al*., 2007). Ratos HOXA10$^{-/-}$ mostraram anomalias cranianas (Daftary & Taylor, 2006).

HOXA11. Envolvida na diferenciação de condrócitos (Gross *et al*., 2012). HOXA11 está intimamente ligada a proliferação, diferenciação e receptividade do endométrio (Daftary & Taylor, 2006; Wang *et al*., 2004).

Da mesma forma, alguns genes do *cluster* HoxA foram detectados na formação do tecido neural. Ratos Hoxa2$^{-/-}$ mostraram alteração na formação dos rombomeros, estruturas que são responsáveis pela padronização do rombencéfalo (Gavalas *et al*., 1997). Expressão de Hoxa3 também já foi observada em rombomeros durante o desenvolvimento neural de galinha (Kato *et al*., 1997).

Em nossa análise de biologia de sistemas foi possível observar que Hoxa10 e Hoxa11 estavam com baixa expressão na EGEP- e LGEP-*Networks* (**S-*Table* 2** (Tabela S2), ver **Material Suplementar 2**), mostrando que outros genes Hox podem ser alterados pelo etanol no tecido cerebral.

É importante ressaltar que o ácido retinóico (AR), principal derivado da vitamina A, é um dos principais sinalizadores de genes Hox durante o desenvolvimento embrionário (Daftary & Taylor, 2006), e a baixa expressão desses genes, como observados em nossas redes, pode ser resultado do efeito do etanol no metabolismo desse composto (como discutido no Capítulo II).

5. CONCLUSÃO GERAL

As análises desenvolvidas neste trabalho visam o avanço na pesquisa básica sobre o abuso de substâncias tóxicas durante o desenvolvimento embrionário, pois essa área carece de informações sobre como esses compostos afetam o desenvolvimento.

As análises de químio-biologia de sistemas, somadas as análises de transcriptomas prospectadas nos interatoma, identificaram diversos genes para serem levados em consideração para o entendimento das anomalias geradas pelos compostos do cigarro. Neste sentido, genes e proteínas envolvidos com o metabolismo de prostaglandinas e leucotrieno, assim como de sinalização e diferenciação celular (dependente de ácido retinóico e HOX) se mostraram afetados negativamente pelos compostos do cigarro. Também foram identificados uma ampla gama de proteínas que atuam na diferenciação osteogênica que são negativamente reguladas pelo abuso de substâncias do tabaco.

Já as quatros redes geradas para as análises para o abuso de etanol, que leva ao quadro patológico de SAF, ajudaram a identificar proteínas envolvidas na biossíntese e metabolismo de ácido retinóico, vitamina D, ácido ascórbico, α-tocoferol, niacina e ácido fólico, como importantes alvos para o entendimento de SAF. Em adição, foi possível identificar que o etanol pode causar quadros de neuroinflamação, problemas em vias sinápticas e na via de glutamato.

Por fim, esses dados forneceram uma visão global de bioprocessos e módulos proteicos que acrescentam importantes informações sobre esses quadros patológicos. Especialmente, nossos estudos mostram que ainda há muitas conexões que não foram feitas ou exploradas no estudo dos efeitos toxicológicos de substâncias químicas durante a gravidez.

6. CONCLUSÕES ESPECÍFICAS

- As redes relacionadas à análise toxicológica relacionada ao abuso de tabaco durante a gravidez permitiram observar quais compostos se ligam a determinados módulos e bioprocessos específicos.
- As análises de químio-biologia de sistemas mostraram que os processos que causam aumento de inflamação, como o acúmulo de leucotrieno, e a inibição da síntese de prostaglandina, afetam a implantação do embrião.
- Os dados indicam que a nicotina é capaz de afetar negativamente a sinalização de ácido retinóico e, consequentemente, a diferenciação celular.
- Os compostos do cigarro diminuem a expressão e inibem proteínas relacionadas diretamente à mineralização e diferenciação osteogênica.
- Os dados apontam para um modelo onde o cluster HOXD, que participa no desenvolvimento do sistema esquelético, é afetado pelo pelos compostos do cigarro.
- As análises transcritômicas corroboraram grande parte das hipóteses geradas.
- Os transcritômas analisados, visando analisar os efeitos do etanol no neurodesenvolvimento, permitiram uma análise dos efeitos desse composto durante o desenvolvimento do sistema nervoso a curto e longo prazo e a identificação de genes que possam explicar o desenvolvimento de SAF.
- Na análise dos efeitos toxicológicos do etanol no neurodesenvolvimento foi observado que o abuso dessa substância afeta criticamente o metabolismo e biossíntese de diversas vitaminas, principalmente de ácido retinóico, vitamina D, ácido ascórbico, α-tocoferol, niacina e ácido fólico.
- O etanol pode causar danos ao sistema nervoso através de um processo de neuroinflamação.

- Os dados apontam que o etanol possui um efeito negativo nos processos de diferenciação celular, sinapse e via de glutamato, podendo esses promover problemas na formação do cérebro e levar a disfunções cognitivas de aprendizado e coordenação motora.
- Alterações no ciclo circadiano e na sinalização de Ca^{++} causados pelo etanol podem interferir no sistema nervoso, causando danos.
- Diversos alvos proteicos foram selecionados e listados, contribuindo para um maior entendimento de como os compostos do cigarro e o etanol afetam o desenvolvimento embrionário foram selecionados e listados.

7. REFERÊNCIAS

ABOLHASSANI, N., IYAMA, T., TSUCHIOMOTO, D. et al. NUDT16 and ITPA play a dual protective role in maintaining chromosome stability and cell growth by eliminating dIDP/IDP and dITP/ITP from nucleotide pools in mammals. *Nucleic Acids Res*. 38(9), 2891-2903, 2010.

BA, A., SERI, B. V., & HAN, S. H. Thiamine administration during chronic alcohol intake in pregnant and lactating rats: effects on the offspring neurobehavioural development. *Alcohol Alcohol*. 31, 27-40, 1996.

BARABÁSI, A.L. & OLTAVAI, Z.N. Network Biology: Understanding the Cell's Functional Organization. *Nat Rev Genet.* 5(2), 101-13, 2004.

BEHR, B., LEUCHT, P., LONGAKER, M.T. et al. Fgf-9 is required for angiogenesis and osteogenesis in long bone repair. *Proc Natl Acad Sci U S A*. 107(26), 11853-11858, 2010.

BENOWITZ, N.L. Nicotine Addiction. *N Engl J Med*. 362(24), 2295-2303, 2010.

BJØRNEBOE, G.E., BJØRNEBOE, A., HAGEN, B.F. et al. Reduced hepatic alpha-tocopherol content after long-term administration of ethanol to rats. *Biochim Biophys Acta*. 918(3), 236-241, 1987.

BORELLO, U., COBOS, I., LONG, J.E. et al. FGF15 promotes neurogenesis and opposes FGF8 function during neocortical development. *Neural Dev*. 3, 17, 2008.

BOSE, M., WHITTAL, R.M., GAIROLA, C.G. et al. Cigarette smoke decreases mitochondrial porin expression and steroidogenesis. *Toxicol Appl Pharmacol.* 227(2), 284-290, 2008.

BOULET, A.M. & CAPECCHI, M.R. Duplication of the Hoxd11 gene causes alterations in the axial and appendicular skeleton of the mouse. *Dev Biol*. 249(1), 96-107, 2002.

BROWN, K.K., VIANA, L.M., HELWIG, C.C. et al. HOXA2 Haploinsufficiency in Dominant Bilateral Microtia and Hearing Loss. Hum Mutat. 2013. "No Prelo".

CHANDRA, N.; PADIADPU, J. Network approaches to drug discovery. *Expert Opin Drug Discov.* 8, 7-20, 2013.

CHANG, S.J., CHEN, S.M. CHIANG, S.L. et al. Association between cigarette smoking and hypoxanthine guanine phosphoribosyltransferase activity. *Kaohsiung J Med Sci*. 21(11), 495-501, 2005.

CHEUNG, V.C, DEBOER, C., HANSON, E. et al. Gene expression changes in the motor cortex mediating motor skill learning. *PLoS One*. 8(4), e61496, 2013.

CHAO, C.C., CHANG, P.Y. & LU, H.H. Human Gas7 isoforms homologous to mouse transcripts differentially induce neurite outgrowth. *J Neurosci Res*. 81(2), 153-162, 2005.

CHAO, C.C. HUNG, F.C. & CHAO, J.J. Gas7 is required for mesenchymal stem cell-derived bone development. *Stem Cells Int*. 2013, 137010, 2013.

CHO, K.W., KIM, J.Y., CHO, J.W. et al. Point mutation of Hoxd12 in mice. *Yonsei Med J*. 49(6), 965-972, 2008.

CHOE, A., PHUN, H.Q., TIEU, D.D. et al. Expression patterns of Hox10 paralogous genes during lumbar spinal cord development. *Gene Expr Patterns*. 6(7), 730-737, 2006.

CONSORTIUM, GENE ONTOLOGY. Creating the gene ontology resource: design and implementation. *Genome Res*. 11(8), 1425-33, 2001.

CONSORTIUM, INTERNATIONAL HUMAN GENOME SEQUENCING. Finishing the euchromatic sequence of the human genome. *Nature*. 431(7011), 931-945, 2004.

CSERMELY, P, KORCSMAROS, T., KISS, H.J.; et al. Structure and dynamics of molecular networks: A novel paradigm of drug discovery: A comprehensive review. *Pharmacol Ther.* 138(3), 333-408, 2013.

DAFTARY, G.S. & TAYLOR, H.S. Endocrine regulation of HOX genes. *Endocr Rev*. 27(4), 331-355, 2006.

da HORA JUNIOR, B.T., POLONI, J. de F., LOPES, M.A. et al. Transcriptomics and systems biology analysis in identification of specific pathways involved in cacao resistance and susceptibility to witches' broom disease. *Mol Biosyst.* 8(5), 1507-1519, 2012.

de FARIA POLONI, J., FELTES, B.C. & BONATTO, D. Melatonin as a central molecule connecting neural development and calcium signaling. *Funct Integr Genomics*. 11(3), 383-388, 2011.

DHAR, M.K., KOUL, A. & KAUL, S. Farnesyl pyrophosphate synthase: a key enzyme in isoprenoid biosynthetic pathway and potential molecular target for drug development. *N Biotechnol*. 30(2), 114-123, 2013.

de MAGALÃES, J.P. & TOUSSAINT, O. GenAge: a genomic and proteomic network map of human ageing. *FEBS Lett.* 571(1-3), 243-247, 2004.

DELPRETTI, S., ZAKANY, J. & DUBOULE, D. A function for all posterior Hoxd genes during digit development? *Dev. Dyn*. 241(4), 792-802, 2012.

DORES, R.M. & BARON, A.J. Evolution of POMC: origin, phylogeny, posttranslational processing, and the melanocortins. *Ann N Y Acad Sci.* 1220, 34-48, 2011.

FEALA, J.D., ABDULHAMEED, M.D., YU, D. et al. Systems biology approaches for discovering biomarkers for traumatic brain injury. *J Neurotrauma*. 30(13), 1101-1116, 2013.

FAVIER, B. & DOLLÉ, P. Developmental functions of mammalian Hox genes. *Mol Hum Reprod.* 3(2), 115-131, 1997.

FELTES, B.C., de FARIA POLONI, J., NOTARI, D.L. et al. Toxicological effects of the different substances in tobacco smoke on human embryonic development by a systems chemo-biology approach. *PLoS One*, 8(4), e61743, 2013.

FELTES, B.C., de FARIA POLONI, J. & BONATTO, D. The developmental aging and origins of health and disease hypotheses explained by different protein networks. *Biogerontology.* 12(4), 293-308, 2011.

FERECATU, I., RINCHEVAL, V., MIGNOTTE, B. et al. Tickets for p53 journey among organelles. *Front Biosci*. 14, 4214-4228, 2009.

FOWLER, P.A., CASSIE, S., RHIND, S.M. et al. Maternal smoking during pregnancy specifically reduces human fetal desert hedgehog gene expression during testis development. *J. Clin Endocrinol Metab.* 93(2), 619-626, 2008.

GAVALAS, A., DAVENNE, M., LUMSDEN, A. et al. Role of Hoxa-2 in axon pathfinding and rostral hindbrain patterning. *Development*. 124(19), 3693-3702, 1997.

GENETTA, T., LEE, B.H. & SOLA, A. Low Doses of Ethanol and Hypoxia Administered Together Act Synergistically to Promote the Death of Cortical Neurons. J Neurosci Res. 85(1), 131-138, 2007.

GERSCH, R.P., LOMBARDO, F., McGOVERN, S.C. et al. Reactivation of Hox gene expression during bone regeneration. *J Orthop Res*. 23(4), 882-890, 2005.

GODBOLE, G.B., MODI, D.N. & PURI, C.P. Regulation of homeobox A10 expression in the primate endometrium by progesterone and embryonic stimuli. *Reproduction.* 134(3), 513-23, 2007.

GOEZ, H.R., SCOTT, O. & HASAL, S. Fetal Exposure to Alcohol, Developmental Brain Anomaly, and Vitamin A Deficiency: A Case Report. *J Child Neurol.* 26(2), 231-234, 2011.

GOGGIN, M., SANGARAJU, D., WALKER, V.E. et al. Persistence and repair of bifunctional DNA adducts in tissues of laboratory animals exposed to 1,3-butadiene by inhalation. *Chem Res Toxicol*. 24(6), 809-817, 2011.

GRIFFIN, J.N., COMPAGNUCCI, C., HU, D. et al. Fgf8 dosage determines midfacial integration and polarity within the nasal and optic capsules. *Dev Biol*. 374(1), 185-197, 2013.

GROSS, S., KRAUSE, Y., WUELLING M. et al. Hoxa11 and Hoxd11 regulate chondrocyte differentiation upstream of Runx2 and Shox2 in mice. *PLoS One.* 7(8), e43553, 2012.

HACKSHAW, A., RODEK, C. & BONIFACE, S. Maternal smoking in pregnancy and birth defects: a systematic review based on 173 687 malformed cases and 11.7 million controls. *Human Reprod Update.* 17(5), 589–604, 2011.

HAYES, J.D., FLANAGAN, J.U. & JOWSEY, I.R. Glutathione transferases. *Annu Rev Pharmacol Toxicol*. 45, 51-88, 2005.

HAGUE, T., NAKADA, S. & HAMDY, R.C. A review of FGF18: Its expression, signaling pathways and possible functions during embryogenesis and post-natal development. *Histol Histopathol*. 22(1), 97-105, 2007.

HEWITT, A.J., KNUFF, A.L., JEFKINS, M.J. et al. Chronic ethanol exposure and folic acid supplementation - fetal growth and folate status in the maternal and fetal guinea pig. *Reprod Toxicol.* 31(4), 500-506, 2011.

HOGEWEG, P. The Roots of Bioinformatics in Theoretical Biology. *PLoS Comput Biol*. 7(3), e1002021, 2011.

HUANG, C., XIANG, Y., WANG, Y. et al. Dual-specificity histone demethylase KIAA1718 (KDM7A) regulates neural differentiation through FGF4. *Cell Res.* 20(2), 154-165, 2010.

HUANG, B.T., CHANG, P.Y., SU, C.H. et al. Gas7-deficient mouse reveals roles in motor function and muscle fiber composition during aging. *PLoS One.* 7(5), e37702, 2012.

IZZOTTI, A., CARTIGLIA, C., LONGOBARDI, M. et al. Gene expression in the lung of p53 mutant mice exposed to cigarette smoke. *Cancer Res*. 64(23), 8566-8572, 2004.

JAUNIAUX, E. & BURTON, G.J. Morphological and biological effects of maternal exposure to tobacco smoke on the feto-placental unit. *Early Hum Dev.* 83(11), 699–706, 2007.

JIANG, W., YU, Q., GONG, M. et al. Vitamin A deficiency impairs postnatal cognitive function via inhibition of neuronal calcium excitability in hippocampus. *J Neurochem.* 121(6), 932-943, 2012.

KAMP, D.W., SHACTER, E. & WEITZMAN, S.A. Chronic inflammation and cancer: the role of the mitochondria. *Oncology (Williston Park)*. 25(5), 400-410, 2011.

KAMEDA, Y. Hoxa3 and signaling molecules involved in aortic arch patterning and remodeling. *Cell Tissue Res.* 336(2), 165-178, 2009.

KAMEDA, Y., ARAI, Y., NISHIMAKI, T. et al. The role of Hoxa3 gene in parathyroid gland organogenesis of the mouse. *J Histochem Cytochem*. 52(5), 641-651, 2004.

KATAOKA, A. & SHIMOGORI, T. Fgf8 controls regional identity in the developing thalamus. *Development.* 135(17), 2873-2881, 2008.

KATO, K., O'DOWD, D.K., FRASER, S.E. et al. Heterogeneous expression of multiple putative patterning genes by single cells from the chick hindbrain. *Dev Biol*. 191(2), 259-269, 1997.

KE, Z.J., WANG, X., FAN, Z. et al. Ethanol Promotes Thiamine Deficiency-Induced Neuronal Death: Involvement of Double-Stranded RNA-activated Protein Kinase. *Alcohol Clin Exp Res*. 33(6), 1097-1103, 2009.

KELLY, T.K., KARSTEN, S.L., GESCHWIND, D.H. et al. Cell lineage and regional identity of cultured spinal cord neural stem cells and comparison to brain-derived neural stem cells. *PLoS One*. 4(1), e4213, 2009.

KILDEGAARD, H.F., BAYCIN-HIZAl, D. LEWIS, N.E. et al. The emerging CHO systems biology era: harnessing the 'omics revolution for biotechnology. *Curr Opin Biotechnol*. 24, 1-6, 2013.

KING, C.D., RIOS, G.R., GREEN, M.D. et al. UDP-glucuronosyltransferases. *Curr Drug Metab*. 1(2), 143-161, 2000.

KIR, S., BEDDOW, S.A., SAMUEL, V.T. et al. FGF19 as a postprandial, insulin-independent activator of hepatic protein and glycogen synthesis. *Science*. 331(6024), 1621-1624, 2011.

KIRSCH, S. H., HERRMANN, W. & OBEID, R. Genetic defects in folate and cobalamin pathways affecting the brain. *Clin Chem Lab Med*. 51, 139-155, 2013.

KIZHNER, T., BEN-DAVID, D., ROM, R. et al. Effects of FGF2 and FGF9 on osteogenic differentiation of bone marrow-derived progenitors. *In Vitro Cell Dev Biol Anim*. 47(4), 294-301, 2011.

KOTURBASH, I., SCHERHAG, A., SORRENTINO, J. et al. Epigenetic mechanisms of mouse interstrain variability in genotoxicity of the environmental toxicant 1,3-butadiene. *Toxicol Sci*. 122(2), 448-456, 2011.

LADAM, F. & SAGERSTRÖM, C.G. Hox regulation of transcription: More complex(es). *Dev Dyn*. 2013. "No Prelo"

LANDER, E.S., LINTON, L.M., BIRREN, B. et al. Initial sequencing and analysis of the human genome. *Nature*. 409(6822), 860-921, 2001.

LAZAKOVITCH, E.M., SHE, B.R., LIEN, C.L. et al., The Gas7 gene encodes two protein isoforms differentially expressed within the brain. *Genomics.* 61(3), 298-306, 1999.

LAEZZA, F., GERBER, B.R., LOU, J.Y. et al. The FGF14(F145S) mutation disrupts the interaction of FGF14 with voltage-gated Na+ channels and impairs neuronal excitability. *J Neurosci*. 27(44), 12033-12044, 2007.

LEE, J.Y., CHANG, M.Y., PARK, C.H. et al. Ascorbate-induced differentiation of embryonic cortical precursors into neurons and astrocytes. *J Neurosci Res.* 73(2), 156-65, 2003.

LEUNG, K. Y., De CASTRO, S. C., CABREIRO, F. et al. Folate metabolite profiling of different cell types and embryos suggests variation in folate one-carbon metabolism, including developmental changes in human embryonic brain. *Mol Cell Biochem.* 378(1-2), 229-236, 2013.

LIN, S., McLENNAN, A.G., YING, K. et al. Cloning, expression, and characterization of a human inosine triphosphate pyrophosphatase encoded by the itpa gene. *J Biol Chem*. 276(22), 18695-18701, 2001.

LIU, F., POGODA, H.M., PEARSON, C.A. et al. Direct and indirect roles of Fgf3 and Fgf10 in innervation and vascularisation of the vertebrate hypothalamic neurohypophysis. *Development*. 140(5), 1111-1122, 2013.

LORTIE, K., HUANG, D., CHAKRAVARTHY, B. et al. The gas7 protein potentiates NGF-mediated differentiation of PC12 cells. *Brain Res*. 1036(1-2), 27-34, 2005.

MACATEE, T.L., HAMMOND, B.P., ARENKIEL, B.R. et al. Ablation of specific expression domains reveals discrete functions of ectoderm- and endoderm-derived FGF8 during cardiovascular and pharyngeal development. *Development.* 130(25), 6361-74, 2003.

MARTIN-MARTIN, N.; SUTHERLAND, J.D.; CARRACEDO, A. PML: Not all about Tumor Suppression. *Front Oncol*. 3, 200, 2013.

MARTINEZ, P. & AMEMYIA, C.T. Genomics of the HOX gene cluster. *Comp Biochem Physiol B Biochem Mol Biol.* 133(4), 571-580, 2002.

MILLER, W.L. Steroid hormone synthesis in mitochondria. *Mol Cell Endocrinol.* S0303-7207(13), 00159-00157, 2013.

MORSE, N. L. Benefits of docosahexaenoic acid, folic acid, vitamin D and iodine on foetal and infant brain development and function following maternal supplementation during pregnancy and lactation. *Nutrients*. 4, 799-840, 2012.

NAG, S, QIN, J., SRIVENUGOPAL, K.S. et al. The MDM2-p53 pathway revisited. *J Biomed Res*. 27(4), 254-271, 2013.

NAGAYAMA, T., OKUHARA, S., OTA, M.S. et al. FGF18 accelerates osteoblast differentiation by upregulating Bmp2 expression. *Congenit Anom (Kyoto)*. 53(2), 83-88, 2013.

NEWMAN, M.E.J. The Structure and Function of Complex Networks. *SIAM Review*. 45, 167-256, 2003.

PFEIFER, G.P., DENISSENKO, M.F, OLIVER, M. et al. Tobacco smoke carcinogens, DNA damage and p53 mutations in smoking- associated cancers. *Oncogene*. 21(48), 7435–7451, 2002.

PIASEK, M.. BLANUSA, M., KOSTIAL, K. et al. Placental cadmium and progesterone concentration in cigarette smokers. *Reprod Toxicol*. 15, 673-681, 2001.

PLIARCHOPOULOU, K., VOUTSINAS, G., PAPAXOINIS, G. et al. Correlation of CYP1A1, GSTP1 and GSTM1 gene polymorphisms and lung cancer risk among smokers. *Oncol Lett.* 3(6), 1301-1306, 2012.

O'LEARY, C.M. Fetal alcohol syndrome: diagnosis, epidemiology, and developmental outcomes. *J Paediatr Child Health.* 40(1-2), 2-7, 2004.

OUZOUNIS, C.A. Rise and Demise of Bioinformatics? Promise and Progress. *PLoS Comput Biol*. 8(4), e1002487, 2012.

PALSSON, B. The challenges of in silico biology. *Nat Biotechnol.* 18(11), 1147-50, 2000.

PIKULEVA, I.A. & WATERMAN, M.R. Cytochromes p450: roles in diseases. *J Biol Chem.* 288(24), 17091-17098, 2013.

RASH, B.G., TOMASI, S., LIM, H.D. et al. Cortical gyrification induced by fibroblast growth factor 2 in the mouse brain. *J Neurosci*. 33(26), 10802-10814, 2013.

RHINN, M. & DOLLÉ, P. Retinoic acid signalling during development. *Development*. 139(5), 843-858, 2012.

ROGERS, J.M. Tobacco and Pregnancy: Overview of Exposures and Effects. *Birth Defects Res C Embryo Today*. 84(1), 1-15, 2008.

ROSADO, J.O., HENRIQUES, J.P. & BONATTO, D. A systems pharmacology analysis of major chemotherapy combination regimens used in gastric cancer treatment: predicting potential new protein targets and drugs. *Curr Cancer Drug Targets*, 11, 849-869, 2011.

SADEU, J.C & FOSTER, W.G. Cigarette smoke condensate exposure delays follicular development and function in a stage-dependent manner. *Fertil Steril*. 95(7), 2410-2417, 2011.

SAKUMI, K., ABOLHASSANI, N., BEHMANESH, M. et al. ITPA protein, an enzyme that eliminates deaminated purine nucleoside triphosphates in cells. *Mutat Res.* 703(1), 43-50, 2010.

SALOMONI, P. The PML-Interacting Protein DAXX: Histone Loading Gets into the Picture. *Front Oncol*. 3, 152, 2013.

SCARDONI, G. & LAUDANNA C. Centralities Based Analysis of Complex Networks. Em *"New Frontiers in Graph Theory"*. Editado por Yagang Zhang, ISBN 978-953-51-0115-4, 2009.

SOBRAL, B.W.S. Bioinformatics and the future role of computing in biology. In: *From Jay Lush to Genomics: Visions for animal breeding and genetics*. 115-123, 1999.

STRANGE, R.C., SPITERI, M.A., RAMACHANDRAN, S. et al. Glutathione-S-transferase family of enzymes. *Mutat Res*. 482(1-2), 21-26, 2001.

SUGATANI, J. Function, genetic polymorphism, and transcriptional regulation of human UDP-glucuronosyltransferase (UGT) 1A1. *Drug Metab Pharmacokinet*. 28(2), 83-92, 2013.

STEVENSON, L.F., SPARKS, A., ALLENDE-VEGA, N. et al. The deubiquitinating enzyme USP2a regulates the p53 pathway by targeting Mdm2. *EMBO J*. 26(4), 976-986, 2007.

TERWISSCHA van SCHELTINGA, A.F., BAKKER, S.C., KAHN, R.S. et al. Fibroblast Growth Factors in Neurodevelopment and Psychopathology. *Neuroscientist*. 2013. “No Prelo”.

TRAPANI, L., SEGATTO, M., ASCENZI, P. et al. Potential role of nonstatin cholesterol lowering agents. *IUBMB Life*. 63(11), 964-971, 2011.

TVEDEN-NYBORG, P., VOGT, L., SCHJOLDAGER, J.G. et al. Maternal vitamin C deficiency during pregnancy persistently impairs hippocampal neurogenesis in offspring of guinea pigs. *PLoS One*. 7(10), e48488, 2012.

van der AKKER, E., FROMENTAL-RAMAIN, C., de GRAAFF, W. et al. Axial skeletal patterning in mice lacking all paralogous group 8 Hox genes. *Development*. 128(10), 1911-1921, 2001.

van der TOORN, M., REZAYAT, D., KAUFFMAN, H.F. et al. Lipid-soluble components in cigarette smoke induce mitochondrial production of reactive oxygen species in lung epithelial cells. *Am J Physiol Lung Cell Mol Physiol.* 297(1), L109-L114, 2009.

VÁZQUEZ, A. Growing network with local rules: Preferential attachment, clustering, hierarchy, and degree correlations. *Phys Rev E Stat Nonlin Soft Matter Phys*. 67(2), 056104, 2003.

VILLAVICENCIO-LORINI, P., KUSS, P., FRIEDRICH, J. et al. Homeobox genes d11-d13 and a13 control mouse autopod cortical bone and joint formation. *J Clin Invest*. 120(6), 1994-2004, 2010.

XIAO, M., BOSCH, M.K., NERBONNE, J.M. et al. FGF14 localization and organization of the axon initial segment. *Mol Cell Neurosci*. S1044-7431(13), 00074-00072, 2013.

XU, B. & WELLIK, D.M. Axial Hox9 activity establishes the posterior field in the developing forelimb. *Proc Natl Acad Sci U S A*. 108(12), 4888-4891, 2011.

WAGNER, G.P., PAVLICEV, M. & CHEVERUD, G.M. The road to modularity. *Nat Rev Genet.* 8(12), 921-931, 2007.

WANG, L.F., LUO, H.Z., ZHU, Z.M. et al. Expression of HOXA11 gene in human endometrium. *Am J Obstet Gynecol*. 191(3), 767-772, 2004.

WEISINGER, K., KOHL, A., KAYAM, G. et al. Expression of hindbrain boundary markers is regulated by FGF3. *Biol Open*. 1(2), 67-74, 2012.

WENTZEL, P. & ERIKSSON, U.J. Altered gene expression in neural crest cells exposed to ethanol in vitro. *Brain Res.* 1305, 50-60, 2009.

WOLF, L.V., YEUNG, J.M., DOUCETTE, J.R. et al. Coordinated expression of Hoxa2, Hoxd1 and Pax6 in the developing diencephalon. *Neuroreport.* 12(2), 329-333, 2001.

YAN, H., PABLO, J.L. & PITT, G.S. FGF14 Regulates Presynaptic Ca(2+) Channels and Synaptic Transmission. *Cell Rep*. 4(1), 66-75, 2013.

YOU, J.J. & LIN-CHAO, S. Gas7 functions with N-WASP to regulate the neurite outgrowth of hippocampal neurons. *J Biol Chem*. 285(15), 11652-11666, 2010.

YU, H., KIM, P. M., SPRECHER, E. et al. The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. *PLoS Comput. Biol.* 3, e59, 2007.

ZÁKÁNY, J., KMITA, M. & DUBOULE, D. A dual role for Hox genes in limb anterior-posterior asymmetry. *Science*. 304(5677), 1669-1672, 2004.

ZANNI, G., BARRESI, S., TRAVAGLINI, L. et al. FGF17, a gene involved in cerebellar development, is downregulated in a patient with Dandy-Walker malformation carrying a de novo 8p deletion. *Neurogenetics*. 12(3), 241-245, 2011.

ZHANG, C., OJIAKU, P. & COLE, G.J. Forebrain and hindbrain development in zebrafish is sensitive to ethanol exposure involving agrin, Fgf, and sonic hedgehog function. *Birth Defects Res A Clin Mol Teratol.* 97(1), 8-27, 2013.

8. ADENDOS

Molecular BioSystems

RSC Publishing

PAPER

View Article Online
View Journal

Combining small molecules for cell reprogramming through an interatomic analysis†

Cite this: DOI: 10.1039/c3mb70159j

Bruno César Feltes and Diego Bonatto*

The knowledge available about the application and generation of induced pluripotent stem cells (iPSC) has grown since their discovery, and new techniques to enhance the reprogramming process have been described. Among the new approaches to induce iPSC that have gained great attention is the use of small molecules for reprogramming. The application of small molecules, unlike genetic manipulation, provides for control of the reprogramming process through the shifting of concentrations and the combination of different molecules. However, different researchers have reported the use of ''reprogramming cocktails'' with variable results and drug combinations. Thus, the proper combination of small molecules for successful and enhanced reprogramming is a matter for discussion. However, testing all potential drug combinations in different cell lineages is very costly and time-consuming. Therefore, in this article, we discuss the use of already employed molecules for iPSC generation, followed by the application of systems chemo-biology tools to create different data sets of protein–protein (PPI) and chemical–protein (CPI) interaction networks based on the knowledge of already used and new reprogramming cocktail combinations. We further analyzed the biological processes associated with PPI–CPI networks and provided new potential protein targets to be inhibited or expressed for stem cell reprogramming. In addition, we applied a new interference analysis to prospective targets that could negatively affect the classical pluripotency-associated factors (SOX2, NANOG, KLF4 and OCT4) and thus potentially improve reprogramming protocols.

Received 19th April 2013,
Accepted 12th August 2013

DOI: 10.1039/c3mb70159j

www.rsc.org/molecularbiosystems

Introduction

Embryonic stem cells are derived from the inner cell mass of the blastocyst and have the ability to generate all cells of the organism, excluding the extraembryonic tissues.[1] The capability of self-renewal and the ability to differentiate into a wide variety of cell types is unique among pluripotent stem cells (PSCs), which not only makes them a good model to study many aspects of development but also offers a clinical solution for the treatment of many diseases, such as leukemia, osteoarthritis, schizophrenia and Parkinson's disease.[2–5]

Unfortunately, difficulties in obtaining PSCs due to ethical and practical issues led to the search of new alternatives to obtain embryonic-like cells. In this sense, somatic cells can be reverted to a pluripotent state after the induction of specific transcription factors, and the resulting cells are called induced pluripotent stem cells (iPSCs).[6]

The use of iPSCs has grown in recent years, and different transcription factors have been discovered to enhance the self-renewal and stability of iPSCs.[7] However, the genetic manipulation of cells to express such transcription factors [*e.g.*, by using integrative vector insertion techniques (reviewed in ref. 8)] does not provide high temporal control over the expression of a protein and its function and is not reversible because the exogenous genes are permanently integrated into the host cell genome. Thus, new approaches that can be applied for iPSCs generation have been developed.

Considering the different methods used for iPSC reprogramming, the application of small molecules has been attracting the attention of researchers. Small molecules are relatively easy to use in cell cultures and can provide a high temporal control over protein functions by varying only their concentration and combination with other molecules.[1,4]

It is interesting to observe that all new molecules discovered for iPSC generation are involved in enhanced reprogramming, prolonged iPSC survival or maintenance of pluripotency. Nevertheless, different experiments that present new compounds for iPSC generation use wide varieties of molecule concentrations and combinations, of exogenous gene insertions, and of different growth media conditions. Testing such combinations is costly and would not necessarily show the same effects when applied in

Centro de Biotecnologia da Universidade Federal do Rio Grande do Sul, Departamento de Biologia Molecular e Biotecnologia, Universidade Federal do Rio Grande do Sul, Avenida Bento Gonçalves 9500 - Prédio 43421 - Sala 219, Porto Alegre, Caixa Postal 15005, RS – Brazil. E-mail: diegobonatto@gmail.com; Fax: +55 51-3308-7309; Tel: +55 51-3308-6080

† Electronic supplementary information (ESI) available. See DOI: 10.1039/c3mb70159j

different cell growth and/or maintenance conditions. Moreover, the use of small molecules for iPSC generation has been widely discussed.[1,4,9–12] Considering that new molecules are continuously being described for iPSC generation, there is no consensus about what small compounds should be used and what best molecular associations are to be employed for cell reprogramming. Although new iPSC-generating molecules are described and screened by high-throughput approaches,[1,4,13–15] their combined use in reprogramming cocktails has not yet been characterized. Additionally, small molecules that are already being extensively used in iPSC reprogramming are normally found in combination with other small molecules to generate iPSCs.[1,4,9,11,16] In addition, the discovery of new molecules by high-throughput techniques *per se* does not add any knowledge about the biochemical mechanisms affected by these compounds leading to iPSC generation. Because the use of small molecules appears to be a breakthrough in cellular reprogramming, the combination of the most beneficial drugs to reprogram different tissues demands an appropriate discussion and analysis of how to create "reprogramming cocktails". Therefore, the use of *in silico* approaches could help to elucidate the mechanisms underlying the effects of small molecules in SC reprogramming and, at the same time, to discover new potential targets to be inhibited for enhanced iPSC generation.

Systems chemo-biology and systems pharmacology tools have been employed for drug discovery and biochemical pathway analysis,[17–19] allowing the researcher to observe the major processes that can be affected by small molecules. A clear example of the potential use of systems pharmacology is the prospection of new antitumoral drugs.[20] Similarly, these tools could also be applied for the analysis and prospection of drugs for iPSC generation.

In this article, we critically discuss the uses of small molecules for iPSC generation. Furthermore, drawing from specific chemical and biological databases, we used prospective proteomic and small molecule data to create protein–protein interaction (PPI) and chemical–protein interaction (CPI) networks. By applying cluster and centrality analyses, we indicated the potential best combination of small molecules and prospected new protein targets for iPSC generation. Gene ontology (GO) analysis was performed to observe the most influenced biological processes by small molecules, and we introduced new potential small molecules for iPSC reprogramming by providing a list of new proteins to be inhibited for successful reprogramming. Furthermore, using an analysis of interference, we also prospectively selected proteins that might negatively affect the classical factors involved in the pluripotency state.

Materials and methods

Interactome data mining and design of the chemo-biology network associated with small compounds for iPSC reprogramming

To design chemo-biology interactome networks and to identify potential new protein targets for the iPSC-associated reprogramming of small molecules, the metasearch engines STITCH 3.1 [http://stitch.embl.de/] and STRING 9.0 [http://string-db.org/][21,22] were used. In this sense, a list of all small molecules described in other reviews and other high throughput original papers[1,4,9,15,16] of commonly used small-molecules in iPSC reprogramming were used as the initial seeds for network selection in STITCH. The STITCH software allows for the visualization of the physical connections among different proteins and chemical compounds, whereas STRING shows protein–protein interactions. Each protein–protein or protein–chemical connection (edge) shows a degree of confidence between 0 and 1.0 (with 1.0 indicating the highest confidence). The parameters used in STITCH software were as follows: all prediction methods enabled, excluding text mining; 20 to 50 interactions; degree of confidence, medium (0.400); and a network depth equal to 1. The results gathered using these search engines were analyzed with Cytoscape 2.8.2.[23] In addition, the GeneCards [http://www.genecards.org/],[24,25] KEGG [http://www.genome.jp/kegg/],[26] PubChem [http://pubchem.ncbi.nlm.nih.gov/] and AmiGO 1.8 [http://amigo.geneontology.org/cgi-bin/amigo/go.cgi][27] search engines were also employed, using their default parameters.

To select prospective protein–protein and chemical–protein interactions (PPI and CPI, respectively), we entered each small-molecule into the STITCH program. Molecules that were not present in the STITCH database (or those that did not show any protein connections) were excluded from the analysis.

Different small CPI and PPI networks were obtained (data not shown), and these networks were further analyzed using Cytoscape 2.8.2. Each network generated by STITCH and STRING was combined into a large network using the Advanced Merge Network function, which was fully implemented in the Cytoscape software.

Module analysis of the CPI–PPI networks associated with small compounds used in iPSC reprogramming

The large CPI–PPI network obtained from the initial search (Fig. 1) was analyzed in terms of the major cluster or module composition using the program Molecular Complex Detection (MCODE),[28] which is available at http://baderlab.org/Software/MCODE. MCODE is based on vertex weighting by the local neighborhood density and outward traversal from a locally dense seed protein to isolate the highly clustered regions, according to given parameters stipulated by the researcher (Bader and Hogue, 2003). The parameters for cluster finding were as follows: loops included; degree cutoff, 2; expansion of a cluster by one neighbor shell allowed (fluff option enabled); deletion of a single connected node from clusters (haircut option enabled); node density cutoff, 0.1; node score cutoff, 0.2; kcore, 2; and maximum network depth, 100. Each cluster generates a value of "cliquishness" (C_i), which is the degree of connection in a given group of proteins. Thus, the higher the C_i value, the more connected the cluster.[28]

Centrality analysis of CPI–PPI networks

Centrality analysis was performed using the program CentiScaPe 1.2.[29] In this analysis, the CentiScaPe algorithm evaluates each network node according to the node degree, betweenness and closeness to establish the most "central" nodes (proteins/chemicals) within the network. Thus, the most relevant node for a determined biochemical pathway or module can be

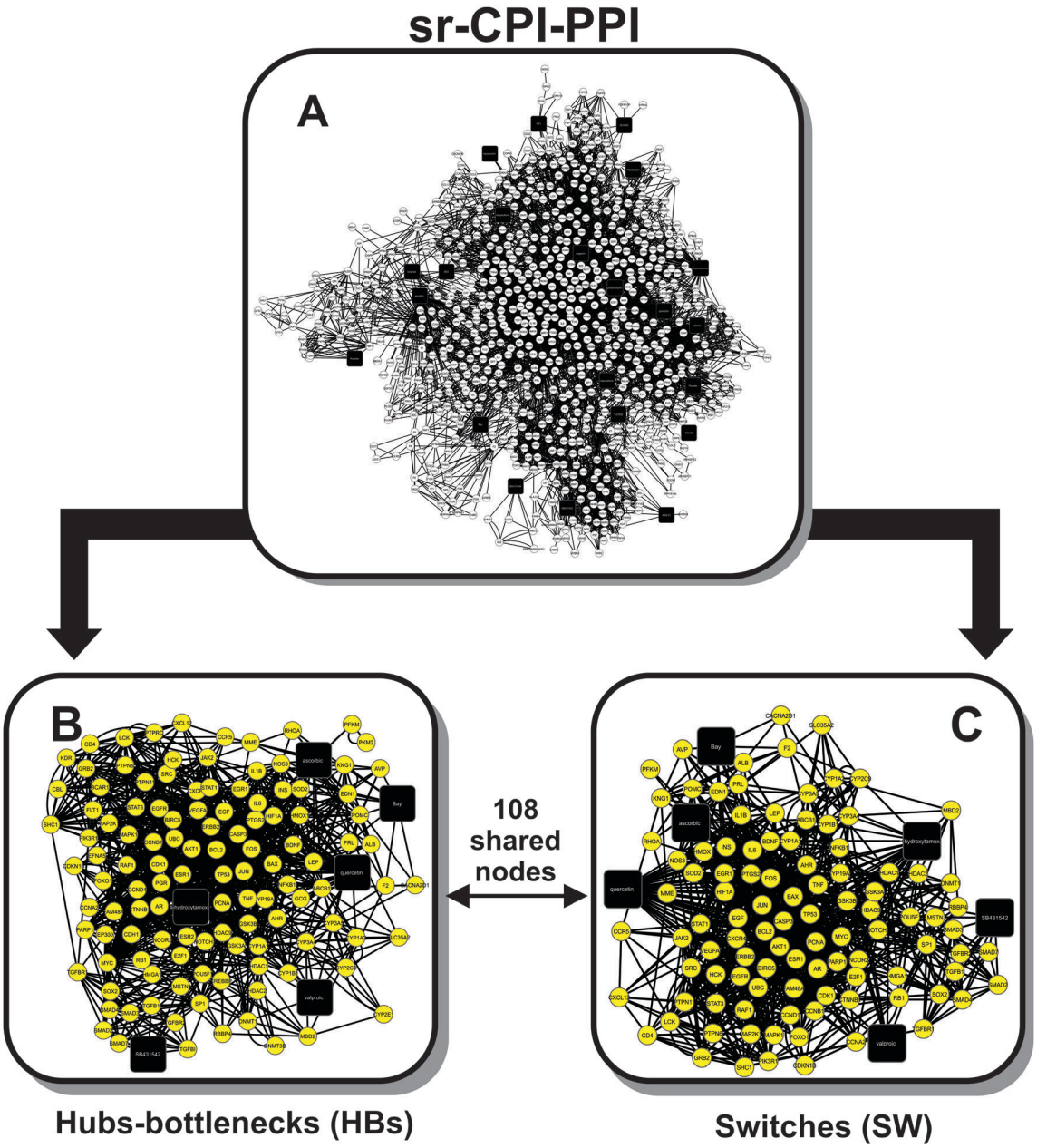


Fig. 1 CPI–PPI network derived from 22 subnetworks containing small-molecules commonly used in SC reprogramming. (A) Major CPI–PPI network that was derived from the merging of the 22 small molecules subnetworks. The network is composed of 858 nodes and 6653 edges. (B) Hubs-bottlenecks (HBs) that were extracted from the major CPI–PPI network in the centrality analysis crosslinking the above average values of node-degree and betweenness. The HB network contains 124 nodes and 1339 edges. (C) Switches (SW) that were also extracted from the major CPI–PPI network in the centrality analysis crosslinking the above average values of closeness and betweenness. The SW network is composed of 108 nodes and 1161 edges.

obtained and further analyzed. In general terms, the closeness analysis (1) indicates the probability that any protein/chemical compound (node in our network) is relevant to another protein/chemical compound (node) in a signaling network or its associated network,[29] as determined using eqn (1):

$$\mathrm{Clo}(v) = \frac{1}{\sum w \in v^{\mathrm{dist}(v,w)}} \quad (1)$$

where the closeness value of node v (Clo(v)) is determined by computing and totalizing the shortest paths among node v and all other nodes (w; dist(v,w)) found within a network (1). The average closeness (Clo) score was obtained by calculating the sum of different closeness scores (Clo$_i$) divided by the total number of nodes analyzed ($N(v)$) (eqn (2)).

$$\langle \mathrm{Clo} \rangle = \frac{\sum_i \mathrm{Clo}_i}{N_{(v)}} \quad (2)$$

The higher the closeness value compared to the average closeness score, the higher the relevance of the protein/chemical

compounds to other protein nodes within the network/module. In turn, the betweenness indicates the number of the shortest paths that go through each node (eqn (3)):[29,30]

$$\mathrm{Bet}(v) = \sum_{s \neq v \neq w} \frac{\sigma_{sw}(v)}{\sigma_{sw}} \quad (3)$$

where σ_{sw} is the total number of the shortest paths from node s to node w, and $\sigma_{sw}(v)$ is the number of those paths that pass through the node. The average betweenness score (Bet) of the network was calculated using eqn (4), where the sum of different betweenness scores (Bet_i) is divided by the total number of nodes analyzed ($N(v)$):

$$\langle \mathrm{Bet} \rangle = \frac{\sum_i \mathrm{Bet}_i}{N_{(v)}} \quad (4)$$

Thus, nodes with high betweenness scores compared to the average betweenness score of the network are responsible for controlling the flow of information through the network topology. The higher a node's betweenness score, the higher the probability that the node connects different modules or biological processes; such nodes are called bottleneck nodes.

Finally, the node degree ($\mathrm{Deg}(v)$) is a measure that indicates the number of connections (E_i) that involve a specific node (v) (eqn (5)):

$$\mathrm{Deg}(v) = \sum E_i \quad (5)$$

The average node degree of a network (Deg) is given by eqn (6), where the sum of different node degree scores (Deg_i) is divided by the total number of nodes ($N(v)$) present in the network:

$$\langle \mathrm{Deg} \rangle = \frac{\sum_i \mathrm{Deg}_i}{N_{(v)}} \quad (6)$$

Nodes with a high node degree are called hubs[29] and have key regulatory functions in the cell.

Gene ontology analyses of the CPI–PPI networks

The CPI–PPI modules generated by MCODE were further studied by focusing on major biology-associated processes using the Biological Network Gene Ontology (BiNGO) 2.44 Cytoscape plugin,[31] available at http://www.cytoscape.org/plugins2.php#IO_PLUGINS. The degree of functional enrichment for a given cluster and category was quantitatively assessed (p-value) using a hypergeometric distribution. Multiple test correction was also assessed by applying the false discovery rate (FDR) algorithm,[32] which was fully implemented in BiNGO software at a significance level of $p < 0.05$. The most statistically relevant processes were taken into account when developing the interaction model.

Interference analyses of the CPI–PPI networks

To analyze all possible proteins that could negatively affect the activity of pluripotency-associated transcriptional factors, such as Yamanaka and Thomson factors (*e.g.*, SOX2, NANOG, OCT4 and KLF4), we used the Cytoscape plug-in Interference 1.0, which is available at http://www.cbmc.it/~scardonig/interference/Interference.php.

This software generates virtual knockout events, in which each node is deleted, and the betweenness is recalculated for the entire network. To accomplish such an analysis, we developed individual networks containing 200–210 proteins for each transcriptional pluripotency-associated factor. The STRING search parameters in this search were the same as those used for the small-molecule networks. Only results that displayed more than 0.2 (20%) of interference were selected.

Results and discussion

Major network for stem cell self-renewal and plasticity maintenance

The data gathered were used to create networks containing small molecules associated with stem cell self-renewal and increasing plasticity. The union of these networks allows the generation of a network named the self-renewal CPI–PPI network (sr-CPI–PPI network) (Fig. S1, see ESI†).

The sr-CPI–PPI network contains 863 nodes (proteins/small molecules), 22 cell reprogramming small molecules that were found in the STITCH database, and 6677 edges (interactions) (Fig. 1A). To understand the modules present in our major CPI–PPI network (Fig. 1A), we applied the Cytoscape plugin MCODE. In addition to module discovery, we defined all small molecules found in more than 4 modules as promiscuous, an important concept in pharmacology, as promiscuous small molecules could affect different biological processes[18] and thus are potentially more effective in iPSC reprogramming.

From this initial analysis, 14 significant clusters were found (Fig. S2 to S15, see ESI†), of which the major promiscuous small molecules are parnate (5 appearances in 14 clusters) (Fig. S2, S5, S7, S11 and S15, ESI†), sodium butyrate (5/14) (Fig. S2, S3, S5, S8 and S11, ESI†), Bayk8644 (5/14) (Fig. S3, S5, S7, S11 and S12, ESI†), SB431542 (4/14) (Fig. S2–S4 and S15, ESI†), and apigenin (4/14) (Fig. S5, S7, S10 and S15, ESI†).

Considering that major compounds act in different clusters that, in turn, are associated with different biological processes, we will discuss each small compound and its associated cluster when relevant.

Small compounds that induce iPSC reprogramming – parnate

Parnate is a monoamine oxidase inhibitor used as an antidepressive.[33] Studies have shown that parnate exhibits an inhibitory effect on lysine-specific demethylase 1 (LSD1/KDM1A) activity[33] and increases the reprogramming efficiency in human neonatal keratinocyte cells transduced with OCT4 and KLF4 and treated with the glycogen synthase kinase 3 beta (GSK3) inhibitor CHIR99021.[33] Inhibitors of the GSK-3β [*e.g.*, 6-bromoindirubin-3′-oxime (BIO) and CHIR99021] (Table 1) are also widely used for iPSC reprogramming. GSK-3β is inactivated upon Wnt signaling pathway induction, resulting in the accumulation of β-catenin and the activation of Wnt target genes.[34] This is an interesting procedure to generate iPSCs because the genes related to the canonical Wnt pathway are found to be

Table 1 List of all small molecules currently used alone or in combination for stem cell reprogramming. All of the molecules involved in promoting self-renewal, maintenance of the pluripotent state or other mechanisms that could improve reprogramming efficiency are indicated. Molecules that only promote de-differentiation towards a specific cell lineage were not included

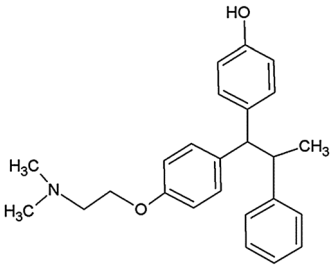
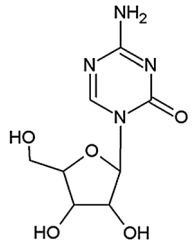
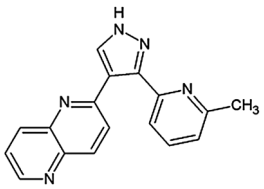
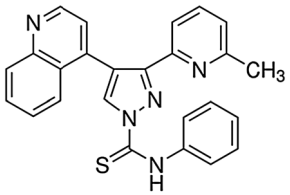
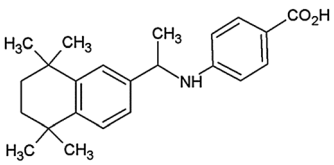
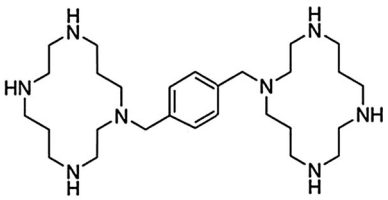
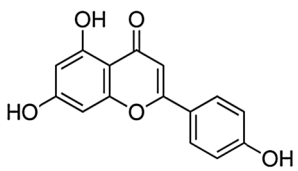
Small molecule structure[a]	Name	Identity	Function in SC reprogramming	Combinations[b]
	4-Hydroxytamoxifen (OHTM)	Estrogen receptor (ER) inhibitor	Stimulates Sox2 mRNA expression and can replace Sox2 activity in MEFs[15]	**Nabumetone, $NiSO_4$, Moclobemide, Corynanthine**
	5-Azacytidine (5′-azaC)	DNA methylation inhibitor	Enhances MEF reprogramming[92]	**Dexamethasone**
	616452	ALK5	Enables reprogramming using only *OCT4* when in combination with VPA, CHIR and Parnate[12]	**VPA, CHIR, Parnate**
	A-83-01	ALK4/5/7 inhibitor	Greatly enhances reprogramming efficiency.[93] Also described to promote alkaline phosphatase (ALP) expression in EpiSC supplemented with LIF, generating colonies similar to ESCs[37]	**PD0325901, CHIR, PD173074, NaB, PS48**
	AM580	Retinoic acid receptor (RAR) agonist	Enhances reprogramming in MEFs[94]	ND
	AMD3100	CXC4 antagonist	Described to mobilize hematopoietic stem cells[95]	ND
	Apigenin	Flavonoid that induces E-cadherin	Induces E-cadherin in MEFs without exogenous pluripotency factors. It also decreases Smad2 expression[49]	ND

Table 1 (*continued*)

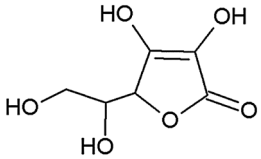
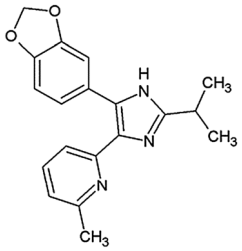
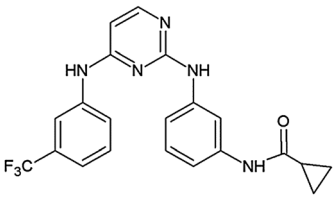
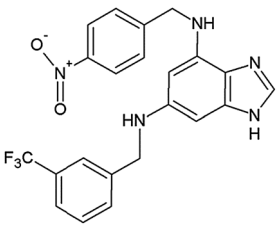
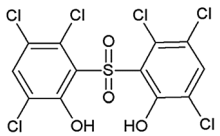
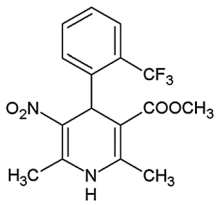
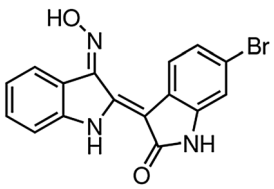
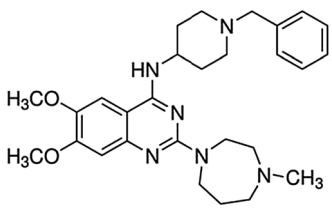
Small molecule structure[a]	Name	Identity	Function in SC reprogramming	Combinations[b]
	Ascorbic acid (AC)	Naturally occurring anti-oxidant (vitamin C)	Enhances reprogramming by diminishing ROS levels, induces the expression of OCT4 and AP and promotes changes in the DNA methylation status[67–69]	**VPA**
	B4	p38 MAPKα inhibitor	Enhanced reprogramming in MEFs[14]	**B8, B6, B4**
	B6	Aurora kinase inhibitor	Enhanced reprogramming in MEFs[14]	**B8, B6, B4**
	B8	Inositol-trisphosphate 3-kinase (IP3K) inhibitor	Enhanced reprogramming in MEFs[14]	**B8, B6, B4**
	B10	p38α and p38β inhibitor	Enhanced reprogramming in MEFs[14]	**B8, B6, B4**
	(+) Bayk8644 (BayK)	L-type calcium channel agonist	BayK alone does not induce self-renewal and pluripotency in MEFs, but when combined with BIX, it increases reprogramming efficacy[47]	**BIX**
	6-Bromoindirubin-3′-oxime (BIO)	GSK-3 inhibitor	BIO can promote the self-renewal of human and mouse ESCs,[34] expressing NANOG, OCT3 and OCT4. However, in HSCs, BIO can either promote the expansion of cell colonies or inhibit it[96]	ND
	BIX01294 (BIX)	G9 histone methyltransferase inhibitor	BIX can reprogram cells without OCT4 in MEFs and could also induce reprogramming in cells transduced with only Klf4 and Oct4[47]	**BayK, RG108**

Table 1 (*continued*)

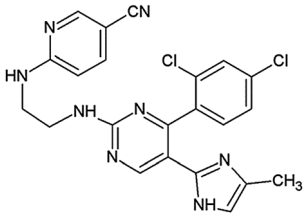
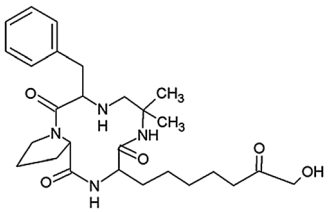
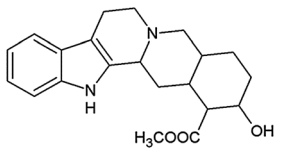
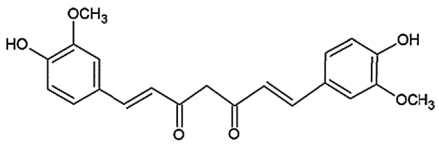
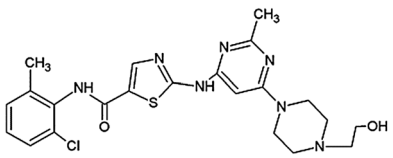
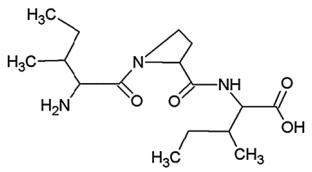
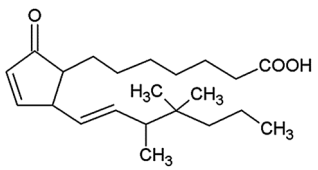
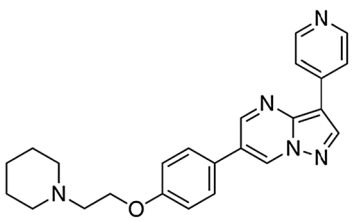
Small molecule structure[a]	Name	Identity	Function in SC reprogramming	Combinations[b]
	CHIR99021 (CHIR)	GSK-3 inhibitor	Induces reprogramming in MEFs and, when combined with Parnate, can induce reprogramming of human neonatal keratinocytes.[33] Can also promote self-renewal and expansion of ESCs[97]	**A-83-01, Parnate, PD0325901, SB431542, SU5402, PD184352**
	Chlamydocin	Histone deacetylase (HDAC) inhibitor	Promotes engraftment, expansion and maintains HSCs in the undifferentiated state[98]	ND
	Corynanthine	α1-Adrenergic and α2-adrenergic receptor antagonist	Enhanced reprogramming when combined with 4 other molecules and lectin[15]	**OHTM, Nabumetone, $NiSO_4$, Moclobemide**
	Curcumin	Naturally occurring anti-oxidant	Reported to enhance reprogramming efficiency[99]	ND
	Dasatinib	Src-family kinase inhibitor	Reported to induce reprogramming of MEFs.[100] However, it induced true successful reprogramming when combined with VPA	**VPA**
	Diprotin A	Dipeptidylpeptidase (DPP4/CD26) inhibitor	Increased stem/progenitor cell homing and HSC engraftment[16,101]	ND
	dmPGE2	PGE_2 receptor, activates the PGE_2 pathway	Reported to expand and modulate the formation of HSCs[102]	ND
	Dorsomorphin	ALK2/3/6 inhibitor	Can promote human ESC self-renewal by decreasing bone morphogenic protein (BMP) signaling[103] but it is described to induce neural differentiation in combination with SB431542 [104]	***SB431542***

Table 1 (*continued*)

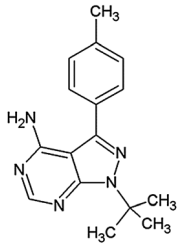
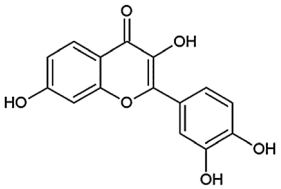
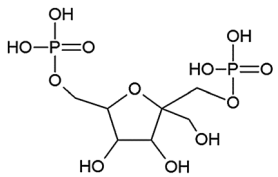
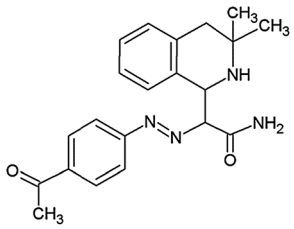
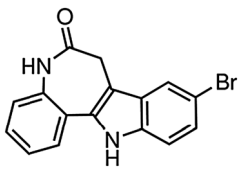
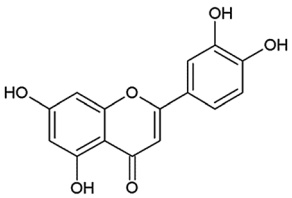
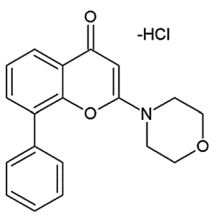
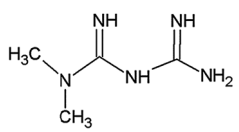
Small molecule structure[a]	Name	Identity	Function in SC reprogramming	Combinations[b]
	EI-275 (PP1)	Src-family kinase inhibitor	Reported to induce reprogramming of MEFs.[59,100] However, it induced true successful reprogramming when combined with VPA	**VPA**
	Fisetin	Sirtuin activator	Reported to enhance reprogramming efficiency[99]	ND
	Fructose-2,6-biphosphate (Fru-2,6-BiP)	Activates phosphofructokinase	Discussed to promote human primary cell reprogramming by modulating glycolysis[4,16]	ND
	IQ-1	Wnt-signaling pathway modulator	Interacts with PP2A to modulate the Wnt pathway. It is also described to promote proliferation and maintenance of the pluripotency state of ESCs without LIF. Together with Wnt3a, it can promote Oct4 and SOX2 expression[105]	ND
	Kenpaullone	GSK-3 and cyclin-dependent kinase (CDK) inhibitor	Described to substitute for Klf4 activity in MEFs. Enhanced reprogramming by expressing NANOG[106]	ND
	Luteolin	Flavonoid that modulates E-cadherin expression	Similar to apigenin, luteolin increased AP^+ colonies and increased E-cadherin expression in MEFs. It also decreases SMAD2 expression[49]	ND
	LY294002	IP3K inhibitor	Reported to enhance reprogramming efficiency.[99] However, it is also described to differentiate human ESCs[107]	ND
	Metformin	AMP-activated protein kinase (AMPK) agonist	Does not induce reprogramming but can reduce the teratomic potential of iPSCs by diminishing Survivin and OCT4 expression.[99,108] Nevertheless, iPSCs treated with metformin maintained their pluripotency	ND

Table 1 (*continued*)

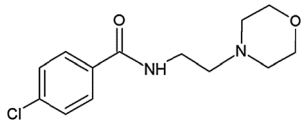
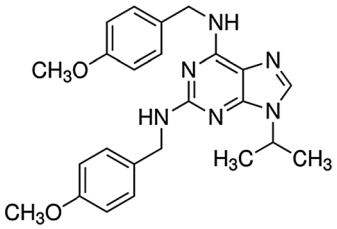
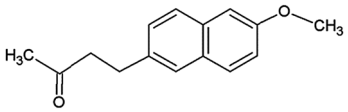
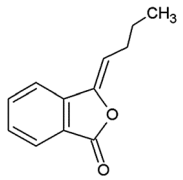
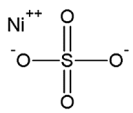
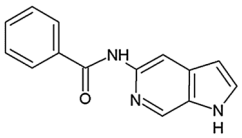
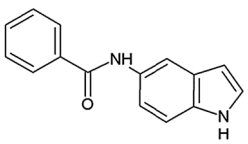
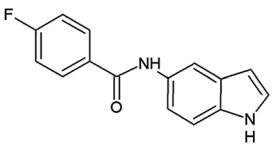
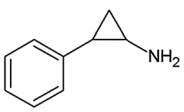
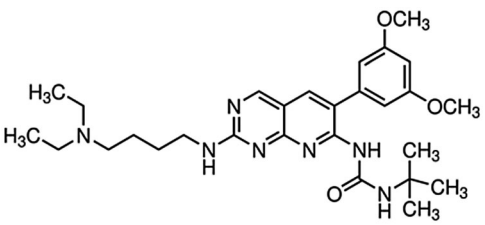
Small molecule structure[a]	Name	Identity	Function in SC reprogramming	Combinations[b]
	Moclobemide	Monoamine oxidase inhibitor	Enhanced reprogramming when combined with 4 other molecules and lectin[15]	**OHTM, Nabumetone, $NiSO_4$, Corynanthine**
	Myoseverin	Multinucleated myotube fission inducer	Myoseverin does not promote the reprogramming of terminally differentiated cells, but in muscle cells, it induces survival by promoting the expression of anti-apoptosis, antioxidant and detoxification genes[109]	ND
	Nabumetone	Non-steroidal anti-inflammatory drug (NSAID)	Targets COX2 and enhances reprogramming efficiency. Substitutes SOX2 activity in MEFs[15]	**OHTM, $NiSO_4$, Moclobemide, Corynanthine**
	*n*-Butylidenephthalide (BP)	Jak2-Stat3 pathway inducer	Reported to increase JAK2-STAT3 signaling and OCT4 and SOX2 expression in MEFs[110]	ND
	$NiSO_4$	Inorganic compound	Enhanced reprogramming when combined with 4 other molecules and lectin[15]	**OHTM, Nabumetone, Moclobemide, Corynanthine**
	Oct4-activating compound (OAC1)	OCT4 inducer	Enhances reprogramming by activating OCT4 and inducing the expression of NANOG and SOX2 expression in MEFs[111]	ND
	OAC2	OCT4 inducer	Enhances reprogramming by activating OCT4 expression in MEFs[111]	ND
	OAC3	OCT4 inducer	Enhances reprogramming by activating OCT4 expression in MEFs[111]	ND
	Parnate	Monoamine oxidase inhibitor	Increased the reprogramming efficiency in human neonatal keratinocyte cells only transduced with OCT4 and KLF4[33]	**CHIR, PD0325901, SB431542**
	PD173074	FGF signaling inhibitor	Expands mouse ESC colony and inhibits the growth of differentiated cells[37]	**CHIR, PD0325901, A-83-01**

Table 1 (*continued*)

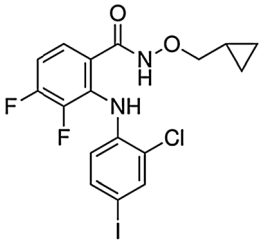
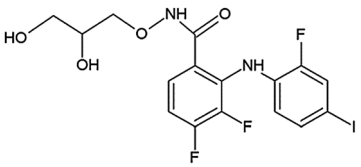
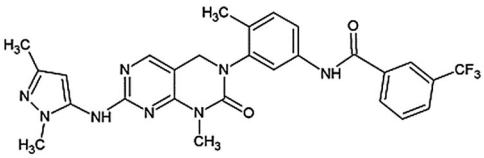
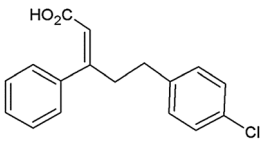
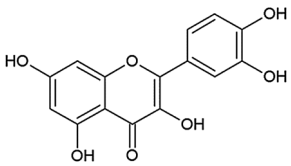
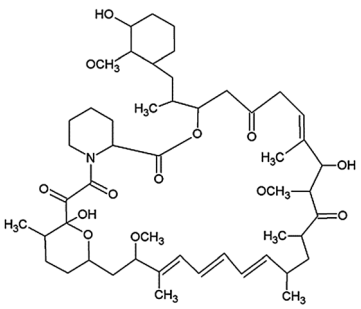
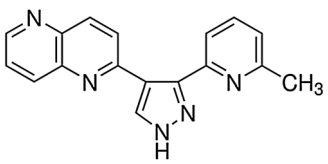
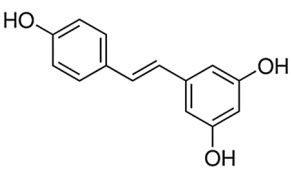
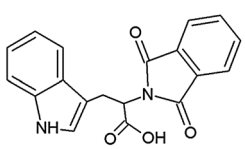
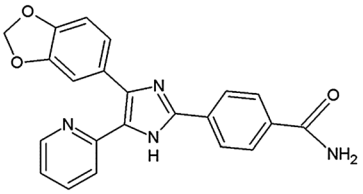
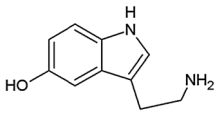
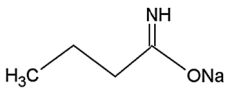
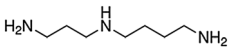
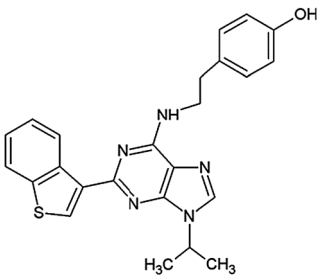
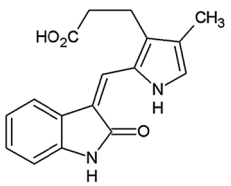
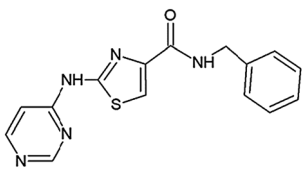
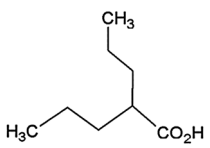
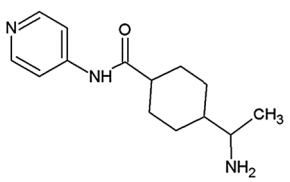
Small molecule structure[a]	Name	Identity	Function in SC reprogramming	Combinations[b]
	PD184352	MEK inhibitor	Can promote self-renewal and expansion of ESCs if combined with SU5402, but not alone[97]	**SU5402, CHIR**
	PD0325901	MEK inhibitor	Reported to enhance reprogramming in different studies[10,37,112]	**A-83-01, CHIR99021, SB431542, PS48, NaB**
	Pluripotin (SC1)	RasGAP and ERK1 dual inhibitor	Supports murine ESC self-renewal and blocks differentiation in the absence of LIF[113]	ND
	PS48	3′-Phosphoinositide-dependent kinase-1 (PDK1) activator	Promotes human keratinocyte reprogramming by shifting mitochondrial oxidation to glycolysis[10]	**A-83-01, NaB, PD0325901**
	Quercetin	HIF pathway activator	Reported to promote human primary cell reprogramming by modulating hypoxia[4,10]	ND
	Rapamycin	Mammalian target of Rapamycin (mTOR) pathway inhibitor	Enhances reprogramming in MEFs[99]	ND
	RepSox (E-616452)	ALK4/5/7 inhibitor	Can induce reprogramming in MEFs and replace SOX2 expression[59]	**VPA**
	Resveratrol	Sirtuin activator	Reported to enhance reprogramming efficiency[99]	ND
	RG108	DNA methyltransferase (DNMT) inhibitor	Reported to improve reprogramming in MEFs[47]	**BIX**

Table 1 (*continued*)

Small molecule structure[a]	Name	Identity	Function in SC reprogramming	Combinations[b]
	SB431542	ALK4/5/7 inhibitor	SB431542 increased reprogramming in mouse fibroblasts transduced with Klf4, OCT4 and c-MYC genes.[59] SB431542 was also efficient in generating iPSCs from human skeletal myoblasts.[41] However, SB431542 can induce the differentiation of human ESCs and iPSCs to neural lineages[60,61]	**VPA, PD0325901, *Dorsomorphin***
	Serotonin	Monoamine neurotransmitter	Induces *ex vivo* cord blood HSC expansion[114]	ND
	Sodium butyrate (NaB)	Histone deacetylase (HDAC) inhibitor	Enhances reprogramming capacity even when Klf4 or c-MYC is absent[10,40]	**PD0325901, PS58, A-83-01**
	Spermidine	Autophagy inducer	Reported to enhance reprogramming efficiency[99]	ND
	StemRegenin (SR1)	*Aryl hydrocarbon receptor* (AhR) antagonist	Promotes engraftment and expansion of human HSCs[115]	ND
	SU5402	FGF signaling inhibitor	Together with PD184352, it maintained the undifferentiated state of mouse iPSCs and can expand through multiple passages[97]	**PD184352, CHIR**
	Thiazovivin	Rho-associated coiled-coil containing protein kinase (ROCK) inhibitor	Improved survival through inhibition of ROCK and increased the expression of E-cadherin in human ESCs[116]	ND
	Valproic acid (VPA)	HDAC inhibitor	Described to enhance reprogramming in multiple experiments[59,67,70,100]	**RepSox, SB431542, EI-275, Dasatinib, Ac**
	Y-27632	ROCK inhibitor	Promotes the survival of human ESCs[116,117]	ND

[a] Molecular structures were drawn using the program ADC/ChemSketch 12 Freeware [http://www.acdlabs.com/resources/freeware/chemsketch/].
[b] Small molecules already tested in combination with a given molecule. The molecules names in bold font indicate synergistic effects (successful reprogramming when used together), whereas the molecules in bold italic font indicate that their combinations showed unwanted effects on iPSC generation. ND, not described.

expressed in undifferentiated cells.[34] However, a study shows that activation of the Wnt pathway, using inhibitors of GSK-3β, such as CHIR99021, can induce cardiomyocyte differentiation, a point that should be addressed in generating iPSC from cardiac tissues.[35] Additionally, suppression of the canonical Wnt signaling was also related to increased self-renewal of human ESC.[36]

The increased efficiency in iPSC reprogramming by parnate may be due to the inhibition of H3K4.[33] Interestingly, low H3K4 trimethylation may result in the lack of binding of OCT4, SOX2 and KLF4 in their target genes,[33] thus blocking cell reprogramming. Parnate has also been described to induce the reprogramming of cells derived from the post-implantation stage epiblasts (EpiSC) back to similar state of murine embryonic stem cells (ESC)[37] in a LIF-supplemented medium.

Small compounds that induce iPSC reprogramming – valproic acid and sodium butyrate

The gene ontology (GO) analysis of the sr-CPI–PPI network allows the identification of proteins related to oxidative demethylation processes, such as the alkylation repair homologs 2 and 3 (ALKBH2-3) (Table S1, see ESI†).

ALKBH1 is a histone dioxygenase that acts specifically on histone H2A.[38] Murine *ALKBH1*$^{-/-}$ embryonic stem cells (ESCs) achieved a prolonged expression of the genes involved in pluripotency maintenance, such as *NANOG* and *OCT4*, and presented delayed differentiation.[38] These changes in stem cell markers may be due to the interaction of ALKBH1 with the histone demethylase KDM2B, which cooperates with OCT4 in cell fate determination.[38] In addition, ALKBH1 interacts with Mrj (DNAJB6) in trophoblastic cells.[39] Interestingly, Mrj mediates gene repression by recruiting class II HDACs,[39] blocking cell reprogramming. However, the full role of ALKBH1 in iPSC generation is not clear, and further investigations of the role of the ALKBH family in cell reprogramming would be of great interest (Table 2).

Considering HDACs as targets to be inhibited for enhanced reprogramming, many inhibitors, including valproic acid (VPA) and sodium butyrate (NaB), have been designed and tested to inhibit HDACs. Valproic acid was described to enhance reprogramming,[9,16] as well as NaB, a naturally occurring fatty acid.[16,40] The effects of NaB in iPSC generation appear to be positively or negatively modulated when in combination with histone acetyltransferase (HAT) or HDAC inhibitors. For example, mesenchymal stem cells (MSCs) treated with C646 (a HAT inhibitor) diminished the basal reprogramming efficiency and the reprogramming effect of NaB.[40] In the same study, it was observed that the reprogramming capacity of NaB is enhanced when *KLF4* or *c-MYC* is absent and when the small molecule BIX01294 (BIX), a G9a histone methyltransferase inhibitor, is present.[40]

NaB was also reported to efficiently reprogram human skeletal fibroblasts, especially when combined with SB431542, a small compound that inhibits activin receptor-like kinase 4/5/7 (ALK4/5/7) in feeder-free conditions.[41] NaB also showed high reprogramming efficiency when combined with SB431542 and a MEK-ERK inhibitor (PD0325901).[42]

Considering the sr-CPI–PPI network (Fig. 1A), NaB was found to be connected to proteins related to the apoptotic pathway, such as BAX, BCL2 and CASP3. In this sense, NaB induces the following: (i) the activation of CASP3 and the death-associated protein kinase (DAPK1/2) in human gastric cancer cells;[43] (ii) CASP7 expression followed by underexpression of the BCL2 and BCLX genes in prostate cancer cells;[44] (iii) cell cycle arrest, CASP3 cleavage, NOTCH1 signaling and apoptosis in rat pheochromocytoma cells;[45] and (iv) CASP3 and CASP9 activity in human colon carcinoma cells.[46] In addition, the apoptotic effect of NaB in those studies indicates that its effects are dose-dependent. NaB appears to be more efficient at low doses to promote cell reprogramming, and appears to be more efficient in combination with other small molecules.

Small compounds that induce iPSC reprogramming – Bayk8644

Bayk8644 (BayK) is an L-type calcium channel agonist[1,4,9,16] that does not exert direct effects on DNA structure or induce histone modifications; however, it seems to act at a cell transduction level by promoting signaling cascades that result in post-translational modifications and gene expression events.[47]

When OCT4/KLF4 transduced-mouse embryonic fibroblasts (MEFs) were treated with BayK, no visible reprogramming was detected.[47] However, the combined use of BayK with BIX led to an increase in the number of colonies and their sizes.[47] A possible biochemical mechanism that correlates with the combined effect of BayK with BIX in inducing cell reprogramming is CREB activation through intracellular Ca^{2+} influx, where the CREB binding protein (CREBBP) is present in the sr-CPI–PPI network (Fig. 1A). Interestingly, CREBBP$^{(+/-)}$ mice were unable to maintain the self-renewal of hematopoietic stem cells (HSCs).[48] Remarkably, the process of Ca^{2+} homeostasis appears to be a significant biological process associated with the sr-CPI–PPI network, indicating that further studies on proteins related to Ca^{2+} homeostasis could be of interest in iPSC reprogramming (Table S1, see ESI†).

Small compounds that induce iPSC reprogramming – apigenin

Another small molecule that had a high occurrence among the clusters was apigenin (Fig. S7, S10 and S15, ESI†). Apigenin is a natural flavone compound that induces E-cadherin expression in MEF cells, a hallmark of reprogramming, even if these cells were not previously transduced with reprogramming factors.[49] However, apigenin does not increase iPSC proliferation.[49]

Apigenin has also been associated with an anti-apoptotic effect on rat kidney cells by lowering the expression of TGFβ.[50] Small molecules that inhibit the TGF signaling pathway (*e.g.*, SB431542, RepSox, and A-83-01) are consolidated as successful reprogramming compounds (Table 1).[1,4,9,16] The TGFβ signaling pathway is responsible for the epithelial-to-mesenchymal transition (EMT), a process that gives rise to all mesenchymal cells in early embryogenesis.[51] During iPSC generation, cells undergo the mesenchymal-to-epithelial transition (MET), a biological process that leads to pluripotency.[51] Therefore, inhibition of

Table 2 Potential targets to be inhibited by small molecules for enhanced reprogramming selected from hub-bottlenecks (HBs), switches (SW) and clusters present in the sr-CPI–PPI network

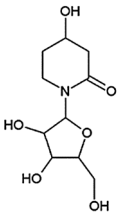
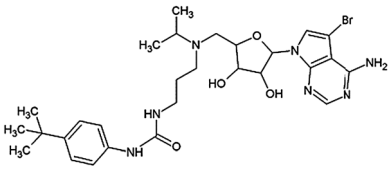
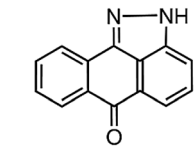
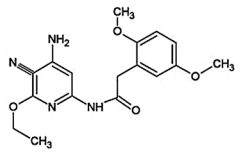
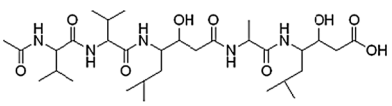
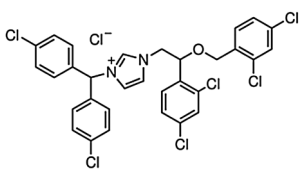
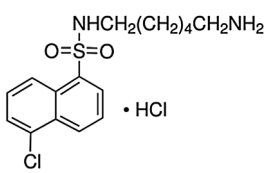
Protein	Function	Inhibitors	Notes	Ref.
E2F1	Transcription factor	 Tetrahydrouridine	Upregulated by p16. It also downregulates *AUF1*. Downregulation of NANOG causes E2F1 expression.	118 and 119
p16 (INK4a)	Cyclin-dependent kinase	ND[a]	Inhibition of p16 causes an increase in *NANOG*, *SOX2* and *OCT4* expression in breast cancer. Suppression of the INK4a/AFR locus also increases reprogramming. p16 also upregulates E2F1.	120–122
CTGF	Growth factor	Curcumin (Table 1)	Differentiates MSCs into fibroblast cells.	81, 123 and 124
DOT1L	Histone methyltransferase	 SGC0946	Required for proper proliferation after differentiation but is not required for ESC viability	88 and 89
JNK1	Serine/threonine-protein kinase	 SP600125  JNK Inhibitor VIII	Not required for embryonic stem cell self-renewal or proliferation but is required for the differentiation of embryonic stem cells.	125–128
VIP	Vasoactive intestinal peptide	 Acetyl-Pepstatin	Found to induce electrophysiological activity in differentiating embryonic stem cells.	129
CALM1	Kinase	 Calmidazolium  W7	Wnt5/calmodulin signaling pathway activates the differentiation of mesenchymal progenitor cells in mesenchymal stromal cells.	130 and 131

Table 2 (*continued*)

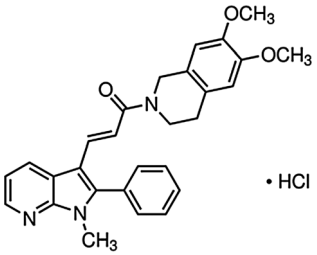
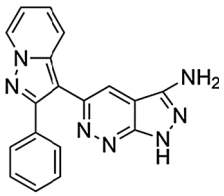
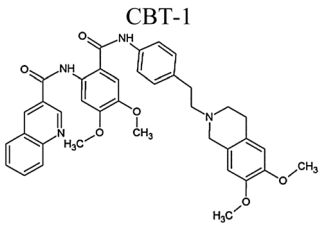
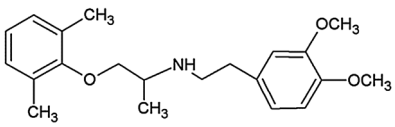
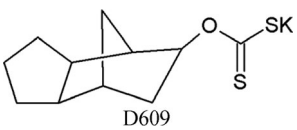
Protein	Function	Inhibitors	Notes	Ref.
SMAD2/3	Transcriptional modulators	 SIS3 (SMAD3)	Related to the induction of EMT.	51, 132 and 133
ERK2	Serine/threonine kinase	 FR180204	Inhibits KLF4 in mouse ESCs together with ERK1, promoting differentiation.	134
ALKBH1/2/3	Dioxygenases	ND	Mouse *Alkbh1*$^{-/-}$ ESCs showed prolonged expression of pluripotency maintenance genes and delayed differentiation	38
LATS1/2	Kinases	ND	LATS1 and LATS2 can decrease the transcriptional activity of TAZ, which increased self-renewal in hESCs	52 and 54
ABCB1(MDR)	P-glycoprotein	CBT-1  Tariquidar	Known to induce multi-drug resistance (MDR) through the regulation of the intracellular drug concentration.	74 and 76
		 1416		
KGN1	Inhibitor of thiol protease	ND	In lung fibroblasts, KNG1 is able to induce TGFβ1 expression and cause the cell to differentiate into a myoblast. KGN1 is also correlated with TGFβ1 secretion in adipose tissue-derived MSC	77 and 78
NCOA3	G-protein	ND	NCOA3 is essential for mouse ESC pluripotency maintenance	79
FOS (c-FOS)	Nuclear phosphoprotein	 D609	FOS protein also shows a synergistic effect with the NF-kB inhibitor (PDTC) in mouse ESCs to promote ERK1/2 signaling and, therefore, differentiation	80 and 135
FAM38A (P38IP)	p38 interacting protein	B10 (for p38α/β) (Table 1)	FAM48A is related to EMT induction by downregulating E-cadherin during gastrulation of mouse embryos	14 and 82
FOXJ3	Transcription factor	ND	Induces neural differentiation	83
DSCAM	Immunoglobulin	ND	Induces neural differentiation	84

Table 2 (*continued*)

Protein	Function	Inhibitors	Notes	Ref.
S100A2	Ca^{2+} binding protein	ND	Is a direct target of p53 during keratinocyte differentiation	85
XPO4	Exportin	ND	Transports SMAD3 to the cytoplasm, where it can play a role in the TGFβ signaling pathway	90
XPO1	Exportin	ND	Related to export proteins of the MAPK family	91

[a] ND, not determined.

genes and proteins related to EMT, such as the SMADs and TGF-associated proteins, is currently being pursued.

Considering that apigenin inhibits the TGFβ pathway, the EMT process could be consequently stimulated, a process already observed in our GO analysis (Table S1, see ESI†). In our GO analysis, we identified the presence of tafazzin (TAZ), a protein regulated by the Hippo pathway, which influences proliferation and apoptosis in various tissues. In the Hippo pathway, the activation of two kinases, LATS1 and LATS2, can decrease TAZ transcriptional activity.[52] TAZ appears to induce connective tissue growth factor (CTGF) protein expression by its interaction with the transcription factor TEA domain family member 1 (TEAD).[53]

Interestingly, a study has shown that synthesized hydrogels made from polyacrylamide displaying glycosaminoglycans (GAGs) increased iPSC formation by stimulating YAP/TAZ activation.[54] GAGs include cell-surface polysaccharides that were already linked to reprogramming.[54]

If the main role of TAZ is in promoting self-renewal, rather than simply inducing differentiation through CTGF expression, small molecules that target LATSs could be of interest for stem cell maintenance (Table 2). It is also interesting to observe that apigenin is a potent neurogenic inducer, promoting adult neural stem cell differentiation in mice.[55,56] Similarly to NaB, low levels of apigenin showed non-toxic effects on MSC, but in high concentrations, apigenin enhances apoptosis in cells treated with LPS.[57]

Small compounds that induce iPSC reprogramming – SB431542

SB431542 is a potent molecule inhibitor of TGFβ signaling, acting on activin receptor-like kinase (ALK5) (TGFβ type I receptor), activin type I receptor (ALK4) and the nodal type I receptor (ALK7).[58] SB431542 has already been related to increased reprogramming in mouse fibroblasts transduced with the *KLF4*, *OCT4* and *c-MYC* genes, indicating that it can partially replace *SOX2* activity by inhibiting TGFβ signaling.[59] SB431542 was also efficient in generating iPSCs from human skeletal myoblasts.[41] However, high doses of SB431542 can induce the differentiation of human ESC and iPSC to neural lineages.[60,61]

sr-CPI–PP network centrality analysis

A centrality analysis was performed in the sr-CPI–PPI, and each node of the network was evaluated in terms of node degree, betweenness, and closeness to establish the most topologically relevant nodes (proteins/small molecules) within the network. This type of analysis allows us to observe which proteins/small compounds could be a relevant target for iPSC reprogramming.

In general terms, closeness analysis indicates the probability that a protein/small molecule is relevant to other proteins/small molecules (nodes) in a signaling network.[30] In turn, betweenness indicates the number of shortest paths that go through each node. Thus, nodes with higher betweenness scores, when compared to the average betweenness score of the network, are responsible for controlling the flow of information through the network, thus characterizing the so-called bottleneck nodes. Finally, node degree is a measure that indicates the number of connections that involve a specific node. Nodes with a high node degree are called hubs[62] and have key regulatory functions in the cell. Finally, the network was scanned for hub-bottleneck (HB) nodes, which combine the bottleneck function (controlling the information flow within the network) and the hub property (number of connections above the average node degree value calculated for the network) into a scale-free biological network, the most important nodes of which are the so-called hub-bottlenecks.[62] Hence, the most relevant node for a determined biochemical pathway or module can be obtained and further analyzed.

In this sense, we created two centrality datasets with different aims. First, we analyzed the sr-CPI–PPI network for HBs. We obtained 124 HB nodes, of which 6 were small molecules (Fig. 1B) [4-hydroxytamoxifen (OHTM), quercetin, ascorbic acid (AC), BayK, VPA and SB431542]. Moreover, we performed a centrality analysis to see which nodes had the highest closeness, betweenness and node degree scores when compared to the average score of each of the selected properties in the network. These nodes were given the names ''switches'' (SWs) (Fig. 1C). This last analysis differs from the HBs because it contains the values of closeness as well. HB proteins have a great number of connections and control the flow of information in the network, but they do not necessarily impact the activity of other proteins.[29] Thus, the activities of proteins with a high closeness value are more likely to impact the function of other proteins.

In this sense, we identified 108 SWs, among which the 6 small molecules (BayK, AC, VPA, SB431542, quercetin and OHTM) gave identical results to those observed for HBs (Fig. 1C).

These results indicate that BayK, AC, VPA, SB431542, quercetin and OHTM (Table 4) should be considered for the formulation of

a reprogramming cocktail, by taking into account the biological processes affected by these compounds.

Potential drugs for iPSC generation when in combination – quercetin

Little is known about the role of quercetin in stem cell reprogramming, although it has been discussed that this small molecule can stimulate glycolysis and enhance cell reprogramming.[14] The use of quercetin alone has been associated with the induction of apoptosis in cancer cells.[63] In addition, cancer stem cells treated with quercetin showed a loss of self-renewal and a loss of stem-like characteristics.[64,65] Interestingly, a decreased expression of OCT4 and NANOG was observed in head- and neck-derived cancer stem cells.[64] It was also found that quercetin inhibited proliferation and promoted the osteogenic differentiation of human adipose tissue-derived stromal cells.[66]

Although quercetin was previously discussed as a reprogramming agent in only one work,[4] and in the sr-CPI–PPI network, it is present in 3/16 clusters (Fig. S2, S3 and S12, see ESI†) with high centrality values, the use of quercetin combined with other reprogramming molecules has not been described until now.

Potential drugs for iPSC generation when in combination – ascorbic acid (vitamin C)

It was described that ascorbic acid (AC) itself appears to greatly enhance reprogramming in MEFs.[67] AC is a naturally occurring anti-oxidant, and during MEF reprogramming (previously transfected with Sox2, Oct4 and Klf4), a lowering in reactive oxygen species (ROS) formation and the promotion of OCT4-GFP$^+$ colonies were observed.[67] ROS formation can lead to cell senescence during reprogramming and constitutes one of the "reprogramming barriers" that should be overcome during iPSC generation. AC also increased the OCT4-GFP$^+$ colonies more efficiently than VPA;[67] the authors also describe that VPA and AC act *via* different mechanisms and have a synergistic effect on increasing OCT4-GFP$^+$ colonies. Moreover, AC can promote DNA demethylation and enhance reprogramming in human ESC.[68] The mechanisms underlying the role of AC in reprogramming have already been reviewed.[69]

Although the role of AC in enhancing reprogramming does not depend only on its anti-oxidative properties,[67] ROS formation can lead to precise cell senescence in iPSC generation, and anti-oxidant enzymes should be taken into account to enhance cell reprogramming. In this sense, responses to oxidative stress and senescence processes were observed in the GO analysis of the sr-CPI–PPI network (Fig. 1A) (Table S1, see ESI†). In addition, it was observed that MAP2K1 (MEK1) connects both oxidative stress and senescence processes.

MEK/ERK inhibitors (*e.g.*, PD184352, PD0325901, and pluripotin) and FGF inhibitors (PD173074 and SU5402) are also commonly used molecules for iPSC reprogramming that are known to inhibit the mitogen-activated protein kinases (MAPK), the extracellular signal-regulated kinases (ERK) and fibroblast growth factor signaling pathways, thus promoting the differentiation of embryonic stem cells.[4,33]

In this sense, PD0325901 is efficient in inducing reprogramming;[4] the addition of this drug together with AC may be beneficial for cellular reprogramming. PD0325901 is present in only 2 clusters of the sr-CPI–PPI network (Fig. S12 and S13, see ESI†), but when closeness analysis was performed alone, it appeared as one of the molecules with above average closeness, together with nabumetone (Table 1) and apigenin (Table 4), indicating that it has a strong action in influencing the activity of other nodes.

Potential drugs for iPSC generation when in combination – 4-hydroxytamoxifen (OHTM)

Another interesting small molecule found in the HB and SW results was 4-hydroxytamoxifen (OHTM) (Fig. 1B and C). There is only one study that shows the role of OHTM in cell reprogramming.[15] However, the results are particularly promising. MEFs transfected with OCT4, KLF4 and c-MYC were treated with OHTM and showed high reprogramming efficiency, indicating that OHTM can replace SOX2 activity through an unknown mechanism.[15] The same authors also observed that OHTM can stimulate endogenous SOX2 mRNA expression.

OHTM is used in chemotherapy for breast cancer treatment by blocking the estrogen receptor (ER) *via* its metabolites, but the authors did not observe ER expression in MEFs.[15] In our GO analysis, we found the response to estrogen stimulus among the bioprocess listed (Table S1, see ESI†). Although ER is not expressed in MEFs, it is expressed in a broad range of somatic tissues [www.genecards.org]; thus, the use of OHTM should not be discarded. In the GO analysis, we also observed that featuring among the proteins in the response to estrogen stimulus are many TGFs (TGFBR1, TGFB3, TGFB1). It is possible that OHTM could regulate reprogramming through its role in TGF signaling.

Potential drugs for iPSC generation when in combination – valproic acid and lithium

VPA is widely used in stem cell reprogramming[1,4,9] and can enhance reprogramming in primary human fibroblasts transfected with *OCT4*, *KLF4* and *SOX2*, as well as those transfected with only *OCT4* and *SOX2*.[70] Interestingly, VPA is applied as an anti-psychotic drug like lithium. Lithium is used as an anti-psychotic drug,[71] and an interesting study showed that VPA together with lithium can delay the differentiation and increase the self-renewal capacity of mice HSCs.[72] In addition, it was observed that lithium alone can enhance reprogramming in MEFs, but not cell proliferation.[71] The authors also report that lithium doubled the reprogramming efficiency in OCT4-infected human umbilical vein endothelial cells that were treated with the small molecules A83-1 (present in our CPI–PPI network), NaB, PS48, CHIR99021 and PD0325901. Lithium has also been reported to be a GSK3β inhibitor,[73] indicating its potential role in reprogramming. Thus, the use of lithium in reprogramming cocktails appears to be promising due to its synergistic effect with other drugs.

Potential drugs for iPSC generation when in combination – ABCB1 inhibitors and bradykinin

In the HB and SW networks (Fig. 1B and C), we observed the ATP-binding cassette, sub-family B, member 1 (MDR/ABCB1)

and bradykinin (KNG1), an inhibitor of thiol proteases. Remarkably, the ATP-binding cassette (ABC) family is known to induce multi-drug resistance (MDR) through the regulation of intracellular drug concentrations.[74] ABCB1 is reported to detoxify antitumoral drugs from tumoral cells, including HDAC inhibitors.[74]

The inhibition of ABCB1 protein by small compounds, such as CBT-1 and tariquidar, is related to a lower efflux of antitumoral drugs.[74] Interestingly, aggressive tumors and embryonic cells share many similarities concerning their proliferative capacity and the expression of genes related to pluripotency maintenance, and many pathways appear to converge between the two cell types.[75] Thus, the inhibition of ABCB1 (Table 2) could be important in overcoming the multi-drug resistance that could be created by the combined use of small molecules in reprogramming.

Despite the beneficial effect of inhibiting ABCB1 proteins to lower multi-drug resistance, ABCB1 is also associated with important processes related to cell survival, such as the activation of anti-apoptotic pathways and DNA repair.[76] However, another drug named 1416 displayed a successful inhibitory effect on ABCB1 proteins, blocking the drug efflux but not ABCB1 gene expression.[76] ABCB1 is expressed in a broad range of somatic tissues [http://www.genecards.org], and the role of its inhibition in iPSC generation is promising and has not yet been explored.

KNG1 is a member of the kinin peptide family and participates in various biological processes, such as inflammation, muscle contraction and vasodilatation.[77] In lung fibroblasts, KNG1 is able to induce TGFβ1 expression and cause the cell to differentiate into a myoblast.[78] Interestingly, KGN1 is also correlated with TGFβ1 secretion in adipose tissue-derived MSCs.[77]

Moreover, we also conducted a careful search in the literature for each of the HBs, SWs and the proteins in each cluster to understand their probable roles in stem cell physiology and reprogramming. We selected the more relevant targets plus the ones already discussed in this work and divided them into protein/genes "to be inhibited" and protein/genes "to be expressed" (Table 2 and Table 3), respectively. We also summarized our topological results for the main small molecules (Table 4). The further discussed targets that were observed in the interference analysis are also included in Table 2. Due to the fact that some proteins have more specific roles in stem cell differentiation (*e.g.*, differentiation into bone tissues or neurogenic potential), we generated another dataset, containing targets to be inhibited in different types of somatic tissue for more specific reprogramming (Table S2, see ESI†).

sr-CPI–PP network merged data

Our systems chemo-biology data discussed so far seem to indicate that those four molecules (BayK, OHTM, quercetin, and SB431542) are the most influential drugs in the reprogramming network. Therefore, we merged the four small molecule networks (Fig. 2) to see what common mechanisms they may share. The merged network, termed the BOQS-network (Fig. 2A), displayed 41 HBs (Fig. 2B).

Both SB431542 and OHTM share the nuclear co-activator 3 (NCOA3). Consistent with the importance of HBs in our analysis, a recent study showed that NCOA3 is essential for maintaining mouse ESC pluripotency through binding to the *Nanog* (but not *OCT4* or *SOX2*) promoter, which increases the levels of histone

Table 3 Potential targets to be expressed or used for enhanced reprogramming, selected from the hub-bottlenecks (HBs), switches (SW) and clusters present in the sr-CPI–PPI network

Protein	Function	Notes	Ref.
AUF1 (HNRNPD)	Nuclear ribonucleoprotein	AUF1 is downregulated by p16 when p16 induces E2F1 (see Table 2)	118
IL6	Cytokine	Supplementation of serum-free medium with IL6 and IL6R chimera (IL6RIL6) was related to increased pluripotent state in ESCs. IL6 also induces the expression of STAT3, which was expressed in equine iPSCs in the absence of c-MYC	136–138
CDK4	Cyclin dependent kinase	Is expressed by LIN28, which also appears to promote CDK4 translation	139
BIRC5 (Survivin)	Inhibitor of apoptosis	Involved in murine ESC survival. Is induced by OCT4	140
PARP1	Poly(ADP-ribose) polymerase	PARP1 expression causes an increase in SOX2 expression by inhibiting *FGF4*	141
CDK2	Cyclin dependent kinase	Involved in the maintenance and survival of pluripotent stem cells	142 and 143
NOV	Insulin-like growth factor	Involved in the regulation of the undifferentiated state of human HSCs	144

Table 4 Small molecules that showed the most preeminent topological results in our sr-CPI–PPI network

Small molecules	SW	HB	Closeness	Closeness *versus* betweenness	Clustering occurrence	Pathway affected
Parnate	No	No	No	No	5/14	Methylation and MAO
Sodium butyrate (NaB)	No	No	No	Yes	5/14	Histone modification
Bayk8644 (BayK)	Yes	Yes	No	No	5/14	Calcium signaling
Apigenin	No	No	Yes	No	4/14	E-cadherin
SB431542	Yes	Yes	No	No	4/14	TGF
4-Hydroxytamoxifen (OHTM)	Yes	Yes	No	No	3/14	COX2
Quercetin	Yes	Yes	No	No	3/14	Glycolytic pathway
Ascorbic acid (AC)	Yes	Yes	No	No	2/14	DNA methylation and antioxidant
Valproic acid (VPA)	Yes	Yes	No	No	2/14	HDAC

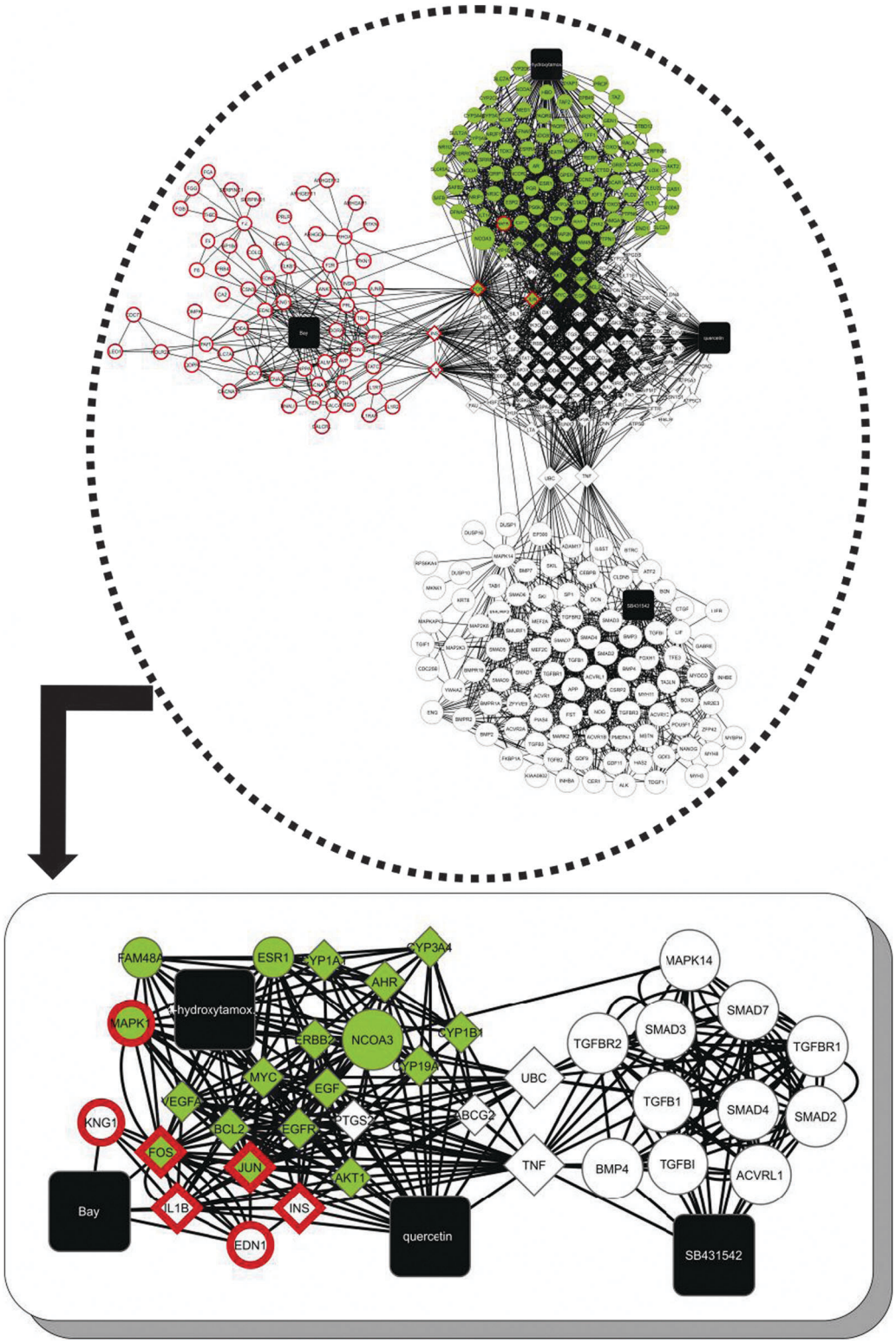


Fig. 2 BOQS (BayK, OHTM, Quercetin and SB431542) network merged data. (A) Red bordered nodes (Bayk8644); green nodes (Quercetin); diamond shaped nodes (4-hydroxytamoxifen); and large nodes (SB431542). The merged network is composed of 339 nodes and 2901 edges. (B) HBs related to the merged networks.

acetylation and histone arginine methylation by recruiting CBP and CARM1.[79] The same authors also showed that the downregulation of *NCOA3* impairs *SOX2*, *OCT4* and *NANOG* expression and upregulates the differentiation markers, such as *NESTIN* (ectoderm), *SOX17* (endoderm), *HAND1*, *NKX2.5* and *T* (mesoderm), and *Cdx2* (trophectoderm).[79] The action of *NCOA3* seems to be dependent on

LIF activity; however, the combined use of PD0325901 and CHIR99021 appears to upregulate *Ncoa3* expression, even in the presence of LIF.[79]

In addition to NCOA3, FOS and JUN proteins were found to be present in three simultaneous subnetworks (Fig. 2A). Differentiation was shown to be promoted in mouse ESC when NF-κB proteins were inhibited by pyrrolidine dithiocarbamate (PDTC), promoting ERK1/2 signaling and, therefore, differentiation.[80] In this situation, the authors described that FOS protein was also present. This relation suggests that the presence of FOS and the inhibition of NF-κB are inadequate for cellular reprogramming. This result is consistent with a report that indicated that the NF-κB pathway may be required to enhance reprogramming,[81] whereas Toll-like receptor 3 (TLR3) is required for the induction of pluripotency genes by altering their chromatin structure to a more open state near the promoters of OCT4 and SOX2. The same authors also discuss that any of the five classes of pathogen recognition receptors have been shown to signal in ways similar to TLR3, where TLR3 may engage NF-κB by TRIF, when the latest is induced by retroviral RNA or mRNA. Unfortunately, the roles of FOS (Table 2) and JUN proteins in iPSC reprogramming are not clear.

Another protein that featured among the HBs is FAM48A (P38IP) (Table 2) that is connected to OHTM (Fig. 1B). Remarkably, FAM48A is related to EMT induction by downregulating E-cadherin during the gastrulation of mouse embryos.[82]

Consistent with our results, the inhibition of p38 by p38α and the p38β inhibitor (named B10) enhanced the reprogramming capacity of MEFs.[14] The authors also describe that B10 has synergistic effects when combined with the inhibitors of IP3K (B8), the aurora kinase inhibitor (B6) and the activin receptor-like kinase 4 and p38 MAPKα inhibitor (B4). AURKA and AURKB are present in the sr-CPI–PPI network (Fig. 1A). It is interesting to highlight that, in the same study, AURKA was shown to enhance the AKT1-mediated inhibition of GSK3B. AKT1 is present in the HB from the BQOS-network (Fig. 2B).

Another important protein target is prostaglandin synthase 2 (PTGS2/COX2), a HB protein present in the BOQS-network (Fig. 2B). Nabumetone is a non-steroidal anti-inflammatory drug (NSAID) inhibitor of COX2, which has been previously reported to greatly enhance the reprogramming of MEFs expressing OCT4, SOX2 and KLF4, suggesting that it can replace c-MYC function.[15] The authors also showed that nabumetone efficiently substitutes for *SOX2* function during the reprogramming of MEFs. However, COX1 inhibitors, such as indomethacin, did not display any significant improvement in reprogramming efficiency.[15] The authors report that Celecoxib and NS-398, both COX2 selective inhibitors, showed similar effects on iPSC generation. Given the role of COX2 inhibition in iPSC generation, it would be interesting to see if other NSAIDs that target COX2 could also affect reprogramming, such as Diclofenac (which selectivity targets COX2).

Interference analysis of the sr-CPI–PPI network

To identify new potential targets to be inhibited for more successful reprogramming, we developed individual networks for each transcription factor (Fig. S16, see ESI†) and used the program Interference to produce virtual knockout events. The betweenness centrality index was chosen to be recalculated after each node deletion. In this sense, there are two types of interference: (i) positive interference ($-n$), which means that the absence of a given node decreases the centrality value of the target node. Thus, the betweenness score for target nodes is affected by the absence of the given node; and (ii) negative interference (n), which indicates that the knockout of a given node increases the centrality value of the target node [http://www.cbmc.it/~scardonig/interference/Interference.php]. Thus, the interference is positive for a target node when a given node is not present.

We focused on the scores of negative interference because those values indicate proteins that could downregulate the pluripotency-associated factors and could, therefore, be inhibited by small molecules.

Interestingly, among the proteins that interfered with OCT4 (Table S3, see ESI†), we found FOXJ3, a member of the forkhead transcription factors.[83] FOXJ3 is expressed during embryogenesis during neural tube development and throughout the neuroectoderm.[83] It could be interesting to understand the role of FOXJ3 during the reprogramming of neural cells. Additionally, another protein, the immunoglobulin DSCAM, is also related to neural development[84] and has been found to interfere with OCT4 (Table S3, see ESI†). Moreover, we also identified the calcium binding protein S100A2 (Table S3, see ESI†). Remarkably, S100A2 appears to be a direct transcriptional target of p53 during keratinocyte differentiation.[85] Because p53 inhibition is already correlated with successful reprogramming,[86] S100A2 could play an important role in iPSC generation.

To corroborate the importance and the potential application of the interference analysis in identifying prospective new targets in a PPI network, one of the proteins that showed high interference (58%) (Table S3, see ESI†) with OCT4 was the chicken ovalbumin upstream promoter transcription factor 2 (NR2F2, also known as COUP-TFII). COUP-TFII is already correlated with increased reprogramming because NR2F2 is inhibited by mir-302, a micro-RNA involved in inducing endogenous OCT4 expression and promoting embryonic self-renewal and pluripotency.[87] Furthermore, a study has shown that the absence of DOT1L, a histone methyltransferase, is related to impaired proliferation in mouse ESCs right after differentiation; however, it is not required for ESC viability.[88] Because DNA methylation inhibitors, such as 5′-azaC, and other compounds that change the methylation status, such as AC, are already employed for iPSC generation, the use of small molecules that can promote the inhibition of DOT1L, such as SGC0946 [89] (Table 2), could be of interest.

Finally, the proteins that caused negative interference in the SOX2 network were both exportins (XPO1 and XPO4) (Table 2 and Table S3, see ESI†). Although these proteins were not correlated with reprogramming or inducing pluripotency, XPO4 has been described to export SMAD3 to the cytoplasm, where SMAD3 can play a role in the TGFβ signaling pathway.[90] This finding is interesting because both SMADs and the TGF signaling pathway are related to the EMT process and should be inhibited for successful reprogramming. Moreover, XPO1 has

been described to export proteins of the MAPK family,[91] which are also related to blocking reprogramming. The evaluation of XPO4 or XPO1 inhibitors might be of interest for iPSC generation.

Conclusions

Using chemo-systems biology tools, we generated PPI and CPI networks and applied clustering, centrality and GO analysis to provide insights into what could be the best "reprogramming cocktails" for use in cell reprogramming. However, it is known that some small molecules, such as PD0325901, VPA, SB431542, CHIR99021, and others, which were not present in our networks due to the lack of data in the STITCH database (*e.g.*, RepSox and Pluripotin), can induce successful reprogramming in all cases studied so far. Nevertheless, identifying prospective new possible combinations is necessary for continuing progress in this field.

Our centrality data combined with the merged data of the CPI–PPI networks highlighted small molecules (BayK, SB431542, quercetin and OHTM) that might have more synergistic effects when used together because they appear to have wide topological roles.

Moreover, finding new targets to enhance cell reprogramming is also a difficult task in the field. Our centrality, interference and clustering data listed potential targets to be inhibited or expressed during reprogramming, and their inhibition by the described drugs or other methods should be explored. These proteins, such as XPO4, XPO1, DOT1L, E2F, ABCB1, and p16, appear to be promising targets to enhance the self-renewal and pluripotency of iPSCs.

References

1 W. Li and S. Ding, *Trends Pharmacol. Sci.*, 2010, **31**, 36–45.
2 C. V. Borlongan, *Exp. Neurol.*, 2012, **237**, 142–146.
3 K. J. Brennand, A. Simone, J. Jou, C. Gelboin-Burkhart, N. Tran, S. Sangar, Y. Li, Y. Mu, G. Chen, D. Yu, S. McCarthy, J. Sebat and F. H. Gage, *Nature*, 2011, **473**, 221–225.
4 W. Li, K. Jiang and S. Ding, *Stem Cells*, 2012, **30**, 61–68.
5 J. M. Murphy, D. J. Fink, E. B. Hunziker and F. P. Barry, *Arthritis Rheum.*, 2003, **48**, 3464–3474.
6 K. Takahashi and S. Yamanaka, *Cell*, 2006, **126**, 663–676.
7 R. Calloni, E. A. Cordero, J. A. Henriques and D. Bonatto, *Stem Cells Dev.*, 2013, **22**, 1455–1476.
8 F. Gonzalez, S. Boue and J. C. Izpisua Belmonte, *Nat. Rev. Genet.*, 2011, **12**, 231–242.
9 Y. Choi and T. G. Nam, *Arch. Pharmacal Res.*, 2012, **35**, 281–297.
10 S. Zhu, W. Li, H. Zhou, W. Wei, R. Ambasudhan, T. Lin, J. Kim, K. Zhang and S. Ding, *Cell Stem Cell*, 2010, **7**, 651–655.
11 J. B. Su, D. Q. Pei and B. M. Qin, *Acta Pharmacol. Sin.*, 2013, **34**, 719–724.
12 P. Hou, Y. Li, X. Zhang, C. Liu, J. Guan, H. Li, T. Zhao, J. Ye, W. Yang, K. Liu, J. Ge, J. Xu, Q. Zhang, Y. Zhao and H. Deng, *Science*, 2013, **341**, 651–654.
13 N. Emre, R. Coleman and S. Ding, *Curr. Opin. Chem. Biol.*, 2007, **11**, 252–258.
14 Z. Li and T. M. Rana, *Nat. Commun.*, 2012, **3**, 1085.
15 C. S. Yang, C. G. Lopez and T. M. Rana, *Stem Cells*, 2011, **29**, 1528–1536.
16 S. Zhu, W. Wei and S. Ding, *Annu. Rev. Biomed. Eng.*, 2011, **13**, 73–90.
17 N. Chandra and J. Padiadpu, *Expert Opin. Drug Discovery*, 2013, **8**, 7–20.
18 P. Csermely, T. Korcsmaros, H. J. Kiss, G. London and R. Nussinov, *Pharmacol. Ther.*, 2013, **138**, 333–408.
19 H. C. Schneider and T. Klabunde, *Bioorg. Med. Chem. Lett.*, 2013, **23**, 1168–1176.
20 J. O. Rosado, J. P. Henriques and D. Bonatto, *Curr. Cancer Drug Targets*, 2011, **11**, 849–869.
21 L. J. Jensen, M. Kuhn, M. Stark, S. Chaffron, C. Creevey, J. Muller, T. Doerks, P. Julien, A. Roth, M. Simonovic, P. Bork and C. von Mering, *Nucleic Acids Res.*, 2009, **37**, D412–D416.
22 B. Snel, G. Lehmann, P. Bork and M. A. Huynen, *Nucleic Acids Res.*, 2000, **28**, 3442–3444.
23 P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski and T. Ideker, *Genome Res.*, 2003, **13**, 2498–2504.
24 M. Rebhan, V. Chalifa-Caspi, J. Prilusky and D. Lancet, *Trends Genet.*, 1997, **13**, 163.
25 M. Safran, I. Dalah, J. Alexander, N. Rosen, T. Iny Stein, M. Shmoish, N. Nativ, I. Bahir, T. Doniger, H. Krug, A. Sirota-Madi, T. Olender, Y. Golan, G. Stelzer, A. Harel and D. Lancet, *Database*, 2010, **2010**, baq020.
26 M. Kanehisa and S. Goto, *Nucleic Acids Res.*, 2000, **28**, 27–30.
27 S. Carbon, A. Ireland, C. J. Mungall, S. Shu, B. Marshall, S. Lewis and the AmiGO Huband the Web Presence Working Group, *Bioinformatics*, 2009, **25**, 288–289.
28 G. D. Bader and C. W. Hogue, *BMC Bioinf.*, 2003, **4**, 2.
29 G. Scardoni, M. Petterlini and C. Laudanna, *Bioinformatics*, 2009, **25**, 2857–2859.
30 M. E. J. Newman, *Social Networks*, 2005, **27**, 39–54.
31 S. Maere, K. Heymans and M. Kuiper, *Bioinformatics*, 2005, **21**, 3448–3449.
32 Y. Benjamini and Y. Hochberg, *J. R. Statist. Soc., Ser. B*, 1995, **57**, 289–300.
33 W. Li, H. Zhou, R. Abujarour, S. Zhu, J. Young Joo, T. Lin, E. Hao, H. R. Scholer, A. Hayek and S. Ding, *Stem Cells*, 2009, **27**, 2992–3000.
34 N. Sato, L. Meijer, L. Skaltsounis, P. Greengard and A. H. Brivanlou, *Nat. Med.*, 2004, **10**, 55–63.
35 X. Lian, C. Hsiao, G. Wilson, K. Zhu, L. B. Hazeltine, S. M. Azarin, K. K. Raval, J. Zhang, T. J. Kamp and S. P. Palecek, *Proc. Natl. Acad. Sci. U. S. A.*, 2012, **109**, E1848–E1857.
36 A. M. Singh, D. Reynolds, T. Cliff, S. Ohtsuka, A. L. Mattheyses, Y. Sun, L. Menendez, M. Kulik and S. Dalton, *Cell Stem Cell*, 2012, **10**, 312–326.
37 H. Zhou, W. Li, S. Zhu, J. Y. Joo, J. T. Do, W. Xiong, J. B. Kim, K. Zhang, H. R. Scholer and S. Ding, *J. Biol. Chem.*, 2010, **285**, 29676–29680.
38 R. Ougland, D. Lando, I. Jonson, J. A. Dahl, M. N. Moen, L. M. Nordstrand, T. Rognes, J. T. Lee, A. Klungland, T. Kouzarides and E. Larsen, *Stem Cells*, 2012, **30**, 2672–2682.

39 Z. Pan, S. Sikandar, M. Witherspoon, D. Dizon, T. Nguyen, K. Benirschke, C. Wiley, P. Vrana and S. M. Lipkin, *Dev. Dyn.*, 2008, **237**, 316–327.
40 P. Mali, B. K. Chou, J. Yen, Z. Ye, J. Zou, S. Dowey, R. A. Brodsky, J. E. Ohm, W. Yu, S. B. Baylin, K. Yusa, A. Bradley, D. J. Meyers, C. Mukherjee, P. A. Cole and L. Cheng, *Stem Cells*, 2010, **28**, 713–720.
41 R. Trokovic, J. Weltner, T. Manninen, M. Mikkola, K. Lundin, R. Hamalainen, A. Suomalainen and T. Otonkoski, *Stem Cells Dev.*, 2013, **22**, 114–123.
42 S. Zhang, Y. Z. Jiang, W. Zhang, L. Chen, T. Tong, W. Liu, Q. Mu, H. Liu, J. Ji, H. W. Ouyang and X. Zou, *Stem Cells Dev.*, 2013, **22**, 90–101.
43 H. Shin, Y. S. Lee and Y. C. Lee, *Oncol. Rep.*, 2012, **27**, 1111–1115.
44 J. Qiu, Z. Gao and H. Shima, *Oncol. Rep.*, 2012, **27**, 160–167.
45 M. A. Cayo, A. K. Cayo, S. M. Jarjour and H. Chen, *Am. J. Transl. Res.*, 2009, **1**, 178–183.
46 L. Wang, H. S. Luo and H. Xia, *J. Int. Med. Res.*, 2009, **37**, 803–811.
47 Y. Shi, C. Desponts, J. T. Do, H. S. Hahm, H. R. Scholer and S. Ding, *Cell Stem Cell*, 2008, **3**, 568–574.
48 S. N. Zimmer, Q. Zhou, T. Zhou, Z. Cheng, S. L. Abboud-Werner, D. Horn, M. Lecocke, R. White, A. V. Krivtsov, S. A. Armstrong, A. L. Kung, D. M. Livingston and V. I. Rebel, *Blood*, 2011, **118**, 69–79.
49 T. Chen, D. Yuan, B. Wei, J. Jiang, J. Kang, K. Ling, Y. Gu, J. Li, L. Xiao and G. Pei, *Stem Cells*, 2010, **28**, 1315–1325.
50 F. W. Chong, S. Chakravarthi, H. S. Nagaraja, P. M. Thanikachalam and N. Lee, *Malays. J. Pathol.*, 2009, **31**, 35–43.
51 J. Chen, Q. Han and D. Pei, *J. Mol. Cell Biol.*, 2012, **4**, 66–69.
52 S. E. Hiemer and X. Varelas, *Biochim. Biophys. Acta*, 2013, **1830**, 2323–2334.
53 H. Zhang, C. Y. Liu, Z. Y. Zha, B. Zhao, J. Yao, S. Zhao, Y. Xiong, Q. Y. Lei and K. L. Guan, *J. Biol. Chem.*, 2009, **284**, 13355–13362.
54 S. Musah, S. A. Morin, P. J. Wrighton, D. B. Zwick, S. Jin and L. L. Kiessling, *ACS Nano*, 2012, **6**, 10168–10177.
55 P. Taupin, *Expert Opin. Ther. Pat.*, 2009, **19**, 523–527.
56 P. Taupin, *Recent Pat. CNS Drug Discovery*, 2010, **5**, 253–257.
57 H. T. Zhang, Z. G. Zha, J. H. Cao, Z. J. Liang, H. Wu, M. T. He, X. Zang, P. Yao and J. Q. Zhang, *Chin. Med. J.*, 2011, **124**, 3537–3545.
58 G. J. Inman, F. J. Nicolas, J. F. Callahan, J. D. Harling, L. M. Gaster, A. D. Reith, N. J. Laping and C. S. Hill, *Mol. Pharmacol.*, 2002, **62**, 65–74.
59 J. K. Ichida, J. Blanchard, K. Lam, E. Y. Son, J. E. Chung, D. Egli, K. M. Loh, A. C. Carter, F. P. Di Giorgio, K. Koszka, D. Huangfu, H. Akutsu, D. R. Liu, L. L. Rubin and K. Eggan, *Cell Stem Cell*, 2009, **5**, 491–503.
60 T. Tra, L. Gong, L. P. Kao, X. L. Li, C. Grandela, R. J. Devenish, E. Wolvetang and M. Prescott, *PLoS One*, 2011, **6**, e27485.
61 S. K. Mak, Y. A. Huang, S. Iranmanesh, M. Vangipuram, R. Sundararajan, L. Nguyen, J. W. Langston and B. Schule, *Stem Cells Int.*, 2012, **2012**, 140427.
62 H. Yu, P. M. Kim, E. Sprecher, V. Trifonov and M. Gerstein, *PLoS Comput. Biol.*, 2007, **3**, e59.
63 J. Duo, G. G. Ying, G. W. Wang and L. Zhang, *Mol. Med. Rep.*, 2012, **5**, 1453–1456.
64 W. W. Chang, F. W. Hu, C. C. Yu, H. H. Wang, H. P. Feng, C. Lan, L. L. Tsai and Y. C. Chang, *Head Neck Oncol.*, 2013, **35**, 413–419.
65 S. N. Tang, J. Fu, D. Nall, M. Rodova, S. Shankar and R. K. Srivastava, *Int. J. Cancer*, 2012, **131**, 30–40.
66 Y. J. Kim, Y. C. Bae, K. T. Suh and J. S. Jung, *Biochem. Pharmacol.*, 2006, **72**, 1268–1278.
67 M. A. Esteban, T. Wang, B. Qin, J. Yang, D. Qin, J. Cai, W. Li, Z. Weng, J. Chen, S. Ni, K. Chen, Y. Li, X. Liu, J. Xu, S. Zhang, F. Li, W. He, K. Labuda, Y. Song, A. Peterbauer, S. Wolbank, H. Redl, M. Zhong, D. Cai, L. Zeng and D. Pei, *Cell Stem Cell*, 2010, **6**, 71–79.
68 T. L. Chung, R. M. Brena, G. Kolle, S. M. Grimmond, B. P. Berman, P. W. Laird, M. F. Pera and E. J. Wolvetang, *Stem Cells*, 2010, **28**, 1848–1855.
69 M. A. Esteban and D. Pei, *Nat. Genet.*, 2012, **44**, 366–367.
70 D. Huangfu, K. Osafune, R. Maehr, W. Guo, A. Eijkelenboom, S. Chen, W. Muhlestein and D. A. Melton, *Nat. Biotechnol.*, 2008, **26**, 1269–1275.
71 Q. Wang, X. Xu, J. Li, J. Liu, H. Gu, R. Zhang, J. Chen, Y. Kuang, J. Fei, C. Jiang, P. Wang, D. Pei, S. Ding and X. Xie, *Cell Res.*, 2011, **21**, 1424–1435.
72 M. A. Walasek, L. Bystrykh, V. van den Boom, S. Olthof, A. Ausema, M. Ritsema, G. Huls, G. de Haan and R. van Os, *Blood*, 2012, **119**, 3050–3059.
73 V. Stambolic, L. Ruel and J. R. Woodgett, *Curr. Biol.*, 1996, **6**, 1664–1668.
74 R. W. Robey, S. Shukla, E. M. Finley, R. K. Oldham, D. Barnett, S. V. Ambudkar, T. Fojo and S. E. Bates, *Biochem. Pharmacol.*, 2008, **75**, 1302–1312.
75 D. E. Abbott, C. M. Bailey, L. M. Postovit, E. A. Seftor, N. Margaryan, R. E. Seftor and M. J. Hendrix, *Cancer Microenviron.*, 2008, **1**, 13–21.
76 Y. Xu, F. Zhi, G. Xu, X. Tang, S. Lu, J. Wu and Y. Hu, *Biosci. Rep.*, 2012, **32**, 559–566.
77 Y. M. Kim, E. S. Jeon, M. R. Kim, J. S. Lee and J. H. Kim, *Cell. Signalling*, 2008, **20**, 1882–1889.
78 C. Vancheri, E. Gili, M. Failla, C. Mastruzzo, E. T. Salinaro, D. Lofurno, M. P. Pistorio, C. La Rosa, M. Caruso and N. Crimi, *J. Allergy Clin. Immunol.*, 2005, **116**, 1242–1248.
79 Z. Wu, M. Yang, H. Liu, H. Guo, Y. Wang, H. Cheng and L. Chen, *J. Biol. Chem.*, 2012, **287**, 38295–38304.
80 Y. E. Kim, J. A. Park, K. H. Nam, H. J. Kwon and Y. Lee, *BMB Rep.*, 2009, **42**, 148–153.
81 J. Lee, N. Sayed, A. Hunter, K. F. Au, W. H. Wong, E. S. Mocarski, R. R. Pera, E. Yakubov and J. P. Cooke, *Cell*, 2012, **151**, 547–558.
82 I. E. Zohn, Y. Li, E. Y. Skolnik, K. V. Anderson, J. Han and L. Niswander, *Cell*, 2006, **125**, 957–969.
83 H. Landgren and P. Carlsson, *Dev. Dyn.*, 2004, **231**, 396–401.

84 K. L. Agarwala, S. Ganesh, K. Amano, T. Suzuki and K. Yamakawa, *Biochem. Biophys. Res. Commun.*, 2001, **281**, 697–705.
85 E. Lapi, A. Iovino, G. Fontemaggi, A. R. Soliera, S. Iacovelli, A. Sacchi, G. Rechavi, D. Givol, G. Blandino and S. Strano, *Oncogene*, 2006, **25**, 3628–3637.
86 L. Yi, C. Lu, W. Hu, Y. Sun and A. J. Levine, *Cancer Res.*, 2012, **72**, 5635–5645.
87 S. Hu, K. D. Wilson, Z. Ghosh, L. Han, Y. Wang, F. Lan, K. J. Ransohoff, P. Burridge and J. C. Wu, *Stem Cells*, 2013, **31**, 259–268.
88 E. R. Barry, W. Krueger, C. M. Jakuba, E. Veilleux, D. J. Ambrosi, C. E. Nelson and T. P. Rasmussen, *Stem Cells*, 2009, **27**, 1538–1547.
89 W. Yu, E. J. Chory, A. K. Wernimont, W. Tempel, A. Scopton, A. Federation, J. J. Marineau, J. Qi, D. Barsyte-Lovejoy, J. Yi, R. Marcellus, R. E. Iacob, J. R. Engen, C. Griffin, A. Aman, E. Wienholds, F. Li, J. Pineda, G. Estiu, T. Shatseva, T. Hajian, R. Al-Awar, J. E. Dick, M. Vedadi, P. J. Brown, C. H. Arrowsmith, J. E. Bradner and M. Schapira, *Nat. Commun.*, 2012, **3**, 1288.
90 H. Zhang, S. Wei, S. Ning, Y. Jie, Y. Ru and Y. Gu, *Exp. Ther. Med.*, 2013, **5**, 119–127.
91 X. S. Puente, M. Pinyol, V. Quesada, L. Conde, G. R. Ordonez, N. Villamor, G. Escaramis, P. Jares, S. Bea, M. Gonzalez-Diaz, L. Bassaganyas, T. Baumann, M. Juan, M. Lopez-Guerra, D. Colomer, J. M. Tubio, C. Lopez, A. Navarro, C. Tornador, M. Aymerich, M. Rozman, J. M. Hernandez, D. A. Puente, J. M. Freije, G. Velasco, A. Gutierrez-Fernandez, D. Costa, A. Carrio, S. Guijarro, A. Enjuanes, L. Hernandez, J. Yague, P. Nicolas, C. M. Romeo-Casabona, H. Himmelbauer, E. Castillo, J. C. Dohm, S. de Sanjose, M. A. Piris, E. de Alava, J. San Miguel, R. Royo, J. L. Gelpi, D. Torrents, M. Orozco, D. G. Pisano, A. Valencia, R. Guigo, M. Bayes, S. Heath, M. Gut, P. Klatt, J. Marshall, K. Raine, L. A. Stebbings, P. A. Futreal, M. R. Stratton, P. J. Campbell, I. Gut, A. Lopez-Guillermo, X. Estivill, E. Montserrat, C. Lopez-Otin and E. Campo, *Nature*, 2011, **475**, 101–105.
92 D. Huangfu, R. Maehr, W. Guo, A. Eijkelenboom, M. Snitow, A. E. Chen and D. A. Melton, *Nat. Biotechnol.*, 2008, **26**, 795–797.
93 T. Lin, R. Ambasudhan, X. Yuan, W. Li, S. Hilcove, R. Abujarour, X. Lin, H. S. Hahm, E. Hao, A. Hayek and S. Ding, *Nat. Methods*, 2009, **6**, 805–808.
94 W. Wang, J. Yang, H. Liu, D. Lu, X. Chen, Z. Zenonos, L. S. Campos, R. Rad, G. Guo, S. Zhang, A. Bradley and P. Liu, *Proc. Natl. Acad. Sci. U. S. A.*, 2011, **108**, 18283–18288.
95 E. De Clercq, *Biochem. Pharmacol.*, 2009, **77**, 1655–1664.
96 J. Jiang, M. Zhao, A. Zhang, M. Yu, X. Lin, M. Wu, X. Wang, H. Lu, S. Zhu, Y. Yu, Z. Mao and W. Han, *Biomed. Pharmacother.*, 2010, **64**, 482–486.
97 Q. L. Ying, J. Wray, J. Nichols, L. Batlle-Morera, B. Doble, J. Woodgett, P. Cohen and A. Smith, *Nature*, 2008, **453**, 519–523.
98 J. C. Young, S. Wu, G. Hansteen, C. Du, L. Sambucetti, S. Remiszewski, A. M. O'Farrell, B. Hill, C. Lavau and L. J. Murray, *Cytotherapy*, 2004, **6**, 328–336.
99 T. Chen, L. Shen, J. Yu, H. Wan, A. Guo, J. Chen, Y. Long, J. Zhao and G. Pei, *Aging Cell*, 2011, **10**, 908–911.
100 J. Staerk, C. A. Lyssiotis, L. A. Medeiro, M. Bollong, R. K. Foreman, S. Zhu, M. Garcia, Q. Gao, L. C. Bouchez, L. L. Lairson, B. D. Charette, L. Supekova, J. Janes, A. Brinker, C. Y. Cho, R. Jaenisch and P. G. Schultz, *Angew. Chem., Int. Ed.*, 2011, **50**, 5734–5736.
101 E. Yoo, L. A. Paganessi, W. A. Alikhan, E. A. Paganessi, F. Hughes, H. C. Fung, E. Rich, C. M. Seong and K. W. Christopherson 2nd, *Transfusion*, 2012, **53**, 878–887.
102 T. E. North, W. Goessling, C. R. Walkley, C. Lengerke, K. R. Kopani, A. M. Lord, G. J. Weber, T. V. Bowman, I. H. Jang, T. Grosser, G. A. Fitzgerald, G. Q. Daley, S. H. Orkin and L. I. Zon, *Nature*, 2007, **447**, 1007–1011.
103 R. Gonzalez, J. W. Lee, E. Y. Snyder and P. G. Schultz, *Angew. Chem., Int. Ed.*, 2011, **50**, 3439–3441.
104 A. Morizane, D. Doi, T. Kikuchi, K. Nishimura and J. Takahashi, *J. Neurosci. Res.*, 2011, **89**, 117–126.
105 T. Miyabayashi, J. L. Teo, M. Yamamoto, M. McMillan, C. Nguyen and M. Kahn, *Proc. Natl. Acad. Sci. U. S. A.*, 2007, **104**, 5668–5673.
106 C. A. Lyssiotis, R. K. Foreman, J. Staerk, M. Garcia, D. Mathur, S. Markoulaki, J. Hanna, L. L. Lairson, B. D. Charette, L. C. Bouchez, M. Bollong, C. Kunick, A. Brinker, C. Y. Cho, P. G. Schultz and R. Jaenisch, *Proc. Natl. Acad. Sci. U. S. A.*, 2009, **106**, 8912–8917.
107 A. B. McLean, K. A. D'Amour, K. L. Jones, M. Krishnamoorthy, M. J. Kulik, D. M. Reynolds, A. M. Sheppard, H. Liu, Y. Xu, E. E. Baetge and S. Dalton, *Stem Cells*, 2007, **25**, 29–38.
108 A. Vazquez-Martin, S. Cufi, E. Lopez-Bonet, B. Corominas-Faja, C. Oliveras-Ferraros, B. Martin-Castillo and J. A. Menendez, *Surf. Sci. Rep.*, 2012, **2**, 964.
109 G. R. Rosania, Y. T. Chang, O. Perez, D. Sutherlin, H. Dong, D. J. Lockhart and P. G. Schultz, *Nat. Biotechnol.*, 2000, **18**, 304–308.
110 S. P. Liu, H. J. Harn, Y. J. Chien, C. H. Chang, C. Y. Hsu, R. H. Fu, Y. C. Huang, S. Y. Chen, W. C. Shyu and S. Z. Lin, *PLoS One*, 2012, **7**, e44024.
111 W. Li, E. Tian, Z. X. Chen, G. Sun, P. Ye, S. Yang, D. Lu, J. Xie, T. V. Ho, W. M. Tsark, C. Wang, D. A. Horne, A. D. Riggs, M. L. Yip and Y. Shi, *Proc. Natl. Acad. Sci. U. S. A.*, 2012, **109**, 20853–20858.
112 Y. Shi, J. T. Do, C. Desponts, H. S. Hahm, H. R. Scholer and S. Ding, *Cell Stem Cell*, 2008, **2**, 525–528.
113 S. Chen, J. T. Do, Q. Zhang, S. Yao, F. Yan, E. C. Peters, H. R. Scholer, P. G. Schultz and S. Ding, *Proc. Natl. Acad. Sci. U. S. A.*, 2006, **103**, 17266–17271.
114 M. Yang, K. Li, P. C. Ng, C. K. Chuen, T. K. Lau, Y. S. Cheng, Y. S. Liu, C. K. Li, P. M. Yuen, A. E. James, S. M. Lee and T. F. Fok, *Stem Cells*, 2007, **25**, 1800–1806.
115 A. E. Boitano, J. Wang, R. Romeo, L. C. Bouchez, A. E. Parker, S. E. Sutton, J. R. Walker, C. A. Flaveny, G. H. Perdew, M. S. Denison, P. G. Schultz and M. P. Cooke, *Science*, 2010, **329**, 1345–1348.
116 Y. Xu, X. Zhu, H. S. Hahm, W. Wei, E. Hao, A. Hayek and S. Ding, *Proc. Natl. Acad. Sci. U. S. A.*, 2010, **107**, 8129–8134.

117 I. H. Park, R. Zhao, J. A. West, A. Yabuuchi, H. Huo, T. A. Ince, P. H. Lerou, M. W. Lensch and G. Q. Daley, *Nature*, 2008, **451**, 141–146.
118 S. C. Choi, J. H. Choi, C. Y. Park, C. M. Ahn, S. J. Hong and D. S. Lim, *J. Cell. Physiol.*, 2012, **227**, 3678–3692.
119 N. Funamizu, C. R. Lacy, K. Fujita, K. Furukawa, T. Misawa, K. Yanaga and Y. Manome, *PLoS One*, 2012, **7**, e37424.
120 Y. Arima, N. Hayashi, H. Hayashi, M. Sasaki, K. Kai, E. Sugihara, E. Abe, A. Yoshida, S. Mikami, S. Nakamura and H. Saya, *Int. J. Cancer*, 2012, **130**, 2568–2579.
121 S. M. Hussein and A. A. Nagy, *Curr. Opin. Genet. Dev.*, 2012, **22**, 435–443.
122 H. H. Al-Khalaf, D. Colak, M. Al-Saif, A. Al-Bakheet, S. F. Hendrayani, N. Al-Yousef, N. Kaya, K. S. Khabar and A. Aboussekhra, *PLoS One*, 2011, **6**, e21111.
123 Y. W. Chen, W. H. Yang, M. Y. Wong, H. H. Chang and M. Yen-Ping Kuo, *J. Periodontol.*, 2012, **83**, 1546–1553.
124 Z. Tong, S. Sant, A. Khademhosseini and X. Jia, *Tissue Eng., Part A*, 2011, **17**, 2773–2785.
125 Y. S. Heo, S. K. Kim, C. I. Seo, Y. K. Kim, B. J. Sung, H. S. Lee, J. I. Lee, S. Y. Park, J. H. Kim, K. Y. Hwang, Y. L. Hyun, Y. H. Jeon, S. Ro, J. M. Cho, T. G. Lee and C. H. Yang, *EMBO J.*, 2004, **23**, 2185–2195.
126 J. Kluwe, J. P. Pradere, G. Y. Gwak, A. Mencin, S. De Minicis, C. H. Osterreicher, J. Colmenero, R. Bataller and R. F. Schwabe, *Gastroenterology*, 2010, **138**, 347–359.
127 T. Ogino, M. Ozaki, M. Hosako, M. Omori, S. Okada and A. Matsukawa, *Leuk. Res.*, 2009, **33**, 151–158.
128 P. Xu and R. J. Davis, *Mol. Cell. Biol.*, 2010, **30**, 1329–1340.
129 M. Chafai, E. Louiset, M. Basille, M. Cazillis, D. Vaudry, W. Rostene, P. Gressens, H. Vaudry and B. J. Gonzalez, *Ann. N. Y. Acad. Sci.*, 2006, **1070**, 185–189.
130 R. Fazzi, S. Pacini, V. Carnicelli, L. Trombi, M. Montali, E. Lazzarini and M. Petrini, *PLoS One*, 2011, **6**, e25600.
131 S. Murasawa, Y. Mori, Y. Nozawa, H. Masaki, K. Maruyama, Y. Tsutsumi, Y. Moriguchi, Y. Shibasaki, Y. Tanaka, T. Iwasaka, M. Inada and H. Matsubara, *Hypertension*, 1998, **32**, 668–675.
132 J. L. Bjornstad, B. Skrbic, H. S. Marstein, A. Hasic, I. Sjaastad, W. E. Louch, G. Florholmen, G. Christensen and T. Tonnessen, *Cardiovasc. Res.*, 2012, **93**, 100–110.
133 M. Jinnin, H. Ihn and K. Tamaki, *Mol. Pharmacol.*, 2006, **69**, 597–607.
134 M. O. Kim, S. H. Kim, Y. Y. Cho, J. Nadas, C. H. Jeong, K. Yao, D. J. Kim, D. H. Yu, Y. S. Keum, K. Y. Lee, Z. Huang, A. M. Bode and Z. Dong, *Nat. Struct. Mol. Biol.*, 2012, **19**, 283–290.
135 G. Schalasta and C. Doppler, *Mol. Cell. Biol.*, 1990, **10**, 5558–5561.
136 M. Amit, J. Chebath, V. Margulets, I. Laevsky, Y. Miropolsky, K. Shariki, M. Peri, I. Blais, G. Slutsky, M. Revel and J. Itskovitz-Eldor, *Stem Cell Rev.*, 2010, **6**, 248–259.
137 K. Khodadadi, H. Sumer, M. Pashaiasl, S. Lim, M. Williamson and P. J. Verma, *Stem Cells Int.*, 2012, **2012**, 429160.
138 K. G. Toth, B. R. McKay, M. De Lisio, J. P. Little, M. A. Tarnopolsky and G. Parise, *PLoS One*, 2011, **6**, e17392.
139 B. Xu, K. Zhang and Y. Huang, *RNA*, 2009, **15**, 357–361.
140 Y. Guo, C. Mantel, R. A. Hromas and H. E. Broxmeyer, *Stem Cells*, 2008, **26**, 30–34.
141 F. Gao, S. W. Kwon, Y. Zhao and Y. Jin, *J. Biol. Chem.*, 2009, **284**, 22263–22273.
142 Z. Koledova, L. R. Kafkova, L. Calabkova, V. Krystof, P. Dolezel and V. Divoky, *Stem Cells Dev.*, 2010, **19**, 181–194.
143 I. Neganova, F. Vilella, S. P. Atkinson, M. Lloret, J. F. Passos, T. von Zglinicki, J. E. O'Connor, D. Burks, R. Jones, L. Armstrong and M. Lako, *Stem Cells*, 2011, **29**, 651–659.
144 R. Gupta, D. Hong, F. Iborra, S. Sarno and T. Enver, *Science*, 2007, **316**, 590–593.

7. Biologia de Sistemas

Capítulo de Livro: Biologia de Sistemas

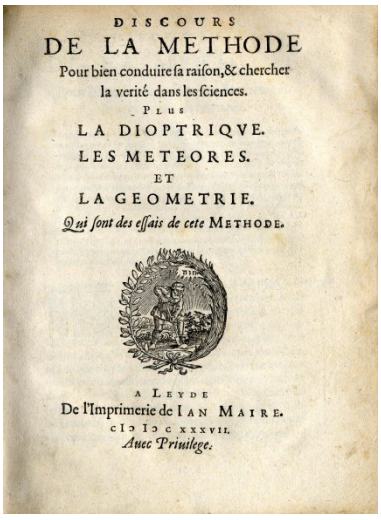
"Pensar a complexidade – esse é o maior desafio do pensamento contemporâneo, que necessita de uma reforma no nosso modo de pensar."

Edgar Morin & Jean-Louis Le Moigne

Joice de Faria Poloni
Bruno César Feltes
Fernanda Rabaioli da Silva
Diego Bonatto

1. Introdução

Uma das posturas metodológicas mais significativas do pensamento científico contemporâneo consiste em reduzir o todo a suas partes componentes. Por exemplo, entendemos o funcionamento de um organismo como fruto da ação de órgãos. Estes por sua vez, são compostos por tecidos, que são compostos por células. As células têm como componentes moléculas que, por fim, são compostas por átomos.


DISCOURS
DE LA METHODE
Pour bien conduire sa raison, & chercher la verité dans les sciences.
Plus
LA DIOPTRIQVE.
LES METEORES.
ET
LA GEOMETRIE.
Qui sont des essais de cette METHODE.
A LEYDE
De l'Imprimerie de IAN MAIRE.
cIↃ IↃ C XXXVII.
Avec Privilege.

Esta abordagem, especialmente importante e difundida na área biológica, é fruto das idéias introduzidas pelo filósofo René Descartes em meados do século XVII, indicando que cada problema encontrado deve ser dividido em tantas pequenas partes quanto for necessário para resolvê-lo de maneira mais parcimoniosa.

É neste contexto que emerge a divisão disciplinar no estudo da natureza. Desde os tempos da escola até a universidade, o conhecimento a ser ensinado manifesta-se na separação das disciplinas. Por exemplo, no meio acadêmico observamos a biologia compartimentada em botânica, zoologia, ecologia, genética, biologia celular e essas, por sua vez, subdivididas em outras áreas. Como aspecto positivo, o estudo das partes forma especialistas e divide o trabalho, facilitando o entendimento de suas partes componentes. Contudo, neste processo tem-se uma redução da complexidade característica dos fenômenos naturais, o que pode comprometer nossa capacidade de entendê-los.

De fato, a complexidade é inerente à biologia, ao funcionamento do nosso organismo e à natureza. Há a necessidade, assim, da construção de uma abordagem que inclua esta complexidade, de forma sistêmica; que interligue as diversas interações presentes e que, ao confrontá-las, consiga encontrar relações mais informativas e completas.

A partir desta premissa, emergem na

década de 50 as primeiras concepções sobre a Biologia de Sistemas (BS). Essa área, pautada nos conceitos de sistema e de complexidade, envolve um estudo sistemático de interações em um sistema biológico.

O conceito de sistema é entendido como um conjunto de partes ou elementos que possuem relações entre si, relações estas que diferem-se daquelas realizadas com outros elementos, fora do sistema. Já a idéia de complexidade é definida como a condição de elementos de um sistema e a relação entre esses elementos em um determinado momento.

Um sistema complexo, por conseguinte, é um sistema composto de partes interconectadas que, como um todo, exibe uma ou mais propriedades que não seriam observadas a partir das propriedades dos componentes individuais, possibilitando assim a observação de novos fenômenos. Portanto, a BS é um campo que investiga as interações entre os componentes de um sistema biológico, buscando contribuir no entendimento de como estas interações influenciam a função e o comportamento do sistema.

A busca da compreensão da biologia em nível de sistema é um tema recorrente na comunidade científica. Norbert Wiener, em 1948, foi um dos proponentes da abordagem sistemática que levou ao nascimento da cibernética, ou biocibernética, consolidada com os estudos do médico neurologista, William Ross Ashby (1903-1972). A partir de 1959, Robert Rosen, sob orientação do professor Nicolas Rashevsky, propôs uma metodologia baseada na "biologia relacional", onde o mais importante na biologia era o estudo da vida em si. Após 20 anos, Ludwig von Bertalanffy (1901-1972) criou a teoria geral dos sistemas, tornando-se o precursor da BS. Em 1966 foi formalizado o estudo da BS, com o lançamento da disciplina "Teoria e Biologia de Sistemas" pelo teórico de sistemas Mihajlo Mesarovic (1928).

A partir do trabalho destes pesquisdores, a teoria geral dos sistemas pode ser definida como a área que estuda a organização abstrata de fenômenos, investigando todos os princípios comuns a todas as entidades complexas (não somente biológicas) e os modelos que podem ser utilizados para a sua descrição.

Com o avanço da biologia molecular nas décadas que se seguiram, juntamente com o nascimento da genômica funcional, grandes quantidades de dados tornaram-se disponíveis e os bancos de dados e ferramentas de análise adaptaram-se ao volume crescente de informações, permitindo construir modelos mais amplos, capazes de lidar com aspectos e fenômenos inacessíveis até então. Assim em 2000, quando o Instituto de Biologia de Sistemas foi fundado, a biologia de sistemas emergiu como um campo próprio, estimulado pelo aumento de dados "ômicos" e pelos avanços da parte experimental e da bioinformática visando o entendimento sistemático da biologia. Desde então, grupos de pesquisas dedicados à BS têm sido formados em todo o mundo.

Para tal, a BS depende de ferramentas interdisciplinares para obter, integrar e analisar diversos tipos dados, exemplificados na Tabela 1. Essa abordagem requer novas técnicas de análise, ferramentas de informática, métodos experimentais e uma nova postura metodológica, articulando partes normalmente estudadas separadamente.

2. Biologia de Sistemas

Em suas análises, a BS relaciona partes individuais de um sistema como representações gráficas de conjuntos de nós ou vértices (V), conectados entre si por conectores ou arestas (E, do inglês edge). Os nós podem representar indivíduos, proteínas ou mesmo lugares, enquanto que os conectores representam a conexão que está presente entre cada par de nós. Esta representação gráfica é denominada de rede.

Muitos exemplos de rede podem ser citados, como redes de cadeia alimentar, amplamente aplicadas na ecologia, redes neurais e de interação protéica usadas na biologia e

Tabela 1: Ferramentas utilizadas no estudo da BS.

Área	Tipo de análise
Bioinformática	Funções biológicas por meio de ferramentas da informática
Genômica	Sequências de DNA
Transcriptômica	Transcritos
Proteômica	Proteínas
Interatômica	Interações protéicas
Interferômica/ microRNômica	RNAi/miRNA
Epigenômica	Modificações na cromatina e no DNA
Metabolômica	Metabólitos
Fluxômica	Alterações dinâmicas de moléculas dentro de uma célula ao longo do tempo
Biômica	Bioma
Glicômica	Totalidade de carboidratos
Farmacogenômica	Genes que definem o comportamento da droga
Nutrigenômica	Relação entre a dieta e os genes individuais
Toxicogenômica	Estrutura e atividade do genoma e os efeitos biológicos adversos na exposição à xenobióticos
Imunômica	Função molecular associada aos transcritos de RNAm relacionados à resposta imune

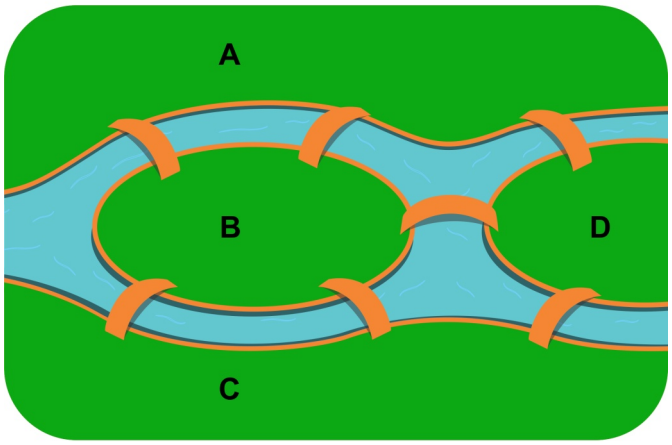
ciências médicas, além da própria World Wide Web, que representa uma das maiores redes funcionais no mundo da comunicação e informática.

A análise matemática de redes é denominada de teoria de grafos, e consiste em um dos principais objetos de estudo da matemática discreta. Desta forma, o termo "rede" representa as interações funcionais de um sistema, enquanto que o termo "grafo" enfatiza as análises matemáticas deste sistema. Neste capítulo, contudo, usaremos ambos os termos como sinônimos.

Historicamente, a teoria de grafos foi desenvolvida em 1736 pelo matemático suíço Leonard Euler na resolução do problema das sete pontes de Königsberg, atualmente conhecida como Kaliningrado, na Rússia. A cidade de Königsberg é atravessada pelo Rio Pregel e consiste de duas grandes ilhas que eram conectadas entre si e com as margens opostas por sete pontes (Figura 1A). O problema apresentado a Euler consistia em descobrir como caminhar pela cidade atravessando cada ponte apenas uma vez. A técnica desenvolvida pelo matemático suíço foi adaptar o mapa de Königsberg, transformando as margens e ilhas em nós e as pontes em conectores (Figura 1B). Euler submeteu a rede que desenvolveu a análises matemáticas, porém não encontrou solução para o problema. Contudo, a metodologia de análise de Euler foi um marco histórico na análise de problemas combinatórios, além de estabelecer o conceito de topologia que é usado em BS (ver adiante).

O emprego da teoria de grafos e suas aplicações têm apresentado um crescimento explosivo devido a sua multidisciplinaridade e ao seu conceito de modelo que permite

A



B

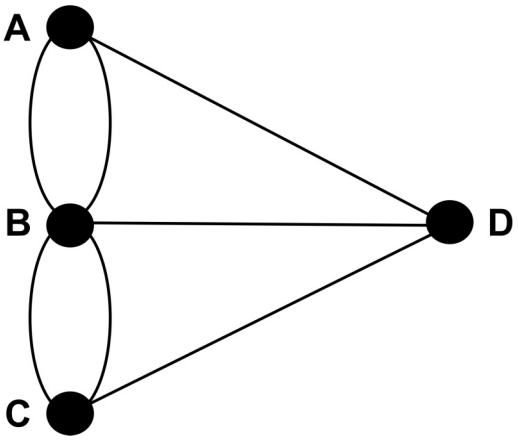


Figura 1: (A) Representação parcial do mapa de Königsberg e suas setes pontes. (B) Ilustração da rede desenvolvida por Euler.

estudar um objeto específico sem negligenciar o meio em que este objeto se encontra. Por exemplo, é possível estudar determinado fármaco considerando a atividade que diversos compostos e enzimas poderiam exercer sobre ele. Nesses estudos pode-se construir uma rede onde os nós representam compostos e enzimas e os conectores representam se há ou não relação entre eles, permitindo analisar:

i) a conectividade dos compostos ou enzimas, ou seja, que tipo de relação duas moléculas aleatórias podem apresentar na rede;

ii) a centralidade, que caracteriza as moléculas que apresentam maior influência sob a ação do fármaco em questão.

Conceitos básicos de grafos

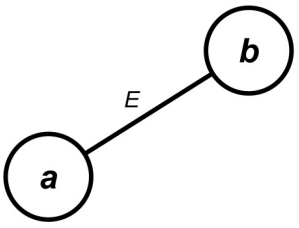
Considerando-se a estreita relação entre a BS e a teoria de grafos, alguns conceitos matemáticos podem nos ajudar a entender e empregar esta área do conhecimento com maior domínio e propriedade. Assim, prosseguiremos com uma breve introdução sobre teoria de grafos e estrutura de rede, apresentando alguns descritores matemáticos frequentemente empregados em BS.

Uma rede (ou grafo) $G = (V, E)$ representa uma combinação de nós (V) e conectores (E) que ligam os nós. Em uma rede, o conjunto de seus nós é denotado por $V(G)$, enquanto o conjunto de seus conectores por $E(G)$. Dessa forma, o número total de nós em G é representado por n, e o número total de conectores é representado por m:

$$n(G) = |V(G)| \text{ e } m(G) = |E(G)|$$

Adicionalmente, conforme apresentado na Figura 2A, um conector E deve apresentar

A



B

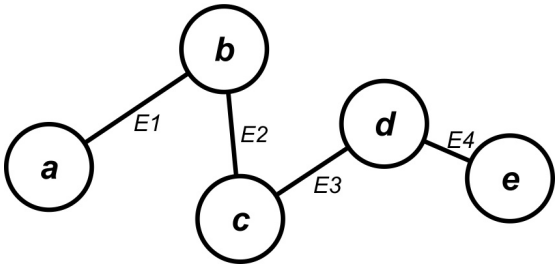


Figura 2: Em (A) a representação da interação de dois nós vizinhos ($V = a, b$) conectados pelo conector $E(a, b)$. Em (B) a rede pode ser descrito como $V = \{a, b, c, d, e\}$ e $E = \{ab, bc, cd, de\}$, com $n = 5$ (5 nós de a a e) e $m = 4$ (4 conectores de 1 a 4).

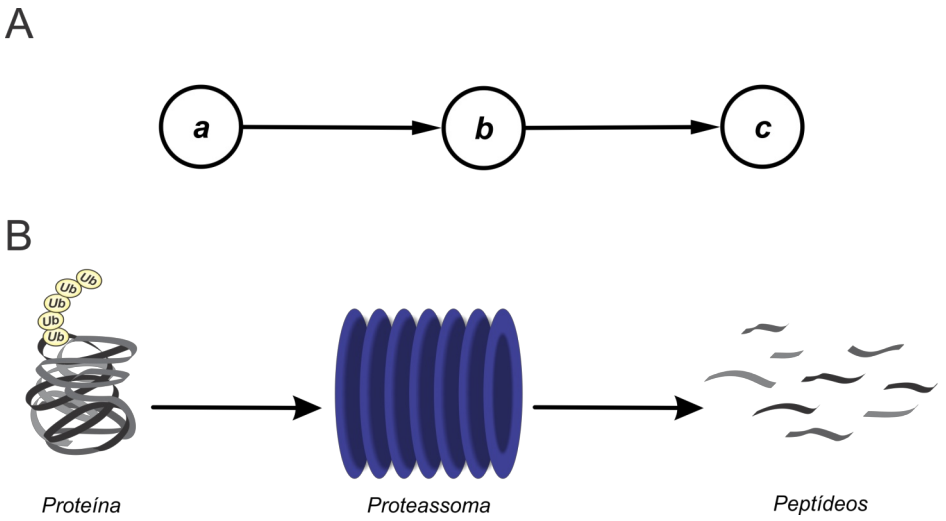


Figura 3: (A) Rede direta; (B) Representação da via de degradação ubiquitina-proteassoma, um dos inúmeros tipos de rede direcionadas encontradas em sistemas biológicos.

suas extremidades ligadas aos nós a e b (a □ V e b □ V), sendo chamado eab, $E(a, b)$ ou apenas ab. Este conector pode ser representado da seguinte forma:

$$E = \{(a, b) \mid a, b \square V\}$$

As redes podem apresentar conectores diretos, ou seja, um conector orientado em determinada direção (exemplo $a \rightarrow b$, $b \rightarrow c$), sendo assim chamadas de redes direcionadas ou dígrafos (Figura 3A). Nos conectores E = (a, b) e E = (b, c), podemos dizer que a é antecessor a b, e b é antecessor a c. Da mesma forma, b é sucessor de a e c é sucessor de b. Um dígrafo é definido por G = (V, E, f), sendo f uma função que associa cada elemento E a um par ordenado de nós em V. Uma rede representando os mecanismos de degradação ubiquitina-proteassoma de uma determinada proteína pode ser um exemplo de rede direta após o reconhecimento da

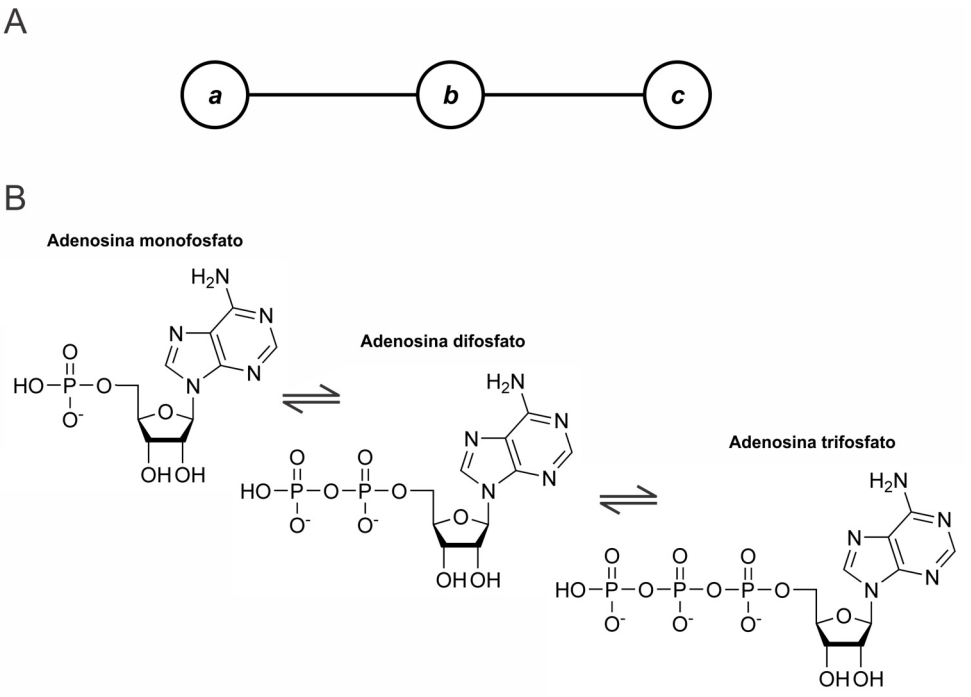


Figura 4: (A) Rede não direcionada; (B) Reação reversa de fosforilação e desfosforilação de adenosina difosfato, representando um exemplo de redes não direcionadas em sistemas biológicos.

proteína ubiquitinada por proteassomas, uma vez que não é possível reverter à degradação da proteína (Figura 3B).

Podem também existir redes não direcionadas (Figura 4A), que apresentam conectores orientados em ambas as direções ($a \leftrightarrow b$, $b \leftrightarrow c$), não sendo possível assim estabelecer antecessor ou sucessor. Um exemplo típico seria a reação reversível de um substrato A para um substrato B em uma via metabólica como, por exemplo, a formação de diferentes moléculas fosforiladas de adenosina conforme a reação AMP ↔ ADP ↔ ATP (Figura 4B).

Em alguns casos, podem existir dois ou mais conectores que ligam os mesmos nós na rede. Esse tipo de interação é chamado multiconector, onde diferentes informações são representadas por cada conector, caracterizando assim um multidígrafo (Figura 5).

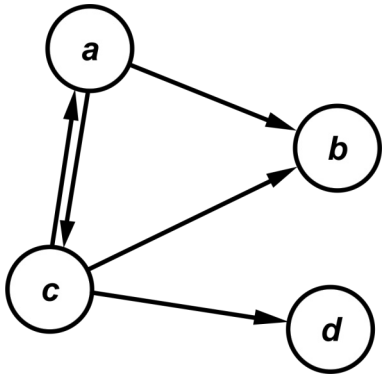


Figura 5: Multidígrafo $G = (V, E)$, onde $V = \{a, b, c, d\}$ e $E = \{ab, ac, ca, cb, cd\}$.

Observa-se, assim, que as redes apresentam interações entre os nós e que essas interações são delimitadas pelos conectores. Portanto, se $E = (a, b)$, logo os nós a e b são vizinhos ou adjacentes, e $E(a, b)$ é incidente aos nós a e b, lembrando que $E(a, b)$ se refere ao conector.

Uma das formas de representar e descrever tais interações entre os nós de uma determinada rede envolve o uso de matrizes. Assim, se considerarmos uma rede G contendo os nós v_a, ... v_n a matriz que descreve os elementos adjacentes em G é dada por:

$$a_{ij} = \begin{cases} 1 \text{ se } viv_j \in E(G) \\ 0 \text{ se } v_iv_j \notin E(G) \end{cases}$$

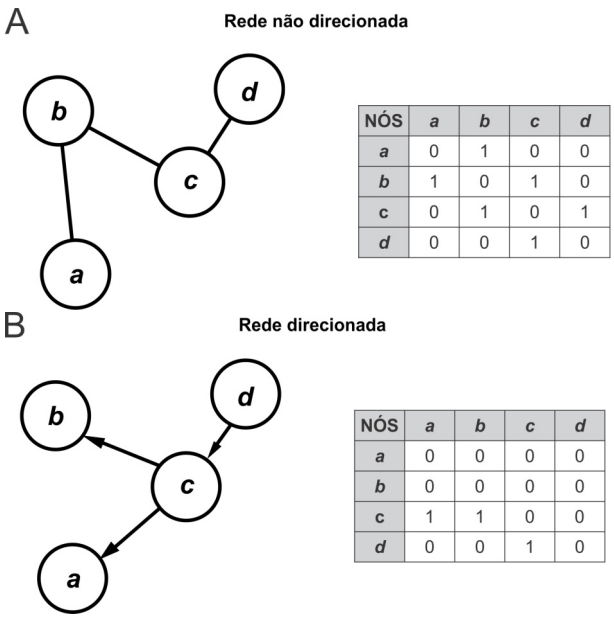
As tabelas representadas na Figura 6 são um mecanismo visual para compreender como a matriz de uma rede é elaborada, tanto para redes não direcionadas (Figura 6A) quanto direcionadas (Figura 6B).

Para as redes não direcionada (Figura 6A) e direcionada (Figura 6B), as matrizes são representadas abaixo:

$$M = \begin{matrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{matrix} \qquad M = \begin{matrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{matrix}$$

Rede direcionada *Rede não direcionada*

Ao analisarmos uma matriz devemos considerar cada nó como uma coluna e uma linha distinta. Na análise da primeira matriz iremos interpor o nó representado na linha 1 (nó a) com o nó representado na coluna 1 (nó a) da mesma forma que as tabelas representadas na Figura 6, e como não há interação de a com a, nos referimos como 0. Da mesma forma, se consideramos a linha 1 (nó a) e a coluna 2 (nó b), há conexão, sendo representado por 1. Perceba que as matrizes são diferentes na rede direcionada e não



NÓS	a	b	c	d
a	0	1	0	0
b	1	0	1	0
c	0	1	0	1
d	0	0	1	0

NÓS	a	b	c	d
a	0	0	0	0
b	0	0	0	0
c	1	1	0	0
d	0	0	1	0

Figura 6: (A) Rede não direcionada $G = (V, E)$, onde $V = \{a, b, c, d\}$ e $E = \{ab, bc, cd\}$ ou $E = \{ba, cb, dc\}$, representados também na tabela pelo número 1, que indica a presença de um conector entre dois nós, exemplo $E = \{ab, ba\} = 1$. A ausência do conector entre dois nós é representada por 0. (B) Rede direcionada $G = (V, E)$, onde $V = \{a, b, c, d\}$ e $E = \{ca, cb, dc\}$. Neste caso, a tabela de interações muda devido ao direcionamento das conexões, por exemplo $E = \{ca\} = 1$, mas $E = \{ac\} = 0$.

direcionada devido à atribuição de uma conexão direcionada. Na matriz direcionada, tanto b está conectado a c quanto c está conectado a b. Contudo, na matriz não direcionada, somente c está conectado a b.

Também podemos definir uma rede como completa se $E(G) = V(G)^{(2)}$, isto é, se dois nós selecionados aleatoriamente na rede G são adjacentes. Assim, uma rede completa tem n nós e é representada por K_n, sendo o número de conectores em K_n representado por $\binom{n}{2}$.

O conjunto de nós e conectores de uma rede pode ser apresentado em uma representação mais complexa e informativa, agregando pesos (atributos) associados aos nós e conectores (Figura 7). Redes que apresentam nós e conectores com atributos são chamadas de redes ponderadas (G, w), onde $G = (V, E)$ e $w = V, E \rightarrow R$, sendo R o conjunto dos números reais e w correspondente a função atributo. Por exemplo, pode-se representar uma rede neural onde o atributo indica a distância que um sinal neural deve percorrer em relação ao local de origem. Assim, se P é uma trajetória na rede, $w(P)$ é considerada a extensão de P. Redes ponderadas são amplamente usadas na bioinformática, onde $_{G, w}(a, b)$ pode representar a quantidade e a fidelidade de informações armazenadas em bancos de dados a respeito da interação entre a e b (Figura 7).

Também podemos nos referir a uma rede como bipartida (Figura 8) onde, em $G = (V, E)$, V pode ser dividido em V_x e V_y. Assim, cada nó de V_x é adjacente aos vértices de V_y. Desta forma, se consideramos E (a, b) significa que $a \in V_x$, enquanto que $b \in V_y$ ou $a \in V_y$ e $b \in V_x$. A aplicação de redes bipartidas na modelagem de redes biológicas pode ser vista em vários contextos, desde a análise de genótipos e SNPs (single-nucleotide polymorphism) em diferentes populações até a representação de conexões ecológicas e reações enzimáticas em vias metabólicas.

O modelo de redes visto até agora, na qual um conector se liga a dois nós, apesar de amplamente utilizado na avaliação da conectividade de redes biológicas, pode ser

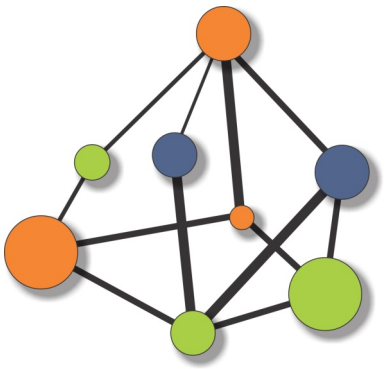


Figura 7: Representação de uma rede ponderada descrevendo: *i*) diferentes tipos de nós,onde cada cor representa diferentes famílias de proteínas (por exemplo, os nós verdes representam serina/treonina cinases, nós azuis representam cinases dependentes de ciclinas e nós laranja representam as tirosina cinases); *ii*) diferentes tamanhos de nós, com atributo $w(a)$, representando o número de artigos w que citam a proteína a; e *iii*) a espessura do conector y, representando a fidelidade w da interação entre duas proteínas distintas.

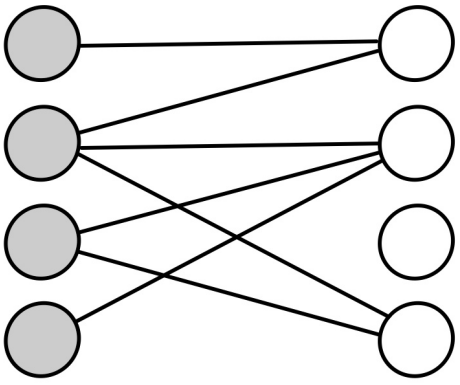


Figura 8: Representação de uma rede bipartida, onde o nós cinzas e brancos representam diferentes grupos de uma análise. Por exemplo, cada grupo pode representar duas linhagens diferentes de E. coli. Para avaliar a eficiência de transformação das linhagens, estas foram divididas em quatro amostras (representadas pelos nós) e cada amostra foi incubada com diferentes plasmídeos. Os conectores apresentam os plasmídeos que obtiveram sucesso na transformação e são comuns entre as duas linhagens.

uma representação simplista quando se trata de redes metabólicas. A organização biológica que caracteriza as redes metabólicas em um contexto bioquímico consiste de complexas interações, frequentemente envolvendo diversos substratos e produtos. Para melhor representar a complexidade de reações bioquímicas, usam-se redes conhecidas como hipergrafos (Figura 9).

Os hipergrafos são caracterizados pela presença de hipervértices, que conectam mais de dois nós com propriedades distintas (Figura 9). Assim, os hipergrafos são frequentemente usados em organizações bioquímicas, devido à intersecção de componentes com atividades em diferentes rotas metabólicas.

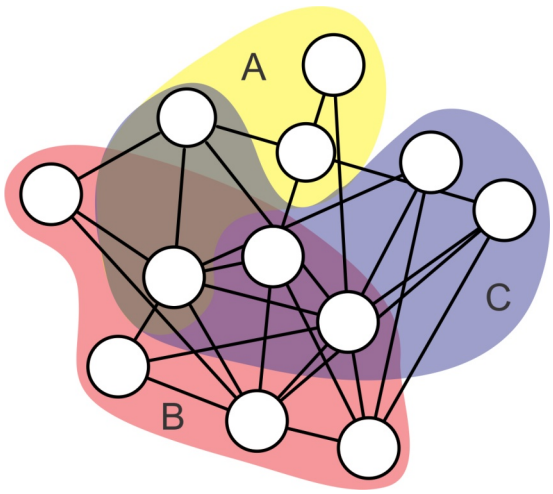


Figura 9: Representação de um hipergrafo. As regiões destacadas em várias cores caracterizam as diferentes propriedades ou atividades bioquímicas representadas na rede. Assim, cada cor estaria representando diferentes vias metabólicas (A, B e C). Os nós da rede indicam componentes presentes em cada uma das vias metabólicas e/ou participando de vias distintas nas regiões intersectadas.

Geralmente, as redes biológicas são extensas, apresentando um grande número de nós. Contudo, análises estatísticas indicam que, dentro de uma rede maior (Figura 10A), podem existir redes menores que participam da composição geral e possuem maior conectividade entre si quando comparados à rede maior (Figura 10B). Essas subredes de G

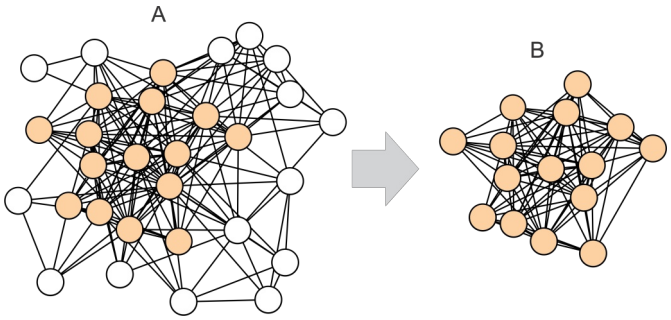


Figura 10: (A) Rede de interações proteína-proteína representando em laranja a subrede, o qual foi destacada em (B).

= (V, E) nada mais são que uma rede G_1 = (V_1, E_1), onde $V_1 \subseteq V$ e $E_1 \subseteq E$.

3. Estrutura de redes

Uma das características de uma rede é sua conectividade (também referida como grau de nó), sendo a conectividade total de uma rede definida por $C = E / N (N - 1)$, onde E representa o número de conectores e N o número total de nós.

Considere os nós V_a e V_e de uma rede. Representamos como um dos possíveis caminhos de V_a a V_e os vértices V_b, V_c e V_d, formando um conector a cada dois vértices sucessivos, caracterizados por E_1, E_8, E_4, E_7 (Figura 11). O nó que originou o caminho é chamado de nó inicial, enquanto que o último nó do caminho é chamado de nó final. Um caminho onde o nó inicial coincide com o nó final, sem repetições de conexões intermediárias, é chamado de circuito. Usando a mesma rede da Figura 11, <d, b, c, e, d> formam um circuito. O comprimento de um caminho ou circuito consiste do número de

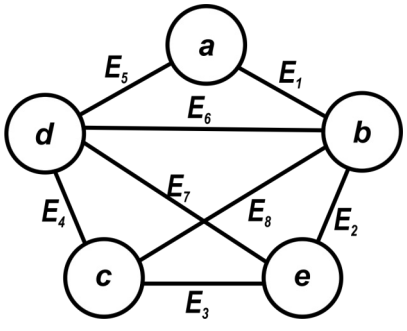


Figura 11: Esquema representando uma rede, onde V = {a, b, c, d, e} e E = {E_1, E_2, E_3, E_4, E_5, E_6, E_7, E_8}.

conectores que pertencem ao caminho (ou circuito) ou, no caso de uma rede ponderada, pela soma dos atributos (ou pesos) dos conectores.

Um caminho de comprimento k tem exatamente $k + 1$ nós, enquanto que um circuito de comprimento k tem $k = v$ nós. Se calcularmos o comprimento de V_a a V_e, com caminho E_1, E_8, E_4, E_7 temos $k = 4$ conectores com $4 + 1$ nós. Para o circuito <d, b, c, e, d> que tem como caminho E_6, E_8, E_3, E_7 temos $k = 4$ conectores, com quatro nós diferentes.

Uma importante análise em uma rede consiste em caracterizá-la conforme sua distribuição de caminhos geodésicos. Um caminho geodésico é definido como a via mais curta dentro de uma rede entre dois nós quaisquer (i e j), sendo representado por δ (i, j) em G. Um bom exemplo disso é o experimento realizado por Stanley Milgram em 1960, onde cartas foram enviadas a indivíduos aleatoriamente. A missão de cada indivíduo era enviar a sua carta a alguém que considerasse capaz de fazer com que as cartas chegassem ao seu destino final. Essa experiência relativamente simples conclui que existem aproximadamente seis graus de separação entre dois indivíduos quaisquer no mundo. Da mesma forma, esse experimento foi a primeira demonstração significativa do efeito "mundo pequeno" (ou do inglês, small world), que estabelece que as redes apresentam nós conectados entre si formando um caminho mais curto entre todos os nós.

O comprimento médio de caminhos entre os nós (i, j) é definido pelo valor médio de conectores entre os nós e pode ser calculado por:

$$\delta = \frac{2}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{N} \delta_{min(i,j)}$$

, assumindo-se que δ_{min} (i, j) é o caminho mais curto entre os nós i e j, sendo N o número total de nós. Adicionalmente, o diâmetro da rede é definido como:

$$D = \max_{i,j} \delta_{min}(i,j)$$

e representa o maior comprimento entre dois nós. Estudos recentes têm revelado que redes biomoleculares, sociais e tecnológicas apresentam valores de comprimento médio de caminhos e diâmetro relativamente pequenos se comparados ao tamanho da rede, apresentando ordem de grandeza log (n) ou menor quando o tamanho da rede é n. Da mesma forma, a densidade de uma rede é calculada com base no número de conexões que cada nó possui, sendo definida como:

$$\rho = \frac{2m}{n(n-1)}$$

Avaliar a densidade de uma rede representa avaliar o nível de conectividade, tornando-se muito importante na definição de suas propriedades, como veremos adiante. Por exemplo, ao analisarmos a rede de interação de uma doença contagiosa, a possibilidade desta doença até então controlada tornar-se uma epidemia depende principalmente de duas variáveis: o tipo de agente infeccioso e a alta densidade de conexões (rotas de transmissão). O procedimento de quarentena (isolamento) quando um determinado indivíduo apresenta os sintomas da doença é justamente reduzir a conectividade da rede de transmissão.

Alguns modelos de rede (como as redes de livre escala e hirárquica, discutidas adiante em 5. Tipos de Rede) podem apresentar clusterização, isto é, os nós tendem a se agrupar. Isso significa que se um nó A se liga ao nó B, e o nó B se liga ao nó C, então há grandes chances de A se ligar a C também. Assim, a rede é composta de centenas de triângulos, ou seja, grupos de três nós conectados entre si, onde cada lateral de um triângulo pode pertencer a outro triângulo.

Podemos quantificar a fração de triplos nós que apresentam um terceiro conector preenchendo um triângulo pelo coeficiente de clusterização:

$$C = \frac{3 \times número\ de\ triângulos\ na\ rede}{número\ de\ nós\ triplamente\ conectados}$$

Na equação, o número três presente no numerador é devido ao fato que cada lateral de um triângulo contribui com outros três triplos nós, além de garantir que C seja $0 \leq C \leq 1$. Dessa forma, o coeficiente de clusterização avalia a probabilidade dos nós i e j serem vizinhos, já que ambos são vizinhos do nó h.

Assim, o coeficiente de clusterização local de um nó i pode ser determinado por:

$$C_i = \frac{2e}{k(k-1)}$$

, onde um nó i tem k vizinhos com e conexões entre eles. Contudo, pode-se também atribuir o coeficiente de clusterização média para a rede total, sendo definido por:

$$C = \frac{1}{N}\sum_i C_i$$

Ao analisarmos uma rede de processos biológicos, notamos que esta apresenta um maior coeficiente de clusterização média quando comparado a uma rede aleatória. Isso possivelmente se deve ao fato de processos celulares ocorrerem de forma dependente da organização de diversos subconjuntos (clusters) de biomoléculas.

Em uma rede consideramos como sendo o grau de um nó o número de conectores k que incidem a este nó. Assim, a distribuição do grau $P(k)$ é definida por ser uma fração de nós com grau k dentro de uma rede. Então sendo k = 0, 1, 2,... $P(k)$ indica a probabilidade de determinado nó ter grau k. A distribuição de grau é definida por:

$$P(k) = \frac{n_k}{n}$$

, onde temos n nós na totalidade da rede e n_k representa a quantidade de nós com grau k.

Uma rede aleatória que apresenta n nós conectados ou não com probabilidade p, tem uma distribuição binominal de grau com parâmetros N-1 e p:

$$P(k_i = k) = C_{N-1}^{k}\, p^k (1-p)^{N-1-K}$$

Outras redes, no entanto, tem distribuição de grau bem diferente. Redes de livre escala (como a maioria das redes biológicas) apresentam distribuição do grau que segue uma Lei de Potência $P(k) \sim k^{-\gamma}$, $\gamma > 1$ (ver adiante neste capítulo 5.Tipos de redes).

Outra estimativa numérica pode ser feita, a função de distribuição cumulativa avalia a probabilidade de um nó ter um grau maior do que k:

$$P_k = \sum_{k'=k}^{\infty} p_{k'}$$

Agora, o que aconteceria se, por acaso, resolvermos excluir alguns poucos nós da rede? Certamente iríamos alterar o comprimento de alguns caminhos e circuitos da rede de forma pouco significativa. Contudo, se formos excluindo mais nós, progressivamente, veremos que a comunicação da rede fica cada vez mais esparsa, até se tornar desconectada. A capacidade de uma rede de tolerar a deleção de nós é chamada de resiliência.

Em 2000, um estudo conduzido por Albert-László Barabási e colaboradores mostrou que a Internet pode ser altamente resiliente na remoção de nós aleatórios. Isso se deve ao fato de que a quantidade de nós com baixo grau de interação é maior em uma rede do que nós com alto grau de interação. Em compensação, se a remoção iniciar a partir dos nós com mais alto grau de interação, a alteração será brusca. Neste caso, observa-se um aumento da distância entre os nós, de forma que apenas poucos nós precisam ser removidos para destruir a comunicação da rede. Assim, fica claro que a Internet apresenta baixa resiliência na remoção de nós com alto grau, tornando-se vulnerável a ataques de hackers. Outro exemplo seriam as redes de interação proteína-proteína. Estas redes geralmente apresentam muitas proteínas com poucas interações e algumas proteínas possuindo muitas interações (chamadas de hubs, ver adiante). Desta forma, redes de interação proteína-proteína são resilientes à deleção de nós aleatórios, porém extremamente vulneráveis a ataques em proteínas hubs.

Os nós de uma determinada rede podem apresentar tendências de conexão. Em outras palavras, duas redes completamente diferentes topologicamente podem apresentar a mesma distribuição do grau. Assim, em uma rede é preciso considerar o padrão de correlação do grau dos nós, onde a conectividade de um nó reflete nas suas possibilidades de ligação.

A tendência de conexão que uma rede apresenta pode ser chamada de assortatividade e desassortatividade. A assortatividade significa que os nós de uma rede apresentam uma tendência a interagirem com outros nós semelhantes, por exemplo, nós do tipo A interagem preferencialmente com nós também do tipo A

(Figura 12A). Vértices com alto grau tendem a interagir com vértices que também apresentam alto grau. No entanto, chamamos de desassortatividade se os nós de uma rede interagem preferencialmente com nós diferentes dele mesmo, por exemplo, nós do tipo A tendem a interagir com nós do tipo B. Neste caso, um nó com alto grau tem tendência a interagir com nós que apresentem baixo grau (Figura 12B).

A correlação de grau dos nós i e j é feita por distribuição de probabilidade conjunta $P(k_i, k_j) = P(k_i) P(k_j)$. Podemos ainda calcular a assortatividade ou desassortatividade da rede como um todo, considerando:

$$r = \frac{\sum_i eii - \sum_i a_i b_i}{1 - \sum_i a_i b_i}$$

Se $r = 1$ a rede é considerada assortativa, enquanto que se $r = -1$, a rede é completamente desassortativa.

Caracteristicamente, redes assortativas são mais resilientes e apresentam hubs bem conectados, enquanto que redes desassortativas são redes mais vulneráveis com nós conexos a hubs esparsos (Figura 12).

A conectividade de uma rede também pode ser avaliada pela teoria da percolação. Essa teoria tem por objetivo estudar a conectividade da rede pela avaliação de sua arquitetura, caracterizando a distribuição do tamanho dos clusters e descrevendo como ocorre a transferência de informações, por exemplo, de A para B.

Redes aleatórias caracteristicamente apresentam baixa tendência em possuir pequenos clusters isolados e uma grande probabilidade em formar um componente conectado gigante. Como visto anteriormente, determinadas redes são altamente resilentes a deleção aleatória de nós. A variação na fração dos nós no maior componente da rede (componente gigante) é a forma mais fácil de calcular a resiliência. Imagine dois nós conectados na rede. Se estes nós pertencem a um componente gigante, há grande probabilidade de se comunicarem com uma extensa proporção de nós da rede. No entanto, nós que participam de pequenos componentes comunicam-se apenas com uma parte reduzida da rede. Essa capacidade de comunicação é responsável pela forma

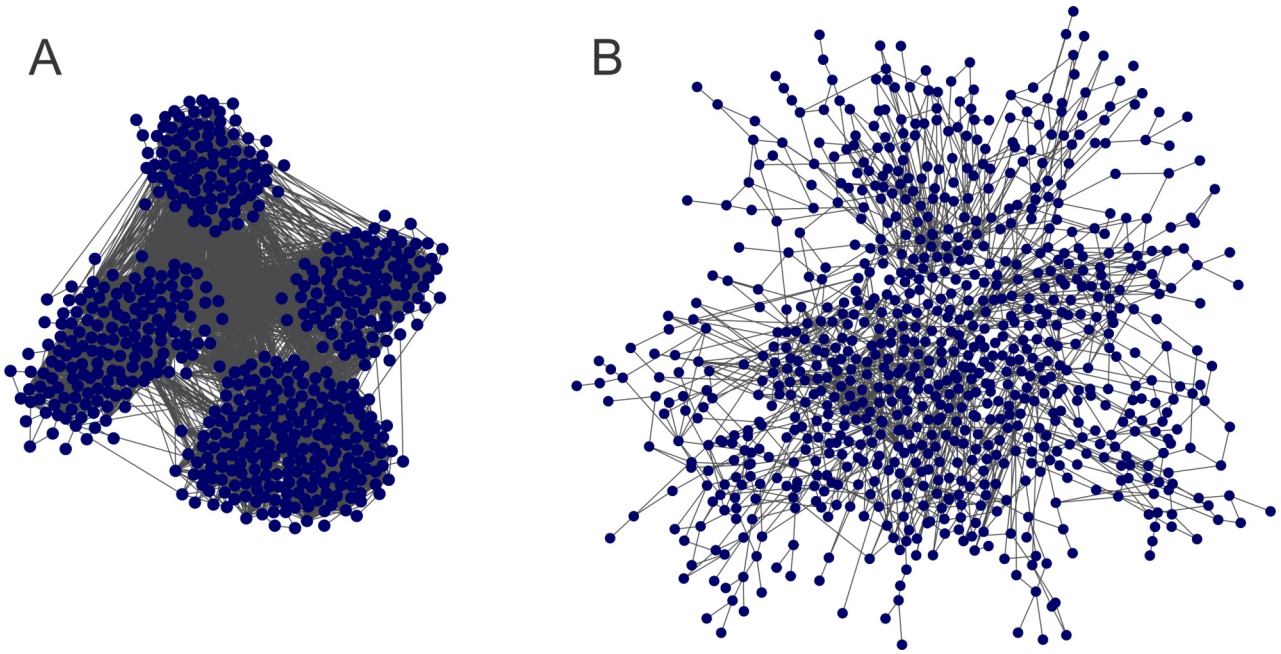


Figura 12: Ilustração representando em (A) uma rede assortativa com nós bem conectados que apresentam conexões com outros nós também fortemente conectados. Em (B), uma rede desassortativa, onde os poucos nós que apresentam mais conexões interagem com nós menos conectados, resultando em uma rede menos densa.

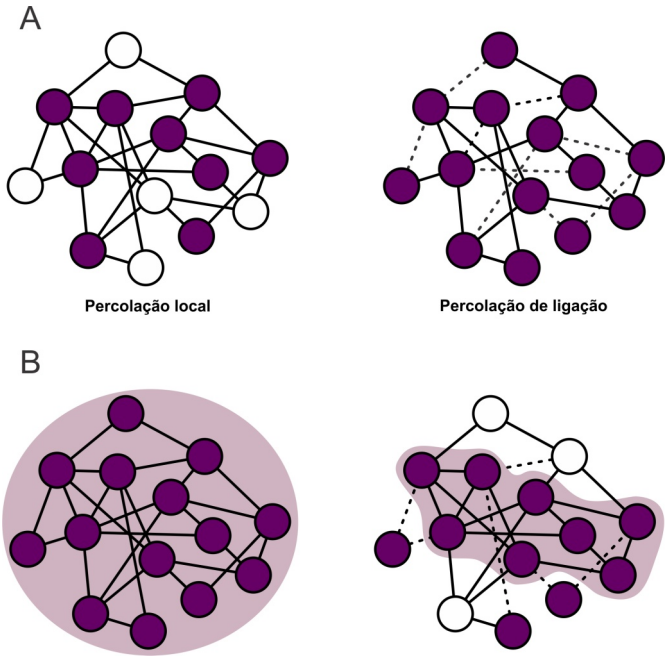


Figura 13: (A) Redes de percolação local e de ligação, onde os nós sólidos estão ocupados ou funcionais, enquanto que os nós brancos são desocupados ou falhos. (B) Representação do componente gigante. Após o surgimento de nós e conectores falhos, sua proporção é alterada e, por conseguinte, as possibilidades de transferência de informações.

como a informação é transferida de um ponto a outro. Dessa forma, associamos a resiliência com a percolação local (refere-se aos nós), enquanto que a percolação de ligação (refere-se aos conectores) está relacionada ao processo de dispersão (Figura 13A). Também podemos considerar os nós de uma rede como ocupados (funcionais) ou desocupados (falhos), dependendo da sua funcionalidade. A probabilidade de um nó estar ou não ocupado pode ser uniforme ou pode depender do grau do nó, sendo que os nós funcionais da rede formam o componente gigante em um modelo de percolação. Assim, os nós ou conectores falhos não participam da transferência de informação, e igualmente, não participam do componente gigante (Figura 13B). Dessa forma, ao observar a propriedade de percolação de um cluster, considerando uma probabilidade de ocupação variável, podemos determinar que isso afeta diretamente a conectividade de uma rede, tornando-a altamente resiliente ou não. Porém, ao combinarmos a percolação local e de ligação, teremos um modelo robusto contra falhas de nós ou conectores.

Os modelos de percolação são utilizados em muitas redes, porém um dos modelos mais interessante é o da dispersão de uma doença. Nesse modelo, cada nó representa o hospedeiro e os conectores representam a capacidade de transmissão da doença entre um hospedeiro e outro. O nó (indivíduo hospedeiro) está ocupado se for suscetível à doença, enquanto que um nó que representa um indivíduo que tomou a vacina seria considerado como desocupado. Da mesma forma, os conectores são considerados ocupados se há possibilidade de transmissão (Figura 14). Levando em conta este modelo, o início de uma epidemia representa a transição de percolação.

Apesar de ter sido originalmente desenvolvida com o objetivo de responder às perguntas em química orgânica, os modelos de percolação têm sido usados com sucesso para estudar diversos fenômenos, como transferência de sinal em neurônios e condutividade elétrica. Em 1987, Robert H. Gardner foi um dos primeiros pesquisadores a usar a teoria de percolação na Ecologia da Paisagem, sendo útil também na avaliação de corredores ecológicos e redes de incêndios florestais.

4. Propriedades de rede

Diversas propriedades são regularmente empregados na análise de redes biológicas, cada uma fornecendo informação sobre as interações e/ou componentes de um determinado sistema. estas propriedades podem ser referente a nós individuais, isto é, grau de nó ou node degree, ou podem contemplar a rede como um todo como é, por exemplo, o caso da modularização e do diâmetro da rede.

Em uma análise de biologia de sistemas, a análise estatísticas destas propriedades possui papel crítico na geração de dados conclusivos e confiáveis, constituindo-se

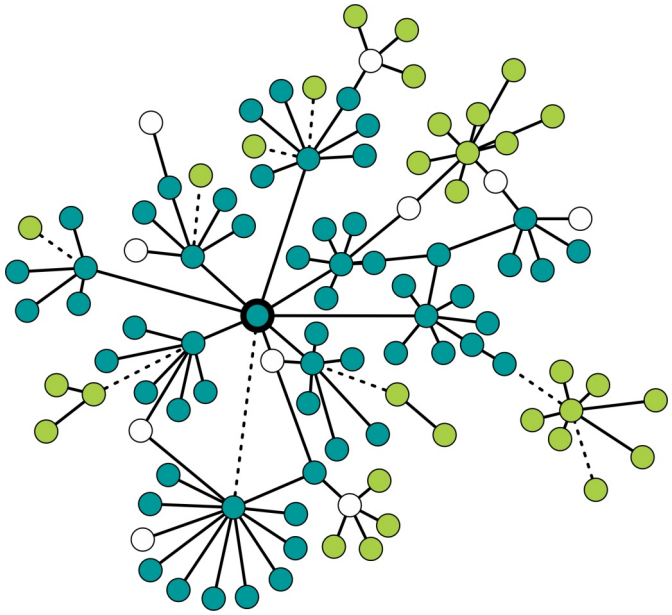


Figura 14: Modelo simplificado de dispersão de uma doença considerando um grupo de trabalho em uma empresa. Suponhamos que o indivíduo central contraiu uma doença viral de fácil transmissão, como a gripe simples. Assim, todos os indivíduos com os quais ele entrou em contato neste período também contraíram a doença (nós azuis), com exceção daqueles que foram vacinados (nós brancos). Neste caso, além de não contraírem a doença, também não a dispersaram. Os conectores pontilhados indicam que não houve interação física durante o período passível de contrair a doença entre o indivíduo saudável com o contaminado. Desta maneira, os indivíduos representados pelo nó verde claro, apesar de não terem sido vacinados, não contraíram a doença por não entrarem em contato com indivíduos contaminados.

assim em redes capazes de descrever com alto grau de fidelidade um determinado modelo biológico, de identificar alvos proteicos críticos na rede ou no desenvolvimento de caminhos moleculares.

Modularidade

Uma das principais características quando nos referimos a propriedades da topologia de redes é a chamada modularidade ou clusterização. O conceito de modularidade é antigo e já amplamente usado em outras áreas do conhecimento, como nas ciências sociais. Dentro das ciências biológicas, é um conceito comum nas áreas da biologia evolutiva, biologia molecular, biologia de sistemas e biologia do desenvolvimento.

Todas as ideias de modularidade giram em torno do conceito de padrões de conectividade, onde seus elementos constituintes estão agrupados em subconjuntos altamente conectados.

De forma geral, a modularidade é um princípio de união entre diferentes tipos de elementos e conexões naturalmente formado no meio biológico, como na interação entre indivíduos de mesma espécie. Um exemplo é a Pollenia rudis, uma espécie de mosca conhecida como cluster fly em decorrência de seu hábito de se agrupar com indivíduos da mesma espécie.

Este princípio é visto em todos os lugares, seja na nossa tendência de formar sociedades e grupos preferenciais de interação interpessoais ou na nossa tendência de organizar objetos por seu tipo, função e cores, dentre outros. Em nível molecular é visto, por exemplo, em elementos que atuam num mesmo processo biológico, como conjuntos de moléculas de RNA responsáveis pela degradação e síntese de ácidos nucleicos ou grupos de proteínas que atuam num mesmo processo biológico como a replicação de DNA e a transcrição gênica.

Existem dois tipos distintos de módulos:

i) Módulo Variacional: apresenta características que variam entre seus componentes e são relativamente independentes de outros módulos, porém possuem um número considerável de ligações com outros módulos;

ii) Módulo Funcional: possui elementos que normalmente atuam juntos em alguma função fisiológica distinta e são semiautônomos (quasi-autonomous) de outros módulos. Esses módulos compreendem a maioria dos módulos vistos em redes biológicas.

Módulos variacionais podem ser

exemplificados na Figura 15B e C, representando a formação de uma mandíbula de rato. Apesar de se tratar da diferenciação de um tecido, podemos usá-la como modelo variacional devido ao fato de diferentes proteínas e genes serem responsáveis pela formação de uma unidade estrutural única (o ramo ascendente e da região alveolar). Desta maneira, é uma unidade estrutural (um único osso) que se origina de diferentes módulos. Assim, o módulo variacional consiste uma integração de vários de genes que dividem efeitos pleiotrópicos entre si e que possuem poucos efeitos pleiotrópicos com outros clusters, sendo praticamente independente.

Módulos de genes de desenvolvimento embrionário, relacionados à diferenciação ou formação de padrões corporais, tendem a ser quase independentes de outros módulos, uma vez que erros na sua expressão ou atuação podem ser letais para o embrião. Por isso, esses módulos de desenvolvimento tendem a depender de elementos dentro do próprio grupo para sua expressão. Podemos visualizar um exemplo de um módulo funcional na Figura 15A.

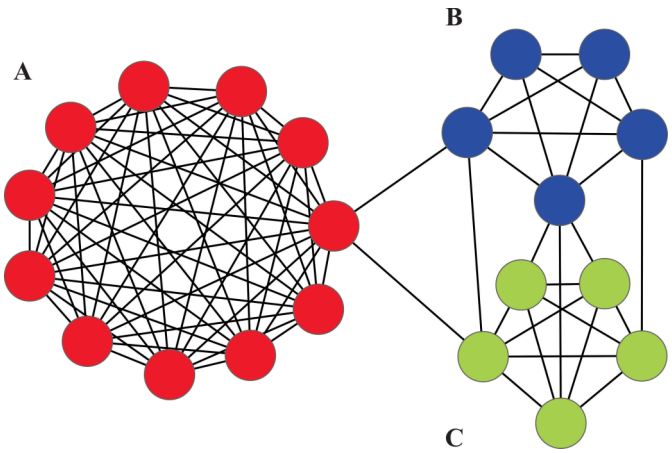


Figura 15: Exemplos de uma rede com diferentes módulos representados. Os módulos variacionais B (azul) e C (verde) se encontram praticamente independentes do módulo A (vermelho), porém possuem proteínas em comuns entre si. Contudo, o módulo A pode ser considerado funcional, uma vez que possui apenas uma conexão com cada outro módulo, sendo praticamente independente.

Ao determinarmos a quantidade e o tipo de módulos presentes em uma rede devemos levar em consideração o coeficiente de agrupamento (C_i) ou clusterização. O coeficiente analisa a tendência de um nó de se associar com seus vizinhos ("cliquishness"), onde "clique" é definido como um grafo maximamente conectado.

Como mencionado anteriormente, a clusterização é dada pela fórmula $C_i=2n/k_i(k_i-1)$, onde k_i é o tamanho da vizinhança de vértices (nós) do vértice i, e n é o número de conectores na vizinhança. Assim, quanto maior o coeficiente de clusterização, mais conectado é o cluster. Evolutivamente, as proteínas que compõem módulos altamente agrupados tendem a ser conservadas ou perdidas juntamente, caso haja uma variação dentro do grupo.

Outro conceito essencial para entender a formação de um cluster em um sistema biológico é a presença de hubs. Os hubs podem ser classificados em dois grupos:

i) party hubs, proteínas altamente ligadas dentro do seu próprio módulo (intra-módulo), ou seja, ligadas no mesmo tempo e/ou espaço,

ii) date hubs, que são hubs que se ligam a diferentes proteínas em diferentes módulos (inter-módulo), ou seja, diferentes tempo e/ou espaços, consequentemente apresentando um papel global na rede (Figura 16). Estes termos podem ainda receber denominações específicas no contexto do conteito de centralidades (ver adiante).

Os party hubs são componentes clássicos de módulos funcionais, uma vez que estes são quase independentes de outros módulos, enquanto date hubs são fundamentais para módulos variacionais, pois estes se ligam a outros módulos.

Assim, uma mutação em um party hub vai afetar principalmente as proteínas referentes ao seu próprio módulo, enquanto a mutação em um date hub (Figura 16) pode afetar vários módulos. Contudo, não existe diferença de importância entre party ou date

hub. A deleção de um hub em um módulo funcional pode ser tão letal quanto à deleção em um módulo variacional.

Baseado em dados estruturais, os hubs podem ser ainda classificados em singlish (com uma ou duas interfaces) e multi-interface (com mais de duas interfaces). Hubs com interface singlish somente se ligam a outras proteínas de maneira alternada e transitória, enquanto hubs multi-interface se ligam a diferentes proteínas concomitantemente.

Ontologias Gênicas

Nos últimos anos, o desenvolvimento e uso de técnicas de análise como microarranjos, ChIP-chip e espectrometria de massas e suas aplicações no estudo de cada vez mais organismos gerou um grande acúmulo de dados genômicos e proteômicos. A leitura e interpretação simples e concisa destes vem requerindo o desenvolvimento de novas abordagens, contexto no qual, em 1990, foi criado o chamado Gene Ontology Project.

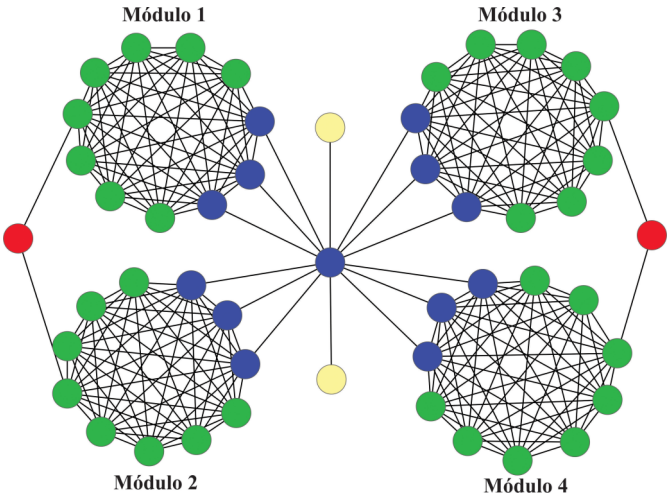


Figura 16: Diferentes tipos de centralidade em uma rede biológica. Em verde são apresentadas proteínas envolvidas em party hubs e encontradas em módulos. Em amarelo encontram-se as proteínas não-hub/não-gargalo, que são aquelas que não possuem alto valor de grau de nó ou betweenness, sendo consideradas componentes funcionais dos módulos. Em azul estão as proteínas hub-gargalo (date-hub) que possuem alto valor de grau de nó e de betweenness, sendo consideradas fundamentais para o funcionamento de redes. Em vermelho estão identificadas as proteínas do tipo gargalo, com alto valor de betweenness e essenciais na ligação entre módulos e processos biológicos.

Ontologia gênica refere-se ao produto de um determinado gene e a função ele desempenha na maquinaria celular. São classificadas em três níveis hierárquicos:

i) Componente celular, descrevendo a localização da proteína na célula;

ii) Processo biológico, referindo-se à série de eventos realizados por uma ou mais funções celulares;

iii) Função molecular, descrevendo a atividade que uma dada proteína desempenha no meio celular.

Essas informações são guardadas em forma de “anotações ontológicas”, onde cada uma possui um número de identificação e se encontram disponíveis em bancos de dados como www.geneontology.org.

Da mesma forma, essas anotações não são restritas a humanos, mas abrangem diversos organismos modelo como Mus musculus, Gallus gallus, Saccharomyces cerevisiae, Caenohabditis elegans e Escherichia coli, além de outros organismos não-modelo mas que já possuem alguma anotação.

De um modo geral, a ontologia gênica tem como função, em uma rede de interação proteína-proteína, agrupar proteínas que façam parte de um mesmo processo biológico. Em biologia de sistemas o emprego de ontologias gênicas pode se mostrar muito útil para direcionar a análise da rede, possibilitando a verificação dos tipos de processos biológicos existentes na rede e das proteínas presentes. Um modelo hipotético de como uma rede poderia se apresentar em termos de ontologias gênicas se encontra na Figura 17, onde diferentes nós poderiam estar relacionados a diversos processos.

A Figura 18 mostra um exemplo de aplicação de ontologias gênicas em uma rede

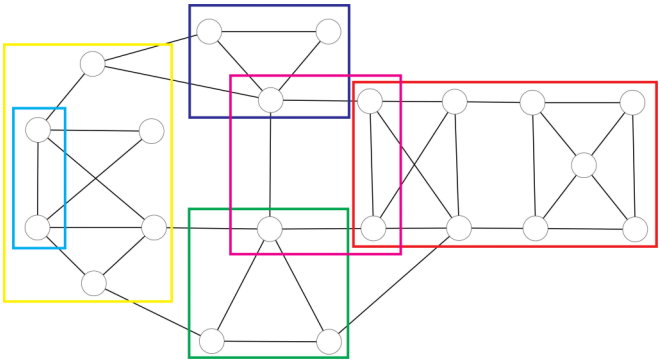


Figura 17: Modelo hipotético da presença de ontologias gênicas em uma rede. Na figura acima, cada cor representa um processo identificado. É importante ressaltar que uma proteína pode estar presente em mais de uma ontologia. Da mesma forma, uma ontologia pode estar dentro de outra. Como por exemplo, o quadrado amarelo poderia significar transcrição, enquanto o quadrado azul claro (inserido no amarelo) poderia significar apenas o complexo de iniciação da RNA polimerase II.

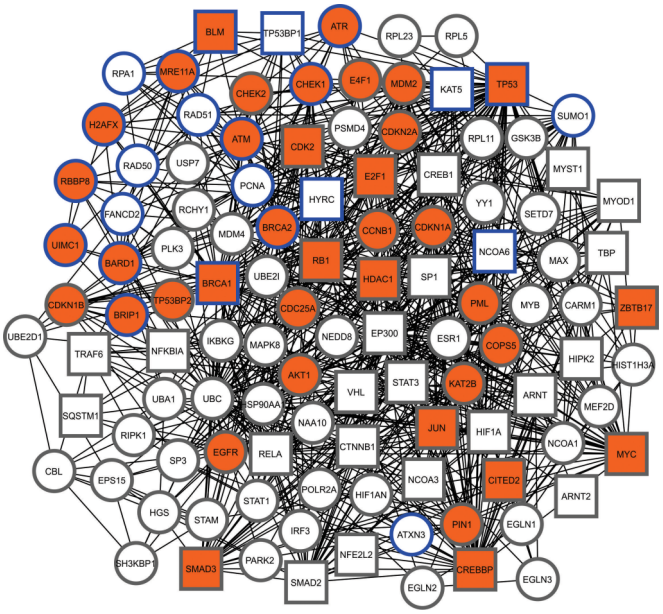


Figura 18: Exemplo de uma rede analisada pelo plugin BiNGO 2.44, o qual analisa as principais ontologias gênicas. A rede mostra três processos biológicos (GOs): *i*) Regulação do ciclo celular (nós de cor laranja); *ii*) Regulação positiva da transcrição (nós de de formato quadrado); *iii*) Resposta a dano de DNA (nós com a linha azul). É possível observar que mais de um nó compõe diferentes GOs.

biológica. Nessa análise foi utilizado o programa Biological Network Gene Ontology (BiNGO) 2.44, um plug-in do programa Cytoscape. É possível, assim, identificar proteínas ou genes com efeitos pleiotrópicos, saber: a proteína Tp53, a proteína breast cancer 1 (BRCA1) e a proteína bloom syndrome protein (BLM), as quais se encontram nas três ontologias da rede (reparo de DNA, regulação positiva da transcrição e ciclo celular).

Centralidades para nós

Como vimos até então, a grande vantagem da biologia de sistemas é permitir a visualização dos componentes moleculares de um sistema biológico de forma dinâmica e global. Contudo, quando falamos de uma rede, temos que levar em consideração todas suas estruturas, como hubs e módulos. Deste modo, o objetivo da análise de centralidades é procurar o elementos mais importantes na topologia geral da rede.

Grau de nó

Um dos parâmetros básicos de análise topológica é o parâmetro de grau de nó (ou node degree) é referente à quantidade de nós adjacentes (diretamente conectados) a outro determinado nó.

Esses nós que apresentam uma grande quantidade de conexões são chamados de hubs, os quais são conectados a outros hubs ou nós com menos conexões (Figura 16). Como veremos posteriormente, uma rede de livre escala é definida por uma lei de potenciação, o que significa que essa rede terá poucos nós altamente conectados. O grau de nó é referente ao valor distribuição de nó, $P(k)$, que informa a probabilidade de um nó ter k conexões, conforme visto em Estrutura de redes.

Numa visão biológica, podemos exemplificar um hub como uma proteína que se liga a várias outras e acaba possuindo uma função regulatória importante na rede. Normalmente, proteínas consideradas apenas hubs se encontram dentro de módulos. A

perda de conexões de uma proteína hub pode lhe tirar esta condição modular. Sua deleção em uma rede de interação proteína-proteína poderia afetar a ação de diversas proteínas vizinhas e até mesmo na formação de módulos.

Betweenness

O parâmetro denominado betweenness é definido como o número de caminhos mais curtos que passam por um único nó, estimando a relação entre eles. Por exemplo, para calcular o valor de betweenness um nó n é calculado o número de caminhos mais curtos entre i e j, e a fração deste caminhos que passam pelo nó n. Deste modo, um nó n pode ser atravessado por diversos caminhos alternativos, que ligam i e j.

Matematicamente, o valor de betweenness é dado pela seguinte fórmula:

$$Bet(n) = \sum_{i \neq n \neq j \in V} \frac{\sigma_{ij}(n)}{\sigma_{ij}}$$

onde σ_{ij} representam caminhos geodésicos entre os nós i e j, e $\sigma_{ij}(n)$ é o total destes caminhos mais curtos que passam por n.

Por exemplo, uma proteína com alto valor de betweenness apresentaria uma elevada capacidade de interação e/ou siinalização com outras proteínas, processos biológicos ou clusters. Uma proteína com tais características é chamada de bottleneck ou gargalo. Na Figura 16, temos dois exemplos de uma proteína com alto valor de betweenness.

Não existe uma maneira óbvia de se encontrar proteínas gargalo. Porém, é possível que rotas de sinalização possuam grande incidência de proteínas gargalo, uma vez que são necessárias para sinalização entre compartimentos e processos biológicos distintos. Contudo, proteínas gargalo não necessariamente possuem um grande número de interações com outras proteínas.

Closeness

O valor de closeness pode ser entendido como o caminho mais curto entre um nó n e

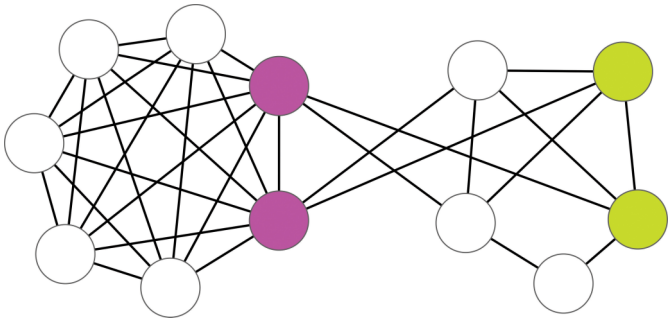


Figura 19: Caracterização de nós com diferentes valores hipotéticos de closeness. Os nós em roxo, dado as suas maiores conectividades com a rede no geral, possuem um valor maior de closeness, enquanto que os nós em verde, por possuírem poucas conexões com a rede, apresentam baixo valor de closeness.

todos os outros nós da rede, uma tendência de aproximação ou isolamento de um nó (Figura 19). Um alto valor de closeness indica que todos os outros nós estão próximos do nó n, enquanto que um baixo valor indicaria que os outros nós encontram-se distantes.

Este parâmetro é dado pela fórmula:

$$Clo(v) = \frac{1}{\sum w \in v^{dist(v,w)}}$$

onde o valor de closeness de um nó v [$Clo(v)$] é determinado através do cálculo e somatório dos caminhos mais curtos entre um nó v e todos outros nós w [$dist(v,w)$] dentro da rede.

Uma proteína com alto valor de closeness poderia ser considerada relevante para muitas proteínas, porém irrelevante para outras. Em termos biológicos, ela seria importante na regulação de muitas proteínas, porém sua atividade pode não influenciar outras. Ao compararmos essas informações com módulos podemos dizer que uma rede com uma média de closeness alta é mais provável de estar organizada como um módulo funcional, enquanto uma com baixo valor de closeness é mais provável de estar organizada como um módulo variacional.

Diâmetro

O diâmetro pode ser considerado um dos primeiros parâmetros referentes à

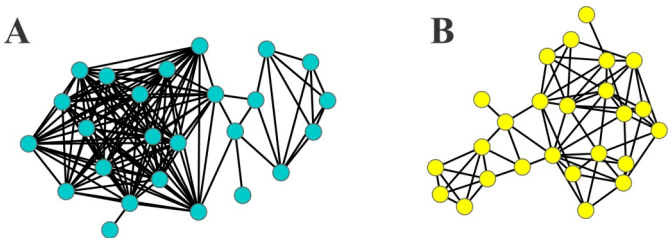


Figura 20: Em (A) uma rede com alto diâmetro e em (B) rede com baixo diâmetro. Pelo fato dos nós da figura A estarem mais interligados entre si, a rede é considerada mais "compacta", pois seus nós mais facilmente podem influenciar uns aos outros. Entretanto, em B, a rede possui muito menos conexões, portanto a deleção de um nó ira afetar a rede de um modo mais sutil.

"compactação", isto é, proximidade dos nós da rede. Ele indica a distância entre os dois nós mais afastados entre si de uma rede. Sendo assim, definimos que uma rede possui um alto diâmetro quando a distância geral entre os nós é muito ampla. Quando a distância entre os nós é pequena, então o diâmetro é baixo. Deste modo, uma rede com baixo diâmetro é considerada mais completa, uma vez que suas proteínas estão mais interligadas entre si.

Um baixo diâmetro pode indicar que as proteínas de uma determinada rede possuem uma maior facilidade de se comunicar e/ou influenciar umas as outras, apontando para uma relação funcional co-evolutiva (Figura 20).

Os parâmetros de centralidades podem ser alterados com a adição ou deleção de nós ou conexões na rede (Figura 21). Como já mencionado, em um sistema molecular, a perda de uma conexão pode ser considerada a mudança de um domínio, impedindo a ligação de duas proteínas ou a mudança de um produto gênico, criando proteínas anormais que não mais farão as mesmas conexões. Contudo, mudanças topológicas nas redes biológicas são processos normais durante a evolução. A deleção e a duplicação de um gene, assim como a perda de interações, sejam pela mudança estrutural ou de função, são processos muitas vezes selecionados e necessários para sobrevivência celular.

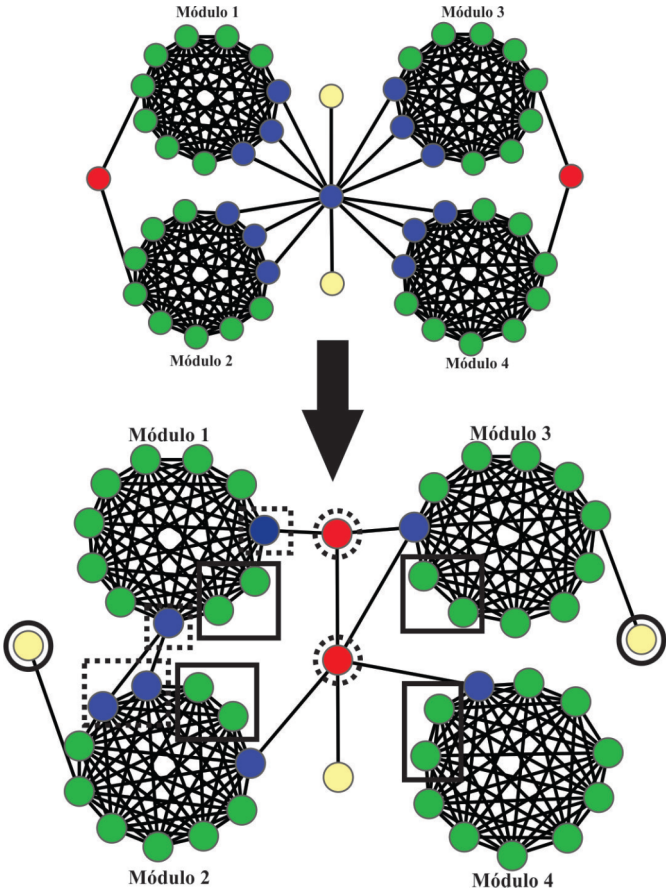


Figura 21: Modificações na topologia de rede podem alterar as centralidades. Devido à perda de conexões com nós fora do módulo, os nós marcados pelos quadrados foram transformados em party-hubs (nós verdes), deixando de ser hubs-gargalos (nós azuis). Porém, marcados pelos quadrados pontilhados, há nós que além de ganharem conexões, passaram a se ligar a outros módulos, saindo do estado de não-hub/não-gargalo para hub-gargalo (nós amarelos). Marcados por círculos, os nós antes gargalos (nós vermelhos), agora pela perda de uma conexão, se tornam não-hubs/não-gargalos. Por fim, os nós marcados pelos círculos pontilhados, devido à perda de muitas conexões (nó central) e ao ganho de uma conexão (nó acima), se tornam gargalos, perdendo os status de hub-gargalo e de não-hub/não-gargalo respectivamente.

Centralidade para conectores

Os elementos mais informativos de uma rede de interação podem ser avaliados através da análise da centralidade. Dentre as possíveis centralidades avaliadas, o

betweenness de um conector pode medir a influência de certos conectores no fluxo de informações entre os componentes da rede.

O betweenness de um conector e é simplesmente o número de caminhos mais curtos entre pares de nós que percorrem e. Se uma rede contém módulos que são conectados por poucos conectores intermodulares, então os caminhos mais curtos entre os diferentes módulos devem passar por estes poucos conectores. Assim, os conectores unindo módulos terão altos valores de edgebetweenness (Figura 22).

Neste caso, os pares de nós unidos pelos conectores serão de diferentes módulos. Se o valor de edgebetweenness de um conector é baixo, esse conector provavelmente fará parte do módulo, uma vez que dentro do módulo os nós são mais interligados entre si. Portanto, edgebetweenness é a frequência de um conector que se coloca sobre os caminhos mais curtos entre todos os pares de nós. Em uma rede proteica, um conector com alto valor de betweenness provavelmente representa o caminho mais curto de comunicação entre dois processos biológicos.

Como conectores com altos valores de betweenness são mais prováveis por posicionarem-se entre módulos, a remoção sucessiva destes conectores pode eventualmente isolar estes mesmos módulos. Essa desordem na rede, conforme será visto adiante, é conhecida como perturbação de conector.

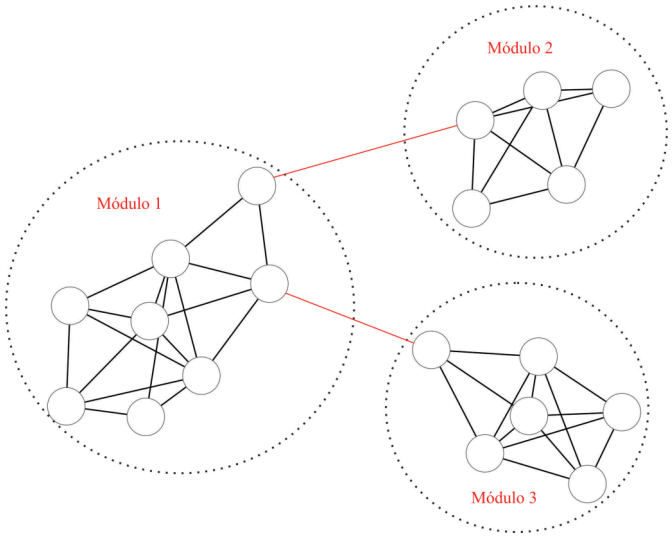


Figura 22: Representação de edgebetweenness. Conectores em vermelho apresentam valores altos de betweenness, pois representam o caminho mais curto do fluxo de informação entre os três módulos representados.

5. Tipos de redes

Rede Aleatória

Os matemáticos Paul Erdös e Alfréd Rényi iniciaram seus estudos sobre redes aleatórias em 1960. Este modelo de rede tem impulsionado o interesse de diversos cientistas ao longo dos anos por ser um dos primeiros modelos de rede descoberto. Porém, apesar de amplamente estudadas, redes aleatórias não capturam a realidade de um sistema biológico (Figura 23).

Essas redes consistem de N nós, com cada par de nós conectados (ou não) com probabilidade p, gerando uma rede de conexões aleatórias com aproximadamente $pN \cdot (N\text{-}1) / 2$. Dessa forma, o grau dos nós segue uma distribuição de Poisson com

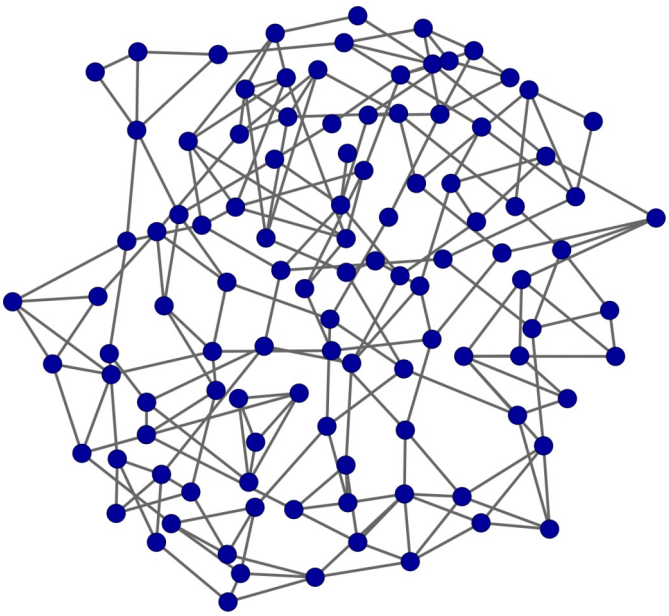


Figura 23: Ilustração de uma rede aleatória consistindo de 109 proteínas. A rede apresenta $P(k)$ 3,8. Observe que as conexões de cada nó são valores próximos a 4, o que está de acordo com $k \approx \langle k \rangle$.

máxima em <k> e a maioria dos nós apresentando aproximadamente o mesmo número de conexões $k \approx \langle k \rangle$, com grau próximo ao da média da rede. Raramente surgem nós que apresentam mais ou menos conexões que <k>.

Adicionalmente, redes aleatórias apresentam a propriedade "mundo pequeno" e distribuição de grau exponencial, sendo estatisticamente homogêneas.

Rede de livre escala

O modelo de rede de livre escala foi introduzido por Barabási e Albert em 1999 onde se observa que redes complexas, como as redes de citações de artigos científicos, redes metabólicas, redes socais e a World Wide Web apresentam distribuição de grau que segue uma lei de potência $P(k) \sim k^{-\gamma}$, $\gamma > 1$. Essas redes são consideradas como livres de escala (Figura 24) pois a lei de potência não permite uma escala característica.

Diferentemente da rede aleatória que apresenta um número fixo de N nós, as redes de livre escala apresentam uma ordem dinâmica de estruturação que permite o crescimento da rede pela adição de novos nós. Assim, a rede aleatória consiste de um sistema aberto que inicia com um pequeno grupo de nós e aumenta de tamanho exponencialmente no tempo devido à inserção de novos nós. A probabilidade deste novo nó se conectar a nós com grande número de conexões é maior, sendo chamada de conexão preferencial. Por exemplo, imagine que você está buscando um artigo sobre determinado assunto na Internet. Certamente os artigos que você encontrará mais facilmente serão publicações com alto grau de conexão por serem mais conhecidos e bem citados quando comparadas a publicações pouco citadas e, consequentemente, menos conhecidas.

Estes dois mecanismos, crescimento da rede e conexão preferencial originaram o algoritmo do modelo Barabási-Albert, que estabelece que o crescimento inicia-se como uma pequena rede, sendo que a cada instante de tempo um novo nó com m conexões é adicionado, onde a probabilidade do novo nó se conectar ao nó i que está previamente presente depende de k_i (grau de i):

$$\Pi(ki) = \frac{k_i}{\Sigma_j k_j}$$

Esse crescimento gera uma rede de livre escala com expoente de grau $\gamma = 3$. Após t instantes de tempo, temos uma rede com $N = t + m_0$ e m_t conectores.

As características da rede de livre escala a tornam uma rede que apresenta um pequeno número de nós altamente conectados (hubs), o que frequentemente determina suas propriedades. Como já mencionado, falhas na rede (ou remoção de nós aleatórios) apresentam poucas consequências, enquanto que o ataque aos nós altamente conectados tornará a rede fragmentada. Em sistemas biológicos, uma rede bioquímica apresenta alta resiliência contra mutações aleatórias, enquanto que os hubs podem ser usados como candidatos importantes para alvo de fármacos. Um exemplo disso seria a proteína EF-Tu. Esta pro-

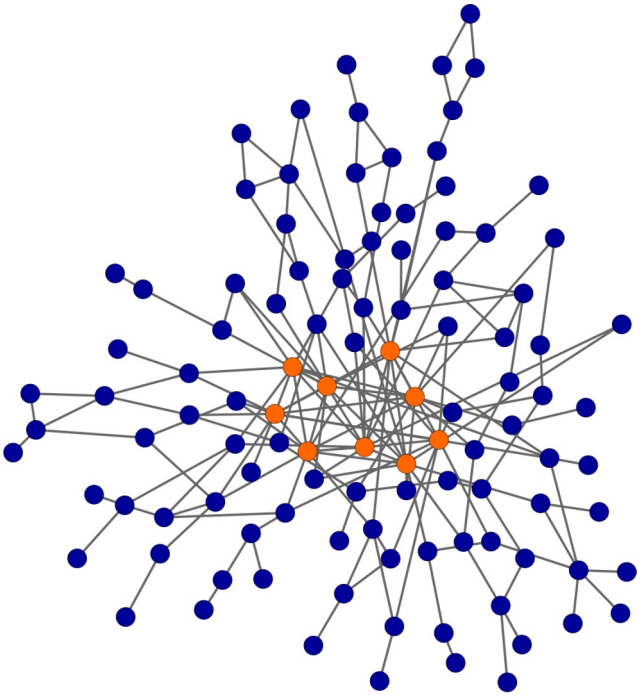


Figura 24: Ilustração de uma rede de livre escala consistindo de 109 proteínas, na qual o grau de distribuição segue uma lei de potência. Neste tipo de rede, as proteínas hubs (nós laranja) tem papel essencial na manutenção da integridade da rede.

teína tem papel essencial durante a elongação da síntese proteica,sendo inibida pelo antibiótico quirromicina, que impede que o complexo EF-Tu-GDP seja liberado do ribossomo.

Rede Hierárquica

Como já vimos anteriormente, uma rede pode ser avaliada pelo grau de agrupamento (clusterização) de seus nós. Na maioria das redes baseadas em um sistema real (chamadas de redes reais), como por exemplo, parte de uma via metabólica, como o coeficiente de clusterização é significativamente maior se comparado a redes aleatórias. Da mesma forma, ocorre a coexistência da propriedade de livre escala e clusterização nas redes reais, como redes metabólicas e de interação protéica. Contudo, grande parte dos modelos propostos para representar estas redes não consegue descrever a livre escala e a clusterização simultaneamente.

Adicionalmente, muitas redes reais apresentam módulos, ou seja, a rede é composta de subredes funcionalmente separáveis. Esses componentes separáveis apresentam densa conectividade entre os seus próprios nós, com conectividade mais dispersa em relação a componentes de outros módulos. Isso ocorre porque cada módulo apresenta a capacidade de executar uma tarefa identificável, diferente de outro módulo. Contudo, essa "separação" de tarefas não significa que um módulo é independente de outro, mas sim que tem funções distintas.

Dessa forma, é necessário combinar a propriedade de livre escala, o alto grau de agrupamento e a modularidade de uma forma interativa, gerando a rede hierárquica. A estrutura hierárquica é convencionalmente representada por um dendrograma ou uma árvore e atua relacionando os nós mais próximos na rede, conforme Figura 25. Essas redes podem ser formadas basicamente pela duplicação de clusters e repetidas indefinidamente, integrando uma topologia livre de escala com alta modularidade, resultando em um coeficiente de clusterização independentes do tamanho do sistema. Muitas vezes, em redes reais, a modularidade não apresenta um limite claro, sendo reconhecida principalmente por nós altamente conectados entre si e conectados a outros módulos.

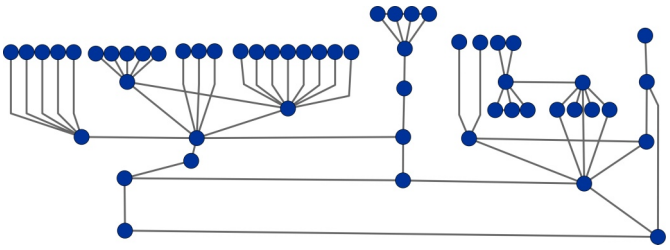


Figura 25: Ilustração de uma rede hierárquica consistindo de 55 proteínas em modelo de dendrograma onde é possível observar sua modularidade intrínseca.

A principal característica dessas redes que não é compartilhada por redes aleatórias ou de livre escala é a hierarquia intrínseca, sendo representada também na sua arquitetura. Essa característica hierárquica pode ser, ainda, analisada quantitativamente, como observado por Dorogovtsev et al. (2002), que construiu um gráfico de livre escala determinístico, na qual o coeficiente de clusterização de um nó que possui k conexões segue a lei de escala $C(k) \sim k^{-1}$.

Portanto, o modelo de rede hierárquico integra uma topologia livre de escala com alta modularidade, resultando em um coeficiente de clusterização independentes do tamanho do sistema.

6. Perturbação e tipos de conectores

Como visto anteriormente, um grafo consiste de um conjunto de nós e um conjunto de conectores que conectam esses nós. Portanto, os nós são as entidades de interesse e os conectores representam as relações entre as entidades.

Quando tratamos de sistemas biológicos, podemos levar em consideração diferentes entidades como, por exemplo, DNA, RNA, metabólitos, pequenas moléculas e/ou proteínas. Estes componentes biológicos não atuam isoladamente, mas sim dependem

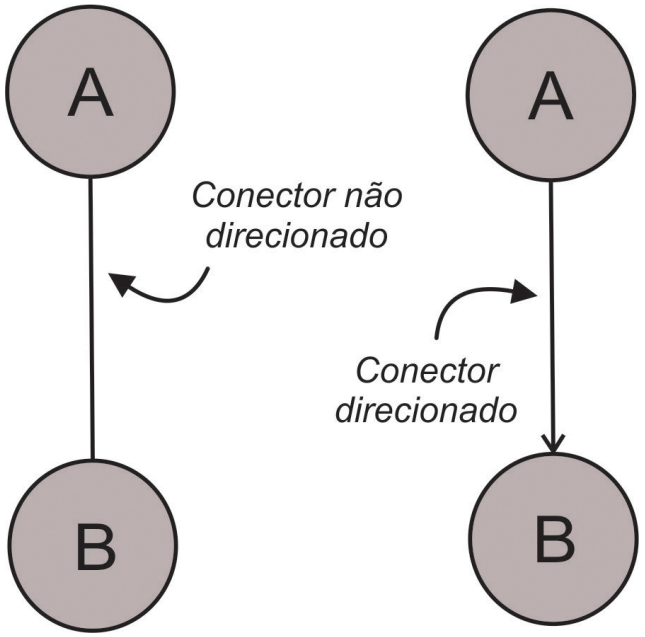


Figura 26: Representação de um conector não direcionado e um direcionado.

da interação com outros componentes. E para que ocorra essa interação (comunicação) é necessária a presença de conectores.

Conectores podem ser interações físicas, bioquímicas ou funcionais. Por exemplo, em redes metabólicas, conectores podem ser reações que convertem um metabólito em outro ou enzimas que catalisam essas reações; em redes de regulação gênica, conectores podem representar a ligação física de um fator de transcrição nos elementos regulatórios; em redes de doenças, conectores podem representar as mutações genéticas associadas à doença; e em redes proteicas, os conectores podem ser ligações físicas entre as proteínas.

Como apresentado anteriormente, as redes podem ser direcionadas e não direcionadas. Esse comportamento da rede depende da natureza da interação e, obviamente, da direcionalidade dos conectores (Figura 26). Em redes direcionadas, a interação entre dois nós tem uma direção bem definida que representa, por exemplo, a direção do fluxo do substrato ao produto em uma rede metabólica. Em redes não direcionadas, a ligação não tem uma direção definida, tal como a interaçao física entre proteínas.

Na abordagem da biologia de sistemas tão importante quanto conhecer os nós que interagem entre si em uma rede é compreender, por exemplo, que tipo de interação pode ocorrer na rede em questão, quais conectores são mais relevantes à rede e qual o impacto da perturbação de um conector. Nesta seção iremos discutir os tipos de conectores entre diferentes componentes de uma rede envolvendo proteínas e as consequências da ruptura nestas conexões.

Interaçao entre proteína-proteína

A interação proteína-proteína é comum e crucial a vários processos celulares, tais como na ligação enzima-inibidor e na interação antígeno-anticorpo.

Os diferentes tipos de complexos protéicos têm sido definidos na literatura como obrigatórios e não obrigatórios. No complexo obrigatório, as proteínas não podem funcionar separadamente, diferindo do complexo não obrigatório onde as proteínas associam-se e dissociam-se dependendo de fatores externos, podendo também exercer funções fora do complexo.

De acordo com a estabilidade e o mecanismo de formação do complexo, incluindo o tipo de conexão entre as proteínas, as interações podem ser conceitualmente separadas em dois grupos: aquelas que são permanentes e aquelas que são temporárias. E, embora não exista um limite bem definido para essa separação, tendências têm sido observadas em relação a suas propriedades biológicas (Figura 27).

Em relação à estrutura, por exemplo, interações temporárias são caracterizadas por interfaces proteicas pequenas, enquanto que as interfaces de proteínas interagindo permanentemente são maiores. Consequentemente, complexos proteicos com interfaces maiores tendem a apresentar um maior grau de mudança conformacional após a ligação. Além disso, componentes de complexos permanentes tendem a ser co-expressos e mais estáveis. Esta estabilidade gera uma pressão seletiva maior e em função

disso, uma taxa evolutiva mais lenta.

Como será já discutido, interação transitória tende a ser date, isto é, as proteínas podem se conectar em diferentes tempos e a interação permanente tende a ser party, isto é, conexão proteica forte e constante.

As proteínas com conectores permanentes existem somente em sua forma complexada e são muito estáveis, enquanto aquelas com conectores transitórios possuem a capacidade de associação e dissociação in vivo. Dentre as proteínas com conectores transitórios, há aquelas em que a associação/dissociação é resultante de uma conexão com baixa afinidade, porém constante (interações temporárias fracas) e aquelas em que a associação/dissociação é desencadeada por um processo ativo (interações temporárias fortes) como, por exemplo, uma mudança conformacional ocorrida em consequência de um fator ligante.

A diferença entre as interações acima citadas é distinguida puramente pelas propriedades da estrutura da interface proteica, isto é, da superfície de contato das proteínas. Essas propriedades conferem afinidade e especificidade, e são determinadas principalmente por forças intermoleculares como complementaridade estérica, força eletrostática, interação hidrofóbica e ligações de hidrogênio.

A complementaridade estérica otimiza as interações de van der Waals entre o complexo. Normalmente, estas interações de fraca energia ocorrem em função da polarização transiente de ligações carbono-hidrogênio ou carbono-carbono e, apesar de fracas, são extremamente importantes para

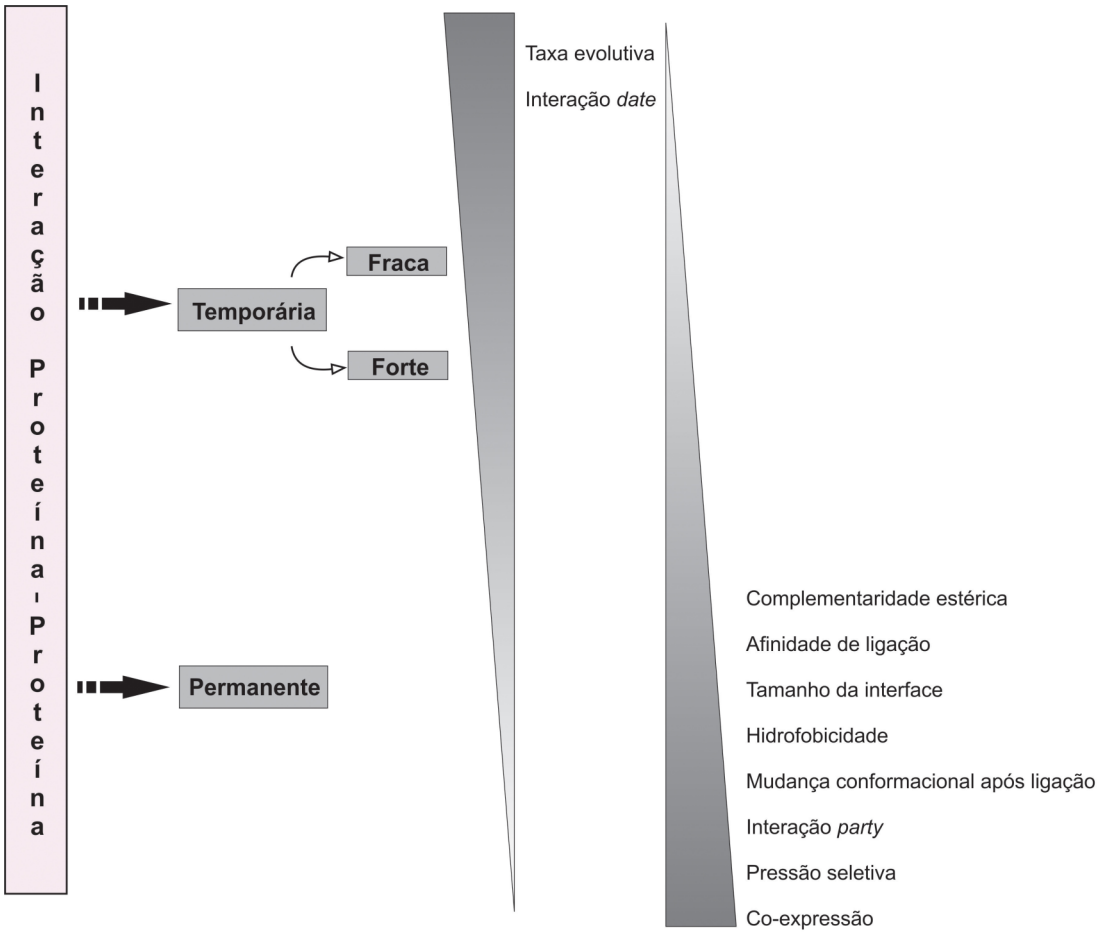


Figura 27: Modelo esquemático representando os diferentes tipos de interações proteína-proteína e as propriedades biológicas relacionadas. Quanto maior o tamanho da base e a intensidade da cor do triângulo, maior é a relação entre o modo de interação protéica e a propriedade biológica.

o processo de reconhecimento intermolecular pois crescem em intensidade com a área de interação. Complexos com conexões permanentes exibem alta complementaridade estérica nas proteínas em contato, enquanto complexos com conexões temporárias demonstram baixa complementaridade.

Como as interações de van der Waals, as interações hidrofóbicas são pontualmente fracas e ocorrem em função da interação entre cadeias ou subunidades apolares. Os complexos com conexões permanentes normalmente persistem no estado ligado, sendo a força hidrofóbica mais significativa. Já em conectores transitórios, a alta hidrofobicidade se torna desfavorável, pois esses complexos permanecem ligados por menos tempo.

As forças de atração eletrostáticas são aquelas resultantes da interação entre dipolos e/ou íons de cargas opostas e representam força significativa na interação proteína-proteína, podendo definir o tempo de vida do complexo.

Dentre as forças intermoleculares discutidas acima, o fator dominante da interação permanente entre proteínas consiste nas interações hidrofóbicas, enquanto várias forças participam de interações temporárias entre proteínas. Além disso, proteínas interagindo de forma temporária possuem interfaces que são menores em tamanho do que as interfaces de proteínas permanentes, os aminoácidos que compõem a interface e a proporção de resíduos hidrofóbicos não diferem drasticamente do resto da superfície proteica e as interfaces são levemente ricas em grupos polares neutros e em água.

O tipo de interação também confere graus diferentes de restrição (pressão seletiva) na evolução da proteína. Proteínas com interação permanente tendem a evoluir em uma velocidade menor comparada a proteínas que formam complexos temporários, bem como possuir pressão seletiva maior e menor plasticidade em sua sequência.

Evidências sugerem que o modelo duplicação-divergência aplica-se à evolução das redes proteicas. Uma das predições é que na duplicação das proteínas algumas ou todas as conexões podem ser herdadas da proteína ancestral. Consistente com esta hipótese, proteínas parálogas tendem a compartilhar padrões de interação em uma frequência maior do que a esperada ao acaso. No entanto, tem sido proposto que depois que a duplicação gênica ocorre, as interações entre as proteínas são rapidamente perdidas. Portanto, duplicações recentes são mais prováveis por compartilhar interações, comparadas a duplicações mais ancestrais.

Outra distinção a cerca da interação proteica refere-se à interação funcional e interação física. A interação funcional pode ou não corresponder a uma interação física direta em algum processo biológico. Assim, na interação física, a proteína A conecta-se a proteína B e, na interação funcional, a proteína A atua com a proteína B. Como exemplo de interação funcional podemos imaginar dois produtos gênicos que interagem em uma mesma via em um processo biológico, mas não se conectam fisicamente.

O tipo de interação tem um papel importante na determinação do comportamento das proteínas. Como já vimos, hubs são proteínas envolvidas em um grande número de interações (altamente conectadas) dentro de uma rede proteica. Algumas proteínas hub são altamente co-expressas com outras proteínas do módulo, o que implica na existência de complexos estáveis (permanentes). Outras proteínas possuem expressão independente, sugerindo a ligação com proteínas em diferentes tempos, de modo transitório. Esses hubs são classificados como party e date hubs, respectivamente.

Na construção de redes proteicas, a diferenciação entre complexos permanentes e transitórios tem importantes implicações. Por exemplo, na prospecção de novos fármacos, a alteração do padrão de interação entre proteínas temporárias por modulação farmacológica ocorre mais facilmente em comparação a proteínas que formam

complexos permanentes. Portanto, uma rede de interação proteica não é um processo estático, mas sim corresponde a um constante fluxo de informações. Por conseguinte, na análise de dados de interação proteína-proteína a discriminação das características da interação e/ou o uso de centralidades de conectores é fundamental para obter modelos mais realísticos.

Interaçao entre proteína-ácidos nucléicos

Proteínas que se ligam a ácidos nucléicos têm um papel central em todos os processos regulatórios que controlam o fluxo de informação genética. Por exemplo, proteínas podem inibir, ativar e coordenar a transcrição do DNA, auxiliar e manter o empacotamento e o rearranjo do DNA e o processamento do RNA, coordenar a replicação do DNA, promover a síntese de proteínas e sinalizar o reparo do DNA, entre outros.

Esses possíveis papéis fisiológicos são determinados pela afinidade e especificidade da interação DNA-proteína, que é a habilidade da proteína em distinguir seu sítio de ligação do restante do DNA. Estas propriedades dependem de interações precisas entre a sequência de aminoácidos da proteína e os nucleotídeos do sítio específico de ligação do DNA.

As proteínas que se ligam a ácidos nucléicos podem ser, de forma simplificada separadas em três grupos de acordo com a função:

i) enzimas, onde a principal função da proteína é modificar a organização do ácido nucléico, como no caso das endonucleases, glicosiltransferases, glicosilases, helicases, ligases, metiltransferases, nucleases, polimerases, recombinases, topoisomerases, translocases e transposases, entre outras;

ii) fatores de transcrição, onde a principal função da proteína é regular a transcrição e a expressão gênica como por exemplo, TFIIA, TFIIB, TFB, entre outros;

iii) proteínas estruturais que liga o DNA, que têm como principal função suportar a estrutura e a flexibilidade do DNA ou agregar outras proteínas, por exemplo, proteínas centroméricas, proteínas envolvidas no empacotamento e na manutenção/proteção do DNA, proteínas de reparo, proteína envolvidas na replicação e proteínas teloméricas, entre outras.

A interação proteína-proteína também é necessária para uma eficiente interação entre proteínas e ácidos nucléicos. A interação proteína-proteína com o DNA pode ocorrer de três modos de acordo com a direção e o eixo da dupla hélice do DNA (Figura 28):

i) a direção da interação entre as proteínas e o eixo da dupla hélice é perpendicular;

ii) a direção da interação da proteína é paralela ao eixo da dupla hélice;

iii) ambos os modos de interação são observados ao mesmo tempo.

Assim como na formação de complexos protéicos, discutido anteriormente, a formação de complexos DNA-proteína ou RNA-proteína também envolve forças intermoleculares, tais como van der Walls, força eletrostática, interação hidrofóbica e ligações de hidrogênio.

A região da proteína que reconhece a sequência do ácido nucléico é denominada motivo da proteína. Os motivos hélice-volta-hélice, dedo de zinco e zíper de leucina são os mais comuns encontrados nas proteínas que interagem com ácidos nucléicos.

O motivo hélice-volta-hélice é um dos elementos normalmente encontrados nos fatores de transcrição e nas enzimas de procariotos e eucariotos, sendo formado por duas α-hélices conectadas por uma volta. O motivo liga-se a cavidade maior do DNA e, em muitos complexos, o contato direto é feito entre a cadeia de aminoácido e a sequência de bases do ácido nucléico.

Já o motivo dedo de zinco é encontrado principalmente em fatores de transcrição de eucariotos. Um dedo de zinco é composto por duas

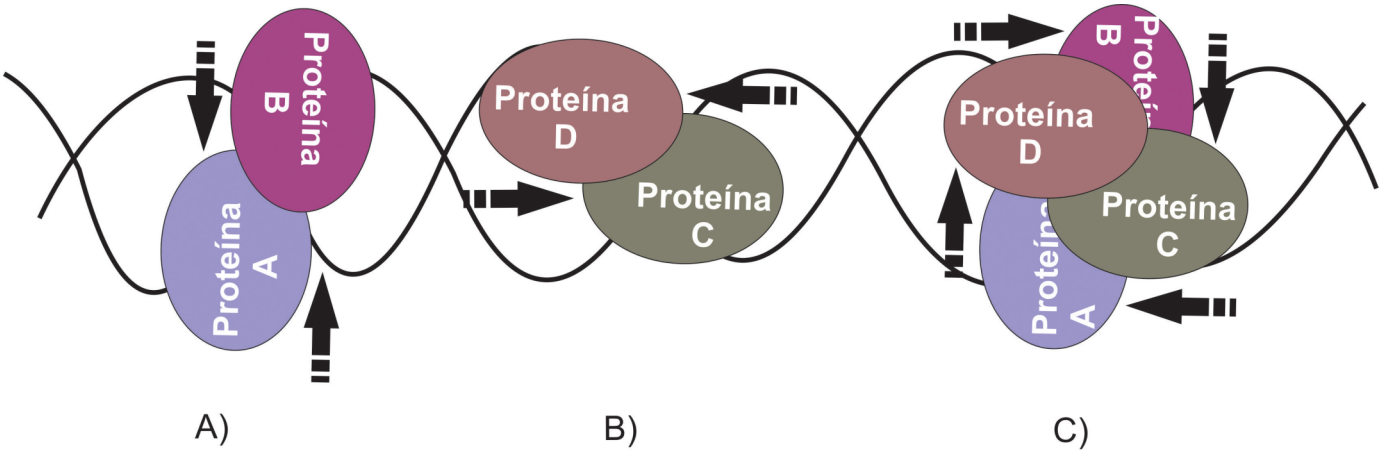


Figura 28: Modos de interação proteína-proteína com a dupla hélice do DNA. A) perpendicular; B) paralela e C) ambas as direções são observadas.

folhas β-pregueadas antiparalelas e uma α-hélice, sendo o íon zinco fundamental para garantir a estabilidade deste tipo de domínio. Subunidades proteicas contêm múltiplos dedos de zinco que se enrolam no DNA formando uma espiral, inserindo a α-hélice na cavidade maior do DNA.

Fatores de transcrição de eucariotos e procariotos também podem conter o motivo zíper de leucina, encontrado em proteínas regulatórias. Esse motivo é formado por duas α-hélice paralelas unidas por resíduos de leucina.

A estrutura do zíper de leucina pode ser dividida em duas partes: a região de dimerização e a região de ligação ao DNA. A dimerização é mediada pela formação de uma estrutura enrolada na região carboxi-terminal de cada hélice com sete resíduos de leucina. A região que se liga ao DNA, também conhecida como região básica, é encontrada na região amino-terminal da hélice que se projeta na cavidade maior do DNA. Embora motivos de diferentes famílias de DNA sejam similares estruturalmente, pouca homologia é observada fora do motivo. Há pouca similaridade entre motivos de diferentes famílias de proteínas e esta variação permite, portanto, o reconhecimento de diferentes conjuntos de sequências de DNA. Além disso, a posição do domínio dentro da cavidade maior do DNA também varia, refletindo a necessidade funcional e estrutural de cada proteína.

A afinidade e a especificidade na ligação de proteínas ao DNA não podem ser endereçados somente a alguns resíduos de aminoácidos, mas o envolvimento de toda a proteína deve ser considerado. Por exemplo, a maioria das proteínas que se ligam ao DNA possuem domínios desordenados que contribuem para o reconhecimento do DNA em vários níveis.

Proteínas com domínios desordenados são proteínas que não apresentam estrutura secundária e terciária sob condições fisiológicas e na ausência de ligantes naturais. Essas proteínas possuem alta especificidade e baixa afinidade na interação, são capazes de interagir com mais de uma proteína e alvos de modificações pós-traducionais, possuindo a capacidade de manter sua função mesmo em ambientes extremos. Na interação com o DNA, o domínio desordenado da proteína não é crucial à formação do complexo, mas pode influenciar o reconhecimento da sequência do DNA, conferindo seletividade e afinidade de ligação.

Além da característica das cavidades na molécula de DNA, da presença de motivos específicos nas proteínas ou ainda da ocorrência de domínios desordenados, outros fatores podem influenciar a interação do DNA-proteína, tais como a flexibilidade e a afinidade da proteína pelo DNA e presença de água no meio.

Muitas proteínas são flexíveis ao ponto de alterar sua conformação quando se ligam ao DNA, enquanto outras são conhecidas por alterar a conformação do DNA após a ligação. A afinidade da interação entre o DNA e uma proteína tende a estar relacionada à relevância funcional da proteína. Por exemplo,

a afinidade de um fator de transcrição por seu sítio de ligação é proporcional à ativação que ele exerce. Ainda, alguns contatos mediados por água foram observados entre proteínas e o DNA, participando de redes de ligações de hidrogênio que conferem estabilidade ao complexo.

Interação entre proteínas e pequenos compostos

Considerando-se que a interação proteína-proteína normalmente envolve superfícies relativamente grandes, pode-se imaginar que moléculas menores não seriam efetivas na modulação da ligação dos complexos por apresentarem áreas menores e, por conseguinte, interações menos intensas. Contudo, ao empregarmos estruturas químicas diferentes de aminoácidos, podemos não só compensar esta redução na área de contato mas produzir moléculas com afinidade maior do que os próprios ligantes fisiológicos envolvidos do processo de interesse.

Adicionalmente, estas moléculas de baixa massa molecular tendem a apresentar muitas vantagens terapêuticas em relação à proteínas, dentre as quais se destaca sua maior estabilidade metabólica e consequente maior biodisponibilidade. Podem atuar diretamente – via inibição da interface proteína-proteína – ou indiretamente – via ligação a um sítio alostérico que induz uma mudança conformacional do alvo da proteína ou da molécula associada.

A busca de novos fármacos deve levar em conta o tipo de complexo protéico alvo. A formação de complexos permanentes pode ser considerada uma continuação do enovelamento da proteína, sendo o dobramento final das subunidades parte deste processo. Assim, esse tipo de complexo é menos propenso à modulação farmacológica, sendo mais interessante explorar o processo de dobramento em si como alvo de pequenos compostos. Já as interfaces das proteínas de complexos temporários são alvos efetivos ao

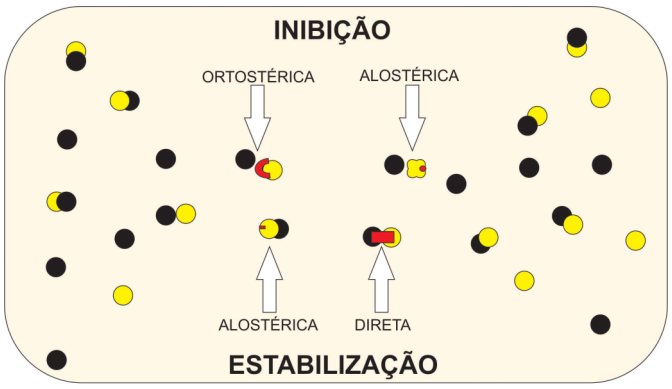


Figura 29: Dois principais mecanismos de modulação da interação proteína-proteína utilizando pequenos compostos. Diferentes proteínas são apresentadas em preto e amarelo. Pequenos compostos são apresentados em vermelho.

planejamento de novos moduladores terapêuticos.

Para que pequenas moléculas modulem a interação proteica, estratégias têm sido estabelecidas e dois principais mecanismos do controle regulatório têm sido utilizados: a inibição e a estabilização (Figura 29). Das estratégias mais exploradas, destaca-se a inibição da interação proteína-proteína.

O modo de ação da maioria dos inibidores de interação proteica é baseado na ligação direta de uma pequena molécula à superfície de interação da proteína ligante, interferindo diretamente nos hot spots críticos da interface e competindo com a proteína original. Esse tipo de inibição é conhecido como ortostérica. Na inibição alostérica, pequenos compostos ligam-se a sítios diferentes, causando mudança conformacional suficiente para interferir na ligação da proteína ligante (Figura 28).

Pequenas moléculas estabilizadoras da interação proteína-proteína também demonstram dois modos gerais de ação Primeiro, um estabilizador pode ligar-se a uma única proteína, no qual aumenta a afinidade de ligação mútua das proteínas do complexo de um modo alostérico. Segundo, a molécula estabilizadora liga-se à superfície do complexo protéico, fazendo contato com ambas as proteínas ligantes e aumentando a afinidade de ligação mútua entre elas. Assim,

a inibição estabilizadora pode ser denominada alostérica (ligada a uma proteína) ou direta (ligada ao menos a duas proteínas).

A ativação por pequenos compostos é, normalmente, um processo mais intrincado pois, além da ligação, é necessário o correto desencadeamento da cascata de ativação. Compostos que induzem a interação proteica são chamados de dimerizadores. Inúmeras vias de sinalização celular iniciam a partir da dimerização proteína-proteína. A principal ideia do uso de dimerizadores é a indução de interação entre duas proteínas por pequenas moléculas que levam a ativação da via de sinalização celular. Na literatura científica foi observado que dimerizadores podem induzir proliferação celular, transcrição e apoptose.

Perturbação dos conectores

Perturbações podem ocorrer em todos os sistemas, e em sistemas biológicos não é diferente. Nos interatomas, essas perturbações podem variar desde a remoção de um ou mais nós até a remoção de conectores. Desta forma, as consequências na estrutura e na função do sistema irão diferir drasticamente dependendo do tipo de perturbação ao qual a rede foi exposta. Como exemplo, podemos imaginar uma rede de proteínas que confere um fenótipo específico (Figura 30).

A remoção do nó não somente incapacita a função deste, mas também a de outros nós, causando a ruptura nas vias de todos os nós vizinhos. Uma perturbação no conector, que remove uma ou poucas interações mas deixa o restante da rede intacta e funcionando, pode ter efeitos mais sutis no sistema, não necessariamente alterando o fenótipo. Contudo, a consequência do desarranjo da rede após a remoção de nós ou de conectores depende da importância do nó e do conector à rede. Essas informações de conectores e nós mais informativos de uma rede podem ser obtidas, por exemplo, pela análise da resiliência e percolação da rede, vista anteriormente.

A distinção entre modelos de remoção de nó e perturbação de conectores - alteração interação-específica e conector-específica (edge-specific ou "edgetic"), respectivamente - pode providenciar novas pistas nos mecanismos básicos de doenças humanas, tais como diferentes classes de mutações que levariam a modos dominantes ou recessivos de herança genética.

Em uma rede proteica, a remoção de um nó pode representar a remoção de uma proteína, causado por uma mutação crítica no gene que desestabiliza a estrutura da proteína. Já a remoção de um conector pode representar uma mudança específica em distintas interações bioquímicas e biofísicas, preservando certos domínios da proteína.

Em relação a genes envolvidos em múltiplas doenças, foi demonstrado que

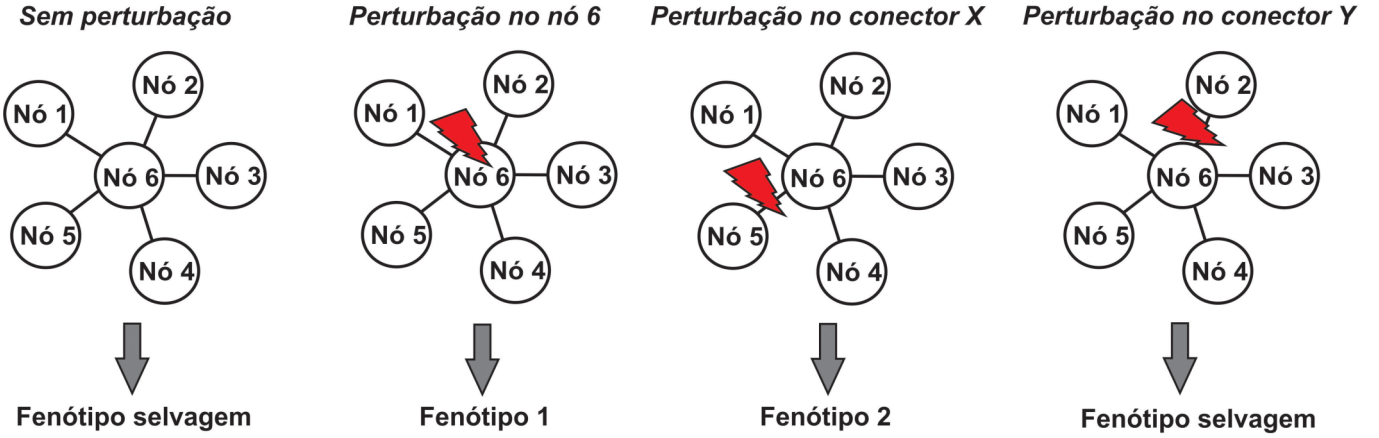


Figura 30: Rede hipotética de proteínas relacionada a um fenótipo específico representando diferentes tipos de perturbação e suas consequências. Neste exemplo o nó 5 e o conector entre os nós 5 e 1 são essenciais à manutenção do fenótipo selvagem.

alelos edgetic responsáveis por diferentes doenças consistem em distintas perturbações edgetic que, por sua vez, tendem a estar localizados em diferentes domínios de interação proteica, conferindo fenótipos diferenciados.

Pesquisadores analisaram cerca de 50.000 alelos mendelianos associados a doenças genéticas hereditárias e observaram que aproximadamente a metade foi potencialmente edgetic. Nesta análise foram consideradas deleções e mutações truncadas dentro dos domínios da proteína que grosseiramente desestabilizaram a estrutura da proteína, como remoção de nó; e mutações com alteração em quadro de leitura que afetaram sítios de ligação específicos e mutações truncadas que preservaram certos domínios da proteína como perturbação edgetic. Alelos truncados foram menos propensos a expressar proteínas estáveis em comparação a alelos que alteraram o quadro de leitura, podendo diferir doenças hereditárias mendelianas envolvendo remoção de nó versus perturbação edgetic.

Um alelo edgetic pode ser identificado pela falta de um subconjunto de interações, quando possuem defeitos nas interações provavelmente devido a mudanças específicas dentro ou próximo a sítios de ligação da proteína ou quando fenótipos in vivo diferem daqueles causados por perturbações nulas (genótipos nulos).

Dependendo da rede, o fenômeno de perturbação de um único conector pode ser mais provável do que da remoção de um nó. Dependendo do conector rompido, o impacto à rede pode ser maior, pois diferentes conectores (interações) têm diferentes níveis de importância (vulnerabilidade). Conectores com alto valor de edgebetweenness podem causar fragmentação da rede em componentes desconectados, caso sejam rompidos, como por exemplo no caso de conectores entre clusters. Esse tipo de conector é assim chamado de cut-edge. Já conectores com baixo valor de edgebetweenness, quando eliminados da rede, podem ser substituídos por vias alternativas, como por exemplo no caso de conectores dentro de clusters. Assim, conectores interclusters tendem a ser mais vulneráveis quando comparados aos conectores intraclusters em uma determinada rede.

7. Conceitos-chave

Assortatividade: tendência de nós interagirem com nós similares a ele mesmo.

Betweenness: parâmetro que estima a relação entre dois nós, ou seja, leva em consideração a quantidade de caminhos mais curtos que passa entre eles.

Biologia de sistemas: área da bioinformática que estuda sistemas moleculares complexos e como as moléculas interagem entre si.

Caminho: seqüência consecutiva de nós em um grafo sem repetições, estando cada nó adjacente interligado por um conector.

Caminho geodésico: definido pela via mais curta dentro de uma rede entre dois nós quaisquer.

Circuito: sequência de nós sem repetição com um conector entre cada par de nós adjacentes na sequência, onde o nó inicial coincide com o nó final.

Clique: é definido como um grafo com alta conectividade entre seus elementos integrantes. Sendo assim, clique também é considerado um sinônimo de cluster.

Closeness: valor que indica os caminhos mais curtos entre um nó n e todos os outros nós da rede, uma tendência de aproximação ou isolamento de um nó.

Complexo protéico: grupo de proteínas formado pela associação de duas ou mais cadeias polipeptídicas.

Comprimento do caminho: definido pelo número de conectores que definem o caminho, ou então, pelo número de nós da sequência menos um.

Conector Cut-edge: conector que quando rompido causa fragmentação da rede.

Date hubs: são hubs que se ligam a diferentes proteínas em diferentes módulos (inter-módulo), ou seja, diferente tempo e/ou espaço, consequentemente, apresentado um papel global na rede.

Desassortatividade: tendência de nós interagirem com nós diferentes dele mesmo.

Diâmetro: indica a distância entre os dois nós mais afastados entre si de uma rede. Sendo assim, definimos que uma rede possui um alto diâmetro quando a distância geral entre os nós é muito ampla. Quando a distância entre os nós é pequena, então o diâmetro é baixo.

Dimerização: corresponde à união de dois monômeros, formando um dímero. Ou seja, é a formação de uma molécula a partir de duas moléculas menores.

Dimerizadores: compostos que induzem a dimerização, neste caso a interação proteica.

Distribuição de Poisson: distribuiçao aplicada a probabilidade de ocorrência de um evento em determinado intervalo de tempo.

Edgebetweenness: parâmetro que indica o número de caminhos mais curtos entre pares de nós que percorrem um determinado conector.

Edgetic: perturbação causada em um conector específico, portanto em uma interação específica na rede.

Forças intermoleculares: força que mantém as moléculas unidas durante a interação.

Gargalo (bottleneck): proteína que apresenta alto grau de betweenness.

Grau de nó (node degree): parâmetro referente à quantidade de nós adjacentes (diretamente conectados) a outro determinado nó.

Hipergrafo: rede caracterizada pela presença de hipervertices.

Hipervertices: Conectores que interligam nós que apresentam propriedades distintas nos hipergrafos.

Hot spot proteico: locais essenciais da interface com alta afinidade de ligação.

Inibição alostérica de uma proteína: na inibição alostérica, pequenos compostos ligam-se a sítios diferentes, causando mudança conformacional suficiente para interferir na ligação da proteína ligante.

Inibição ortostérica de uma proteína: inibição causada pela ligação direta de uma pequena molécula à superfície de interação da proteína ligante, interferindo diretamente nos hot spots críticos da interface e competindo com a proteína original.

Interface proteica: área através da qual as macromoléculas se comunicam e exercem sua funcionalidade.

Modularidade (clusterização): padrões de conectividade, onde seus elementos constituintes estão agrupados em subconjuntos altamente conectados.

Motivo proteico: associação de estruturas secundárias da proteína formando estruturas terciárias

Multiconector, interações: quando há dois ou mais conectores ligando os mesmos nós na rede em redes direcionadas

Multidígrafo: rede direcionada com a presença de multiconectores.

“Mundo pequeno”, efeito: define que existe um caminho mínimo entre um nó de origem e um nó de destino.

Ontologia gênica: tipo de análise que tem como função, em uma rede de interação proteína-proteína, agrupar proteínas que façam parte de um

mesmo processo biológico

Party hubs: proteínas altamente ligadas dentro do seu próprio módulo (intra-módulo), ou seja, ligação no mesmo tempo e/ou espaço.

Pleiotrópico, efeito: proteínas pleiotrópicas são aquelas que apresentam múltiplos efeitos em um sistemas biológico.

Rede: representação gráfica da interação entre nós por meio de vértices.

Rede bipartida: existe uma partição da rede, por exemplo, partição A e partição B, sendo os nós presentes na partição A adjacentes apenas a nós da partição B, e vice-versa.

Rede direcionada: apresentam conectores que direcionam o fluxo da informação em uma direção.

Rede não direcionada: os conectores desta rede não apresentam uma direção orientada.

Rede ponderada: são redes que se caracterizam pela presença de atributos associados a conectores e nós.

Resiliência: capacidade de uma rede a tolerar a deleção de seus nós por falha ou ataque.

Taxa evolutiva: medida das mudanças ocorridas numa entidade (gene, proteína, organismo, população) evolutiva ao longo do tempo.

Teoria da Percolação: tem por objetivo investigar o comportamento das propriedades de conectividade de uma rede.

Topologia de redes: estrutura e disposição de conexões entre os nós.

Vulnerabilidade do conector: grau de importância do conector.

8. Leitura recomendada

BARABÁSI, Albert-László; OLTVAI, Zoltán N. Network biology: understanding the cell's functional organization. Nature reviews Genetics. 5, 101-113, 2004.

GURSOY, Attila; KESKIN, Ozlem; NUSSINOV, Ruth. Topological Properties of Protein Interaction Networks from a Structural Perspective. Biochemical Society transactions. 36, 1398-1403, 2008.

LEVY, Emmanuel D.; PEREIRA-LEAL, Jose B. Evolution and Dynamics of Protein Interactions and Networks. Current opinion in structural biology. 18, 1–9, 2008.

MASON, Oliver; VERWOERD, Mark. Graph theory and networks in Biology. IET Systems Biology. 1, 89-119, 2007.

NEWMAN, Mark E.J. The structure and function of complex networks. SIAM Review. 45, 167-256, 2003.

YU, Haiyuan; KIM, Philip M.; SPRECHER, Emmett; TRIFONOV, Valery ; GERSTEIN, Mark. The Importance of Bottlenecks in Protein Networks: Correlation with Gene Essentiality and Expression Dynamics. PLoS computational biology. 3,e59, 2007.

WAGNER, Günter P.; PAVLICEV, Mihaela; CHEVERUD, James M. The road to modularity. Nature reviews Genetics. 12, 921-931, 2007.