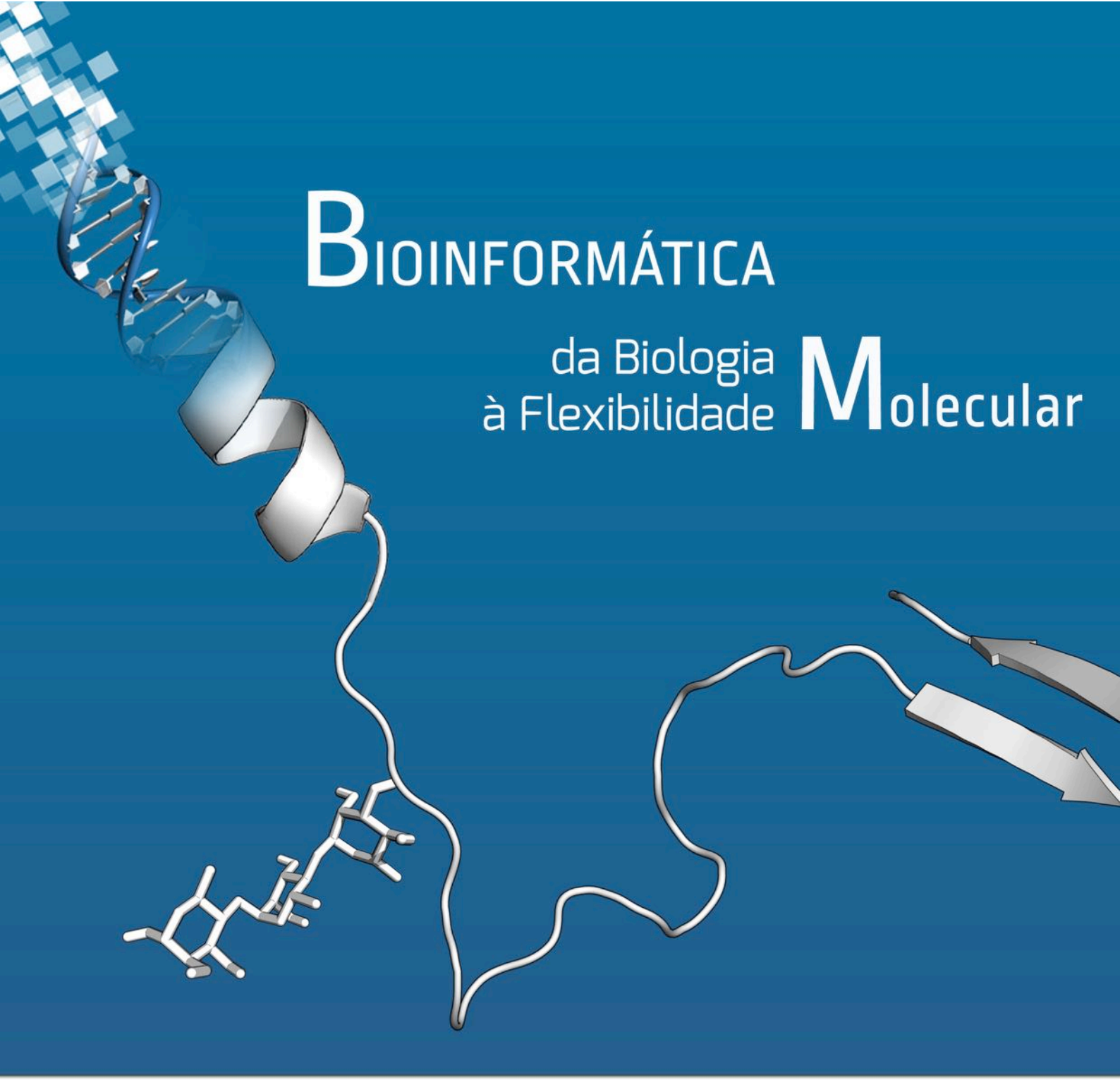


Bioinformática
da Biologia à Flexibilidade Molecular

Hugo Verli (Org.)

1ª edição
São Paulo, 2014

ISBN 978-85-69288-00-8
9 788569 288008

Sociedade Brasileira de Bioquímica e Biologia Molecular — SBBq

Apoio:
CNPq
CAPES
FAPERGS
UFRGS
UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

Hugo Verli Organizador

Bioinformática: da Biologia à Flexibilidade Molecular

1ª Edição

São Paulo

Sociedade Brasileira de Bioquímica e Biologia Molecular - SBBq

2014

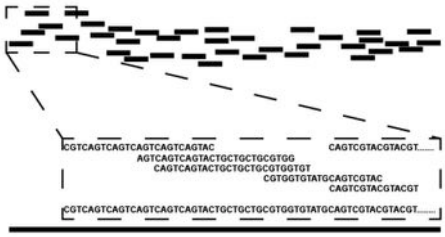
Ficha catalográfica elaborada por Rosalia Pomar Camargo CRB 856/10

B615 Bioinformática da Biologia à flexibilidade molecular / organização de Hugo Verli. - 1. ed. - São Paulo : SBBq, 2014. 282 p. : il.

1. Bioinformática 2. Biologia Molecular

CDU 575.112
ISBN 978-85-69288-00-8

4. Projetos Genoma



Representação da montagem de genomas.

Charley Christian Staats
Guilherme Loss de Morais
Rogério Margis

4.1. Introdução

A análise *in silico* das sequências nucleotídicas de cromossomo(s) de um dado organismo, ou simplesmente genoma, constitui uma da mais importantes aplicações da bioinformática. Tem como objetivo desenvolver e utilizar ferramentas para identificar e caracterizar genes, elementos genéticos móveis e outros elementos presentes em um determinado genoma, assim como fazer intercorrelações entre diferentes genomas com o intuito de buscar aspectos evolutivos comuns.

O primeiro organismo a ter a sequência de nucleotídeos de seu genoma determinado foi a bactéria Gram negativa *Haemophilus influenzae*, em um projeto liderado por J. Craig Venter. Desde 1995, ano de publicação desta análise genômica, as sequências de milhares de genomas de outros organismos já foram determinadas e analisadas, não apenas de espécies, mas também de variedades de espécies, raças e linhagens, entre outros.

Com a grande disseminação de estratégias de sequenciamento cada vez menos onerosas, muito tem se investido na geração de algoritmos e programas para analisar as sequências genômicas geradas. Previamente às análises do genoma de *H. influenzae*, programas para montagem de genomas já existiam, tendo sido desenvolvidos para análise de volumes de sequências relativamente pequenos, como os dos fagos λ e CMV, com tamanhos de aproximadamente 48.000 pares de bases (pb) e 229.000 pb, respectivamente. Para genomas maiores, novos programas tiveram que ser desenvolvidos em virtude da maior complexidade e quantidade das sequências analisadas. Neste capítulo, serão abordados os conceitos básicos e as principais ferramentas para montagem e anotação de genomas, assim como alguns programas para a sua análise.

4.2. Montagem de genomas

Nos primeiros anos da era genômica, o sequenciamento de genomas era baseado na metodologia de Sanger, ou método dideóxi. Para obtenção da sequência dos genomas, os fragmentos de DNA gerados após fragmentação química, física ou enzimática eram subclonados em vetores plasmidiais. Esta estratégia, denominada sequenciamento *shotgun*, é baseada na fragmentação aleatória dos cromossomos em fragmentos de DNA com tamanho relativamente pequeno. Estes fragmentos, cujo tamanho geralmente variava de 2.000 a 5.000 pb, eram submetidos ao sequenciamento. As sequências obtidas a partir de cada clone (chamadas de *reads*), com tamanho médio de 600 a 800 pb, eram submetidos a um processamento para retirada de sequências de baixa qualidade e, então,

utilizadas na montagem de *contigs* e genomas (ver abaixo).

Com o advento das metodologias denominadas *next-generation sequencing* – NGS (pirossequenciamento, Illumina, SOLiD, dentre outros), também ocorre fragmentação aleatória do DNA genômico, mas geralmente não são necessários os passos de clonagem. Comparativamente, estes novos métodos permitem a obtenção de *reads* de maneira muito mais rápida. Entretanto, o tamanho dos *reads* é menor, variando de algumas dezenas a poucas centenas de pares de base, dependendo da metodologia. Assim como no sequenciamento por Sanger, os *reads* obtidos passam por um controle de qualidade e então podem ser utilizados na montagem de genomas.

Independente da metodologia de sequenciamento utilizada, como resultado se tem uma grande lista de sequências nucleotídicas - os *reads* - de tamanhos que podem variar de 50 a 800 pb. Para montagem das sequências genômicas a partir destes *reads*, diferentes estratégias são utilizadas, dependendo da metodologia empregada. Para o sequenciamento convencional (Sanger), cada um destes reads é alinhado entre si na procura de regiões de identidade ou de sobreposição, de maneira a construir fragmentos contíguos (*contigs*), os quais podem ser definidos como a união de duas ou mais sequências (*reads*) formadas por sobreposição de elementos comuns a pelo menos duas sequências (Figura 1-4).

Os primeiros algoritmos para montagem de genomas se baseavam no alinhamento dos *reads* e na concatenação de sequências obtidas dos *reads* com os maiores alinhamentos. O processo se dava de forma cíclica, concatenando as sequências com o maior alinhamento até que todos estes alinhamentos fossem utilizados. Esta montagem de genomas a partir de *reads* tem como base os seguintes passos:

i) cálculo de alinhamentos aos pares de todos os fragmentos;

ii) escolha de dois fragmentos com a maior sobreposição;

iii) fusão dos dois fragmentos;

iv) repetição dos passos anteriores até obtenção de uma única sequência.

Para as novas metodologias de sequenciamento, devido ao tamanho relativamente menor dos fragmentos, algoritmos diferentes foram desenvolvidos. Os

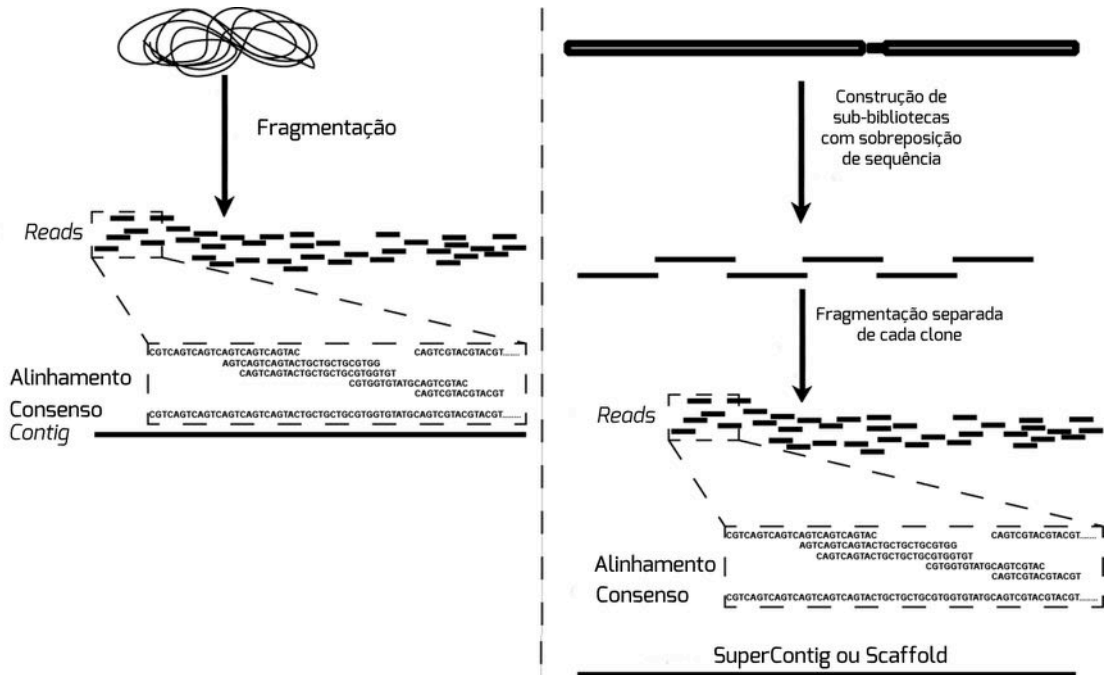


Figura 1-4: Montagem de genomas utilizando a estratégia de sequenciamento de genomas por *shotgun*. O painel à esquerda ilustra um esquema utilizado para genomas de menor tamanho e reduzido conteúdo de sequências repetitivas. O painel à direita ilustra uma estratégia mais complexa, usado para organismos com genoma maior.

programas de montagem atuais utilizam grafos de sobreposição ou grafos de Bruijn. Estes grafos identificam *reads* com possibilidade de compartilharem trechos de sobreposição entre si utilizando uma estratégia baseada no alinhamento em sementes.

Com esta abordagem, pequenos fragmentos de comprimento fixo obtido de cada *read*, os k-*mers*, são usados como um índice, e apenas pares de leituras que partilham uma semente são posteriormente avaliados. Os grafos de Bruiijn baseiam-se na decomposição de reads em k-*mers* (por exemplo dodecâmeros, ou seja fragmentos de 12 nucleotídeos), os quais são utilizados como nodos destes grafos. Uma ligação direta entre os nodos indica que estes k-*mers* ocorrem consecutivamente em um ou mais *reads*.

Uma série de programas foram desenvolvidos para a montagem de genomas, utilizando diferentes algoritmos (Tabela 1-4). No caso de sequenciamento de genomas procarióticos, ao final do processo é esperada a obtenção de uma sequência única, a qual representa toda a sequência nucleotídica do cromossomo. Sabe-se, todavia, que plasmídeos podem ser encontrados em diversos micro-organismos. Assim o número de *contigs* será dependente do número de plasmídeos e, em casos menos frequentes, do número de cromossomos presentes naquela bactéria.

Ao ser analisado o genoma de organismos eucariotos, nos quais se encontra uma grande variação no número de cromossomos, um número maior de *contigs* é esperado. Teoricamente, cada cromossomo deveria ser representado por um *contig*. Entretanto, nos passos iniciais de montagem de genomas são observados dezenas a centenas de *contigs*, dependendo da complexidade do organismo cujo genoma esta sendo sequenciado. Os genomas de eucariotos, em especial de eucariotos superiores, possuem pelo menos duas características que tornam o processo de montagem mais complexo:

i) uma quantidade considerável de sequências repetitivas que dificulta o processo de montagem devido a alinhamentos de alto escore com diversas sequências;

ii) o seu tamanho, podendo chegar a mais de 3 bilhões de pares de base (caso do genoma humano).

Tabela 1-4: Principais programas utilizados na montagem de genomas e transcriptomas.

Nome	Análise
ABySS	grandes genomas
ALLPATHS-LG	grandes genomas
Celera WGS *Assembler*	grandes genomas
CLC *Genomics Workbench*	genomas e trancriptomas
Geneious	genomas
Newbler	genomas e transcriptomas
Phrap	genomas e transcriptomas
SOAPdenovo	genomas e transcriptomas
Staden gap4 *package*	genomas pequenos e transcriptomas
Trans-ABySS	transcriptomas
Velvet	genomas pequenos e transcriptomas

Para sobrepujar estas dificuldades, passos intermediários se tornam necessários, como a construção de sub-bibliotecas genômicas. Cada uma destas sub-bibliotecas é sequenciada, de forma a gerar *contigs*. O conjunto de diferentes *contigs* oriundos de diferentes sub-bibliotecas será utilizado para a geração de *scaffolds* (Figura 1-4). Geralmente, são necessários passos adicionais de clonagens de regiões específicas do genoma e posterior sequenciamento destas para o "fechamento" do genoma.

Um dos maiores desafios, entretanto, para o sequenciamento de genomas reside na adequada montagem de regiões repetitivas. No genoma humano, por exemplo, existem pelo menos seis classes de sequências repetitivas:

i) minissatélites, microssatélites ou satélites;

ii) SINEs (elementos nucleares pequenos intercalados);

iii) LINEs (elementos nucleares longos intercalados);

iv) transposons;

v) retrotransposons;
vi) *clusters* de genes DNAr (genes responsáveis pela síntese dos RNA ribossômicos – RNAr).

Estas diferentes classes, cujos tamanhos podem variar de centenas de pares de base, caso de microssatélites e SINEs, a dezenas de milhares de pares de base, observado em *clusters* de genes DNAr, podem constituir mais de 50 % do tamanho de cada cromossomo humano.

O grande desafio na montagem de sequências genômicas com alto conteúdo de elementos repetitivos se refere a correta quantificação e localização destes elementos nos cromossomos. Desta forma, o desafio central da montagem de genomas reside na resolução destas sequências repetitivas, estando este desafio diretamente associado à metodologia de sequenciamento utilizada. Por exemplo, se forem obtidos *reads* de tamanho menor que uma unidade de repetição, todos estes *reads* serão utilizados para formar um *contig* que contém apenas a sequência de repetição. Entretanto, ao serem obtidos *reads* com tamanho maior que a unidade de repetição, os mesmos podem ser utilizados na resolução da localização destas sequências repetitivas em um determinado cromossomo.

Alguns programas permitem montar genomas complexos com repetições baseados em *reads* maiores (como os obtidos pela metodologia de Sanger ou pirosequenciamento). Para tal, estes programas realizam a montagem em duas ou mais fases distintas, nas quais as sequências repetitivas são processadas separadamente. Em um primeira fase do processo de montagem, *reads* contendo sobreposição de sequências não ambíguas são agrupados em *contigs*, cujas extremidades contém as regiões limítrofes das sequências de repetição. A segunda fase se caracteriza pela montagem de *contigs* não ambíguos em sequências maiores, usando dados de *reads mate-pair*.

Dados de sequenciamento *paired-end* oferecem a possibilidade da determinação exata de sequências que flanqueiam uma determinada sequência de repetição. Em experimentos tradicionais associados ao sequenciamento de Sanger, um protocolo *paired-end* inicia-se com longos fragmentos de DNA clonados em vetores para sua replicação em *Escherichia coli*. As extremidades destes fragmentos poderiam assim ser facilmente determinadas por sequenciamento. Protocolos *paired-end* para as estratégias de sequenciamento atuais não requerem passos de clonagem em *E. coli*. Entretanto, os mesmos se baseiam na circularização do fragmento de DNA do tamanho desejado, sendo as extremidades posteriormente reconhecidas devido à etiqueta (*tag*) utilizada para propiciar a circularização por meio da ligação. Com a determinação das sequências flanqueadoras de uma repetição, há maior chance de conseguir determinar a sua localização em um genoma.

A qualidade de montagem do genoma pode ser acompanhado por alguns índices. A cobertura reflete a quantidade de *reads* associados a um determinado fragmento de DNA. Por exemplo, uma cobertura de 10X indica que, para o genoma sendo avaliado, cada nucleotídeo foi encontrado em pelo menos 10 *reads*.

Outro valor importante refere-se ao N50. Trata-se de uma medida estatística muito utilizada para avaliar a qualidade da montagem, visto que revela o quanto de um genoma é coberto por *contigs* grandes. Um valor de N50 igual a n significa que 50% dos reads estão montados em um *contig* de tamanho n ou maior. Por exemplo, na montagem do genoma de cão doméstico, depositado no NCBI sob o número de acesso AAEX03, o sequenciamento dos 40 cromossomos, com uma sequência total de 2.410.976.875 bases gerou 27.106 *contigs* com um N50 de 267.678. Isto significa que mais de 50% dos *reads* estão associados a *contigs* de 267.678 bases ou maiores.

4.3. Montagem de transcriptomas

Em análises de novos genomas, um ponto importante se refere à identificação de transcritos. Além de fornecer indícios sobre quais genes estão sendo expressos em uma determinada situação fisiológica a qual as células ou tecidos estão sendo expostos, o sequenciamento de transcritos tem uma aplicação importante na procura de sequências codificantes em genomas. Esta estratégia tem uma aplicabilidade muito grande em organismos em que o conteúdo de íntrons por gene é grande, como em eucariotos mais complexos.

Ao contrário de genomas, em transcriptomas o material de partida geralmente é

cDNA, obtido a partir de transcrição reversa de RNA. A grande maioria dos trabalhos se dá em torno de RNAm mas, cada vez mais, RNAs não codificantes, com possível papel regulatório, estão sendo avaliados por esta metodologia (ver abaixo). O *pool* de cDNAs pode então ser subclonado e ser submetido ao sequenciamento pela metodologia de Sanger ou diretamente fragmentado e ser submetido ao sequenciamento NGS. Uma grande lista de *reads* é então obtida, os quais podem ser utilizados para realizar a montagem do transcriptoma *de novo* ou ser ancorados a sequência de um genoma para ajudar na identificação de sequências codificantes e de extremidades éxon/íntron.

No caso da montagem *de novo*, os *reads* são alinhados e aqueles que apresentam alinhamento positivo são fusionados, dando origem a *contigs*. Entretanto, diferentemente da análise de genomas, muitos *contigs* são gerados, cada um possivelmente representando um mRNA maduro.

Adicionalmente, alguns programas podem, além de realizar a montagem de transcriptomas ou alinhamento a genomas, fazer uma análise da representatividade de cada transcrito dentro do conjunto total de RNA analisado, por meio do cálculo da frequência relativa de cada transcrito identificado. Com estes cálculos é possível realizar análises de expressão diferencial de genes. Dentre os pacotes de programas utilizados, podem ser citados Cufflinks-Cuffdiff, DegSeq, DESeq, EdgeR, entre outros.

A análise desta expressão relativa de transcritos pode ser realizada com base em duas estratégias principais:

i) mapeamento a uma sequência genômica previamente conhecida;

ii) análise *de novo*, independente da sequência genômica e baseada na montagem dos transcritos diretamente a partir dos *reads*.

Na primeira estratégia, os *reads* são mapeados ao genoma, ou seja, as regiões de identidade nucleotídica são ancoradas à sequência genômica, sendo identificadas por metodologias de sequenciamento que levam em consideração o número de *reads* mapeados em relação à porção do genoma que contém um gene. Alguns dos programas para este tipo de mapeamento incluem Bowtie, Tophat e SOAP, dentre outros. Como resultado, uma determinada sequência do genoma é representada por um grande número de *reads*, no caso de genes mais expressos, ou um baixo número de *reads*, no caso de genes menos expressos.

Deve ser levado em consideração, entretanto, que quanto maior o tamanho do gene mais se espera encontrar *reads* associados a este gene. Desta forma, a maneira mais comum para se calcular a expressão relativa de um determinado gene é o RPKM (*reads per kilobase of transcript per million mapped reads* – *reads* por kilobase de transcrito por milhões de *reads* mapeados). Esta abordagem permite uma análise comparativa baseada em uma série de análises estatísticas para comparação de transcritos com diferentes RPKMs de diferentes amostras biológicas ou diferentes tempos de tratamento, por exemplo.

Quando são considerados organismos cujo genoma ainda não foi determinado, uma construção do transcriptoma a partir de dados de RNAseq é realizada (*de novo*). A partir das sequências dos transcritos gerados, é possível então fazer o cálculo do RPKM de cada transcrito identificado.

4.4. Identificação/anotação gênica

A anotação de genomas é o passo seguinte à montagem dos genomas. Trata-se de um conjunto de protocolos e fluxos de trabalho utilizados para delimitar, em uma determinada sequência genômica, possíveis genes e predizer a sua função com base na similaridade com sequências conservadas. Basicamente, existem dois grande grupos de genes avaliados nestas metodologias. O primeiro grupo se refere àqueles cujo produto é reconhecido pelos ribossomos e dará origem a uma proteína (ou seja, RNAm). Já o segundo engloba os genes cujo produto terá funções estruturais e funcionais dependentes da própria molécula de RNA, como RNAt e RNAr. Diferentes abordagens são utilizadas para identificar as sequências de cada um destes grupos de genes, como será visto abaixo.

Identificação de regiões codificantes

O mecanismo de delimitação da sequência gênica é drasticamente influenciado pelo Domínio ao qual pertence o organismo cuja sequência genômica foi determinada. Isto se deve ao fato de que existe uma grande diferença nas estruturas de genes procarióticos e eucarióticos.

Genes procarióticos codificantes de proteínas são colineares com seus produtos gênicos. Esta característica permite inferir que toda região delimitada por um códon de inicío e um códon de término, região esta denominada de ORF (*Open Reading Frame*), potencialmente constitui uma região codificante de uma proteína em um genoma procariótico.

Por sua vez, genes eucarióticos codificantes de proteínas são mais complexos, geralmente sendo caracterizados pela presença de sequências intervenientes ou íntrons. Até pouco tempo, acreditava-se que íntrons constituíam um produto da evolução que povoou as sequências gênicas com o chamado "DNA lixo", de modo que uma mutação que eventualmente viesse a acontecer tivesse maior possibilidade de ocorrer em regiões do gene que não têm capacidade codificante. Recentemente, contudo, determinou-se que os íntrons exercem um importante papel regulatório na expressão gênica.

Íntrons são elementos gênicos que, durante o processo de expressão gênica, são excisados durante o processamento do RNA, em um grande complexo de reações denominado *splicing*. Os íntrons podem variar em número e tamanho, dependendo da complexidade do organismo. Assim, em organismos mais simples, como leveduras e fungos filamentosos, o número de íntrons por gene é pequeno (geralmente de 1 a 4 por gene), assim como o seu tamanho (geralmente girando em torno de 50 pb).

Ao contrário, em organismos mais complexos como humanos e plantas, tanto o número de íntrons por gene quanto o seu tamanho aumentam significativamente, de forma que grande parte do gene é constituído por íntrons (mais de 90%, dependendo do organismo). Um comparativo entre as estruturas básicas de genes codificantes de proteínas procarióticos e eucarióticos, assim como os seus respectivos processos de expressão, é apresentado na Figura 2-4.

Associado ao grande número de íntrons, genes de organismos eucarióticos mais complexos geralmente são caracterizados pelo

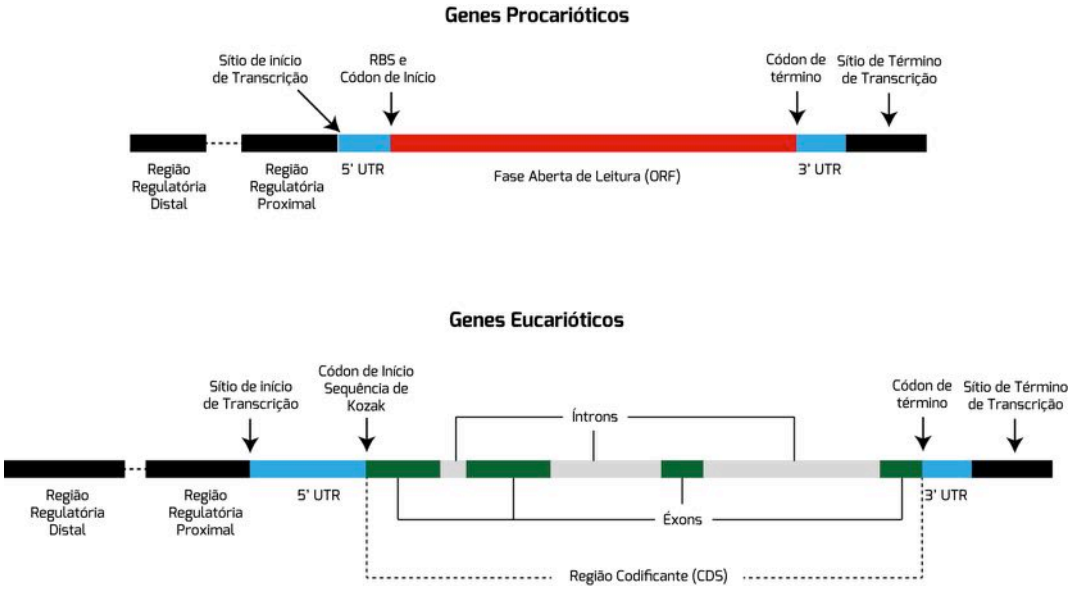


Figura 2-4: Esquema representando os elementos encontrados em genes procarióticos (quadro superior) e eucarióticos (quadro inferior). Os genes estão representados no sentido 5'-3' e podem ser notadas as principais diferenças entre estas classes de genes, como a presença de íntrons e regiões regulatórias mais complexas em eucariotos.

splicing alternativo. Este processo é caracterizado pela incorporação diferencial de íntrons e éxons no RNAm maduro, de forma a produzir diferentes proteínas a partir do mesmo gene.

Diferentes estratégias para procura de genes em genomas foram desenvolvidas considerando estas características diferenciais na estrutura de genes procarióticos e eucarióticos. A procura de ORFs em genomas procarióticos constitui uma estratégia simples e direta. Entretanto, é uma estratégia sujeita a uma diversidade de erros.

Nestas predições, não são considerados elementos canônicos clássicos presentes na estrutura de genes (isto é, sequências conservadas para ligação do fator sigma, região de ligação do ribossomo, sítio de início de tradução e sítio de término de tradução) e operons, os quais poderiam auxiliar na procura *ab initio* (ou seja, diretamente a partir de sequência, sem informações experimentais diretas sobre o produto gênico) de genes em genomas procarióticos. Assim, a procura de genes baseada apenas na identificação de ORFs geralmente leva a um número grande de resultados falsos positivos e falsos negativos (Figura 3-4).

Para sobrepujar estas limitações, mecanismos de delimitação das sequências gênicas em genomas procarióticos foram então desenvolvidos e se baseiam em algoritmos característicos para detectar, na sequência de DNA, dois tipos fundamentais de informações: sinais e conteúdo. Estes mecanismos foram então expandidos para procura de genes em organismos eucarióticos.

Os detectores de sinais procuram por caracteres funcionais específicos de genes, tanto associados à transcrição quanto à tradução. Sinais transcricionais incluem sequências canônicas conservadas que delimitam as regiões necessárias para que se inicie o processo de transcrição. Os sinais mais comumente descritos em procariotos são as regiões -35 e -10 e as sequências de associação com a RNA Polimerase. Já os sinais procurados em sequências eucarióticas geralmente constituem a região TATA box, assim como o sítio de clivagem e poliadenilação, que caracteriza o terminador.

Os sinais traducionais, por sua vez, se referem basicamente às regiões importantes para recrutamento de ribossomos, como o RBS (*ribosome binding site*, ou sitio de ligação a ribossomos) em procariotos. Como este mecanismo é diferente em organismos eucarióticos, uma região conservada, denominada sequência de Kozak, é utilizada como sinal traducional em eucariotos. Estas duas regiões se localizam imediatamente a montante (*upstream*) aos respectivos códons de início, e desempenham um papel importante nos mecanismos de delimitação de genes.

Adicionalmente, a detecção de sinais que delimitam os íntrons também são utilizados pois, como abordado anteriormente, os genes de eucariotos são amplamente povoados por íntrons. Desta forma, a correta predição da posição de íntrons é fundamental para correta anotação do gene, sendo que os principais sinais a serem avaliados são os nu-

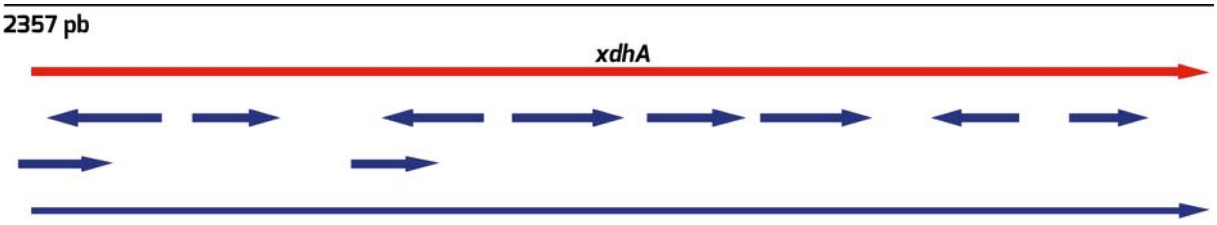


Figura 3-4: A simples procura de ORFs pode gerar resultados falso positivos na procura de genes em organismos procarióticos. Como exemplo, uma sequência de DNA de 2357 pb da bactéria *E. coli* HS (nucleotídeos 3027764 ao 3030120 – Código de Acesso junto ao NCBI NC_009800.1), o qual contém o gene xdhA, foi avaliada quanto à presença de ORFs com mais de 150 pb com o programa ORF Finder. A sequência anotada do gene encontra-se em vermelho, ao passo que as possíveis ORFs estão demarcadas em azul.

cleotídeos que compõem as extremidades conservadas 5' e 3' do íntron, mais comumente GT e AG (ver abaixo).

Já os detectores de conteúdo classificam a sequência de DNA em codificante e não-codificante. Como região não-codificante entendem-se íntrons, regiões intergênicas e regiões não traduzidas dos genes. Os detectores de conteúdo podem ainda ser subdivididos em detectores extrínsecos e detectores intrínsecos. Os detectores de conteúdo extrínsecos se baseiam no fato de que regiões codificantes são mais conservadas em relação às não-codificantes propiciando, desta forma, a identificação de éxons conservados com base em procuras por homologia.

O mecanismo básico desta busca é através do programa BLAST (ver capítulo 3). Contudo, uma limitação nesta metodologia se refere à avaliação adequada da presença de ortólogos diretos. Desta forma, a distância filogenética (isto é, evolutiva, ver capítulo 5) entre o organismo cujo genoma está sendo analisado e aqueles organismos cujas sequências estão depositadas nos bancos de dados pode influenciar diretamente no resultado.

Detectores de conteúdo intrínseco, por sua vez, tem como foco principal algumas características inatas do DNA, as quais permitem a predição do potencial de uma sequência codificar ou não uma proteína. Como exemplos de características avaliadas em detectores intrínsecos podem ser citados:

i) em muitos organismos há uma preferência das bases G ou C em relação às bases A ou T na terceira posição do códon;

ii) a utilização diferencial de códons sinônimos, ou seja, diferentes códons que codificam para o mesmo aminoácido;

iii) frequência de distintas sequências nucleotídicas hexaméricas;

iv) a periodicidade de ocorrência de bases, dentre outros.

Estes caracteres são utilizados, por exemplo, em modelos de Markov para a construção de modelos capazes de reconhecer sequências codificantes. Com base nos mecanismos discutidos acima, dois principais sistemas para procura de genes em genomas de eucariotos foram construídos, denominados empírico e *ab initio*.

Procura empírica de genes

A predição empírica ou baseada em evidência leva em consideração buscas por similaridade com outros bancos de dados (genômicos, transcritômicos ou proteômicos) para identificar e delimitar as sequências gênicas. Métodos de identificação de genes baseados em similaridade são considerados de alta confiabilidade para localizar e construir modelos gênicos, desde que existam relatos prévios de estruturas gênicas do próprio organismo (como, por exemplo, sequências de RNAm) ou baseado em análises de conservação provenientes de alinhamentos de genomas de espécies filogeneticamente relacionadas.

Especialmente para o caso de organismos eucarióticos, alinhamentos de sequências oriundas de bancos de dados de proteínas ou de transcritos contra o genoma em anotação permitem aferir que, geralmente, os *gaps* constituem os íntrons. Esta premissa é frequentemente acompanhada pela observação de que as sequências limítrofes dos íntrons identificados constituem os dinucleotídeos consenso GT e AG, característicos sítios 5' e 3' dos íntrons. Estes alinhamentos geram forte evidência dos componentes das estruturas dos genes, muitas vezes definindo completamente a localização de cada éxon e cada íntron (Figura 4-4).

Procura ab initio *de genes*

A predição *ab initio*, por sua vez, depende tanto da informação de detectores de sinais quanto de conteúdo para delimitar a sequência gênica. Para tal, os algoritmos que se valem desta estratégia utilizam redes neurais, transformadas de Fourier e, mais comumente, modelos de Markov. Para realizar estas detecções, os algoritmos são treinados

com sequências conhecidas do genoma em questão. Por exemplo, a Figura 5-4 ilustra o grau de conservação dos nucleotídeos presentes na sequência de Kozak de *Drosophila melanogaster*, perfil este que pode ser utilizado na predição de novas sequências codificantes neste organismo. Outro exemplo pode ser observado no grau de conservação das regiões 5' e 3' provenientes de íntrons de genes humanos (Figura 6-4).

Dentre as limitações da predição *ab initio* está o fato de que, usualmente, o resultado obtido se refere às regiões codificantes, sem informações sobre regiões não traduzidas ou transcritos provenientes de *splicing* alternativo.

Assim, para sobrepujar estas limitações a combinação das duas estratégias parece ser a mais eficaz nos fluxos de trabalho utilizados para predição de genes em genomas sequenciados. Para tanto, alguns destes algoritmos são treinados com modelos gênicos já conhecidos, de organismos filogeneticamente próximos e, assim, provavelmente possuem uma estrutura gênica muito parecida com a do organismo que está em análise.

Anotação de regiões codificantes

O passo seguinte à identificação de sequências que possivelmente constituem genes é a sua anotação. A anotação manual foi bastante utilizada na análise dos primeiros genomas. Entretanto, devido à complexidade e ao alto número de sequências genômicas disponibilizadas a cada dia, há um consenso de que a anotação automática está se tornando indispensável.

A forma mais simples de anotação automática se dá pela análise de uma série de diferentes mecanismos de predição e delimitação de sequências gênicas e, então, utilização de um algoritmo de seleção, também denominado de *combiner*. Este algoritmo tem a função de selecionar a predição que melhor represente os modelos gênicos frente os algoritmos utilizados. Para tanto, os *combiners* estimam os tipos e as frequências de erros oriundos de cada programa de predição, escolhendo posteriormente as combinações de evidências que minimizam tais erros. Após as predições *ab initio* e baseados em evidência, alguns dos *combiners* devem ser treinados com sequências não previamente utilizadas nos programas de predições de genes.

Os *combiners* mais atuais utilizam técnicas que combinam evidências não estocásticas ponderadas (*nonstochastic weighted evidence*) que computam tanto o tipo quanto a abundância de uma evidência para o cálculo da sequência gênica consenso. Uma lista dos algoritmos mais utilizados para confecção de fluxos de trabalho para identificação de genes está disponível na Tabela 2-4.

A anotação da função de genes é um processo basicamente comparativo, sendo utilizados bancos de dados de proteínas, como o NCBI ou o UniProt (trEMBL + Swiss-Prot)

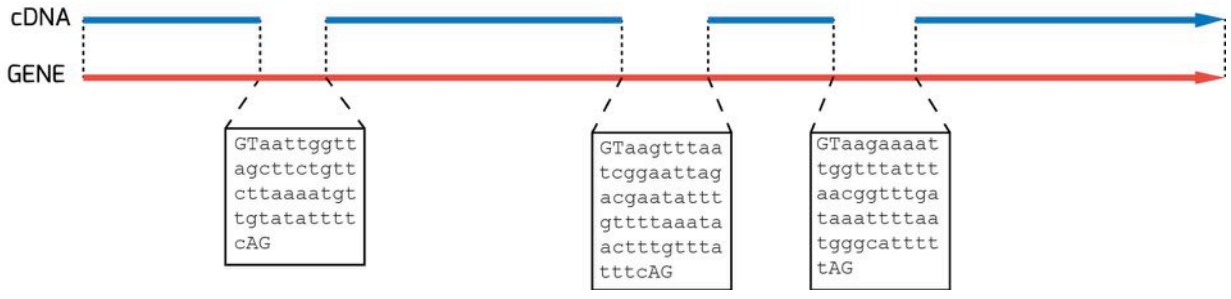


Figura 4-4: Identificação de genes baseada em evidência. Utilizando BLASTn com base em dados de transcritoma (cDNA, em azul), pode ser alcançada uma aproximação da sequência do gene (vermelho), inclusive permitindo a delimitação de éxons e íntrons. As regiões de identidade estão delimitadas por traços verticais. Com base na sequência de íntrons (quadros na porção inferior), é possível construir modelos para sua predição. Modelo construído com base no gene F10E9.5 de *Caenorhabditis elegans* (código de acesso NCBI NC_003281).

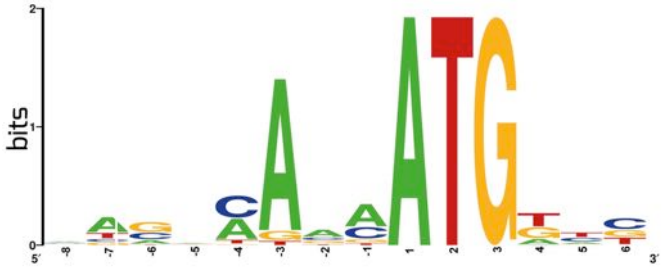


Figura 5-4: Padrão de conservação de nucleotídeos da sequência de Kozak, baseado no alinhamento de 30 sequências de cDNA obtidas de *D. melanogaster* e analisados junto ao servidor WebLogo. A medida de conservação é refletida pela altura da base. Os números abaixo representam o códon de início de tradução (1 a 3), o segundo códon do mRNA (4 a 6) e a região a montante (-8 a -1).

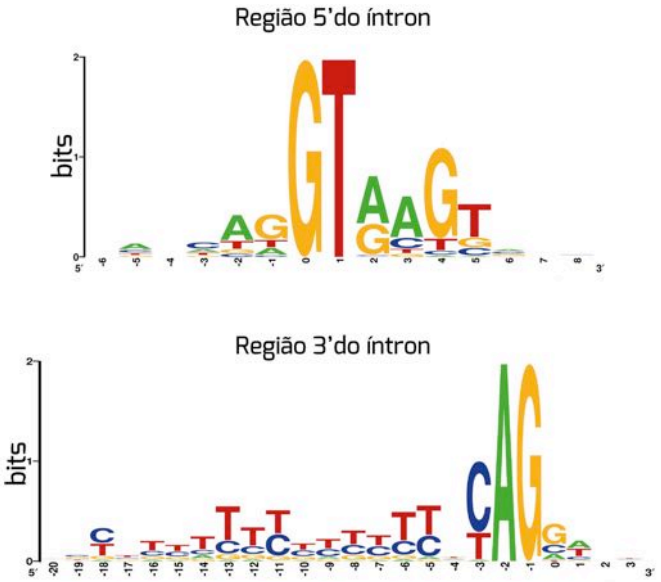


Figura 6-4: Padrão de conservação de nucleotídeos nas regiões 5' (painel superior) e 3' (painel inferior) de íntrons humanos. Resultado obtido pelo alinhamento de 100 sequências intrônicas e analisados junto ao servidor WebLogo. A medida de conservação é refletida pela altura da base. Os números abaixo de cada esquema indicam o início e o fim do íntron (0 e 1 no esquema superior; -2 e -1 no esquema inferior), assim como as regiões adjacentes.

ou de domínios proteicos (PFAM, NCBI CDD, Interpro). Uma das vantagens da utilização do Swiss-Prot como banco de dados para identificação dos produtos gênicos se refere ao fato deste ser um banco de dados manualmente curado, ou seja, inspecionado contra possíveis erros decorrentes da anotação automática. Com base nestas análises, quatro grupos distintos de anotações podem ser realizadas:

i) a existência de um ortólogo direto previamente caracterizado, revelado por BLAST, gerará a anotação com base no nome do ortólogo;

ii) a inexistência de um ortólogo direto, mas a presença de um domínio proteico conservado, revelado por análises em PFAM ou Interpro, gerará a anotação "*domain containing protein*" ou proteína contendo o domínio;

iii) a inexistência de ortólogos diretos previamente caracterizados ou domínios conservados confere as anotações proteína predita (*predicted protein*) ou proteína hipotética (*hypothetical protein*);

iv) quando um gene codificante de proteína hipotética possui ortólogos diretos, eles são denominados codificadores de proteína hipotética conservada (*conserved hypothetical protein*).

Outro passo na anotação da função de genes se refere à predição da localização da proteína codificada por este gene. Por exemplo, se uma proteína possui muitas regiões hidrofóbicas, compatíveis com sua inserção em membrana, possivelmente esta será uma proteína integral de membrana. Adicionalmente, proteínas secretadas ou endereçadas a alguma organela geralmente apresentam uma sequência sinal.

Diversas ferramentas estão disponíveis para localização de domínios transmembrana (TMHMM, TMPred, HMMTOp), baseando-se em métodos estatísticos para aferição da presença destes domínios. Métodos mais robustos para determinar a localização celular de um produto gênico foram desenvolvidos e se baseiam em uma diversidade de métodos estatísticos, geralmente treinados com sequências proteicas conhecidamente pertencentes a algum sub-compartimento celular (Tabela 3-4). De uma maneira geral, todas estas ferramentas são utilizadas na constru-

Tabela 2-4: Principais algoritmos utilizados na predição de genes e a sua funcionalidade.

Algoritmo	Descrição	Aplicação
Predições *ab initio* e baseados em evidência		
Augustus	Aceita evidências baseadas em transcriptomas e banco de dados de proteínas	Eucariotos
FGNESH	Arquivos para treino derivados de análise do fabricante	Eucariotos
fgenesB	Predição de genes e operons em bactérias baseadas em padrões e cadeias de Markov	Procariotos
Genemark	Arquitetura de busca baseada em *self-training*	Procariotos e eucariotos
Twinscan	Extensão do algoritmo Genscan que utiliza homologia entre dois genomas para guiar a predição de genes	Eucariotos
GenomeScan	Extensão do algoritmo Genscan que utiliza BLASTx para guiar a predição de genes	Eucariotos
Glimmer	Utiliza modelos de Markov interpolados	Procariotos
Combiners		
Evidence Modeler	Tem como resultado um modelo gênico pela combinação de evidências obtidas a partir de alinhamento de dados transcriptômicos e proteômicos com predições *ab initio*	Eucariotos
Evigan	Algoritmo de evidências probabilísticas que usa redes Bayesianas para pontuar e integrar predições *ab initio* e baseadas em evidência para produzir modelos gênicos.	Eucariotos

ção de fluxos de trabalho que integram diferentes ferramentas para analisar o resultado da predição de cada gene, conferindo uma anotação geral (Figura 7-4).

4.5. Identificação/anotação RNAnc

Considerando o dogma central da biologia molecular, no processo de síntese proteica (tradução) há a participação direta de pelo menos três classes distintas de RNAs:

i) o RNA mensageiro, que servirá de molde para síntese da proteína;

ii) o RNA ribossômico que, como indica o nome, é um componente estrutural e funcional dos ribossomos;

iii) o RNA transportador, que funciona como adaptador, carreando aminoácidos para serem incorporados na cadeia nascente da proteína durante o processo de tradução.

A anotação de genes de RNAs não codificantes - RNAnc (RNAt, RNAr, dentre outros) ainda não apresenta um grande número de programas quando comparada às estratégias disponíveis para anotação de genes codificantes de proteínas. Isto se deve, principalmente, à grande heterogeneidade e à pequena conservação dos RNAnc quando comparados a sequências de proteínas. Ao contrário de genes codificantes de proteínas, RNAnc geralmente não apresentam conservação de sequência 1^{ária}, dificultando a detecção destes genes.

Um dos mecanismos mais utilizados na busca de RNAt em genomas é o tRNAscan-SE. Este algoritmo se baseia em uma série de cálculos estatísticos que avaliam, entre outros parâmetros, o potencial local para formação das estruturas 2^{árias} típicas de tRNAs em forma de trevo, assim como a presença de bases invariantes que definem regiões conservadas presentes nos promotores destes genes. Outro mecanismo de busca de RNAts se refere ao algoritmo ARAGORN. A

Tabela 3-4: Principais algoritmos utilizados na predição da localização celular de proteínas.

Algoritmo	Descrição	Aplicação
BaCelLo	Com base na composição de aminoácidos e sequências de treino, prediz em 5 localizações (secretada, citoplasmática, nuclear, mitocondrial e cloroplástica)	Plantas, animais e fungos
LOCtree	Com base na sequência N-terminal, prediz a localização em secretada, citoplasmática, nuclear, mitocondrial, cloroplástica e organelar.	Eucariotos e procariotos
TARGETp	Com base na sequência N-terminal, prediz a localização como secretada, mitocondrial e cloroplástica, dentre outras.	Eucariotos e procariotos
Wolf PSORT	Com base na sequência N-terminal e regras empíricas, classifica o endereçamento em cloroplástico, citosólico, citosesqueleto, retículo endoplasmático, extracelular, golgi, lisossômico, mitocondrial, nuclear, peroxissomal, membrana plasmática e membrana vacuolar. Permite localização múltipla.	Animais, fungos e plantas
Cell-PLoc	Permite realizar a localização de proteínas em mais de 25 diferentes locais, baseados em treino com sequências cuja proteína tem localização conhecida.	Eucariotos, procariotos e vírus

estratégia deste programa para a procura de tRNAs em sequências nucleotídicas se baseia em algoritmos heurísticos para a predição da estrutura do tRNA baseada na homologia com sequências conservadas, assim como a potencialidade de formar estruturas 2^{árias} típicas do tRNA. Por fim, o tRNAfinder se baseia em cálculos para detecção da estrutura 2^{ária} do RNA predito para identificar genes de tRNA.

Já a predição de RNArs é baseada em conservação de sequências. Ao passo que organismos procarióticos possuem geralmente três moléculas de RNAr (23S, 16S e 5S) completamente maduras e funcionais, eucariotos possuem quatro (28S, 18S, 5.8S e 5S). Cada uma destas sequências apresenta grande grau de conservação com os ortólogos de diferentes organismos. Desta forma, ferramentas baseadas em Modelos Ocultos de Markov, como o RNAmmer, foram construídas para delineamento dos genes responsáveis pelos RNArs. Adicionalmente, um grande banco de dados com famílias de RNA foi construído, e a cada ano novas adições de sequências de RNAs são feitas ao RFam. Estas famílias podem ser classificadas em três grandes grupos:

i) RNAs não codificantes (RNAnc);

ii) elementos estruturais regulatórios em *cis*, característicos de alguns RNAm que desempenham função de regulação da expressão gênica principalmente por meio da formação de estruturas 2^{árias};

iii) RNAs que podem sofrer o processo de *auto-splicing*.

Cada uma destas famílias é representada por alinhamentos múltiplos, consensos de estruturas 2^{árias} e modelos de covariância. Por meio de comparação de sequências com os consensos obtidos para os modelos de cada família, é possível identificar genes responsáveis pelos rRNAs, tais como os snoRNAs, que são componentes do spliceossomo. Existe ainda, contudo, uma grande gama de outros RNAnc que não apresentam grau de conservação necessário para formar uma família.

Identificação de pequenos RNAs

O termo "pequeno RNA" é, conceitualmente, muito vago e acaba englobando diferentes classes destes, como microRNAs, siRNAs, TAS-siRNAs, tRFs, entre outras. Contudo, existem características dos pequenos RNAs que podem ser utilizadas para identifi-

car as classes distintas: não codificam proteínas (apesar de alguns serem originados de regiões codificadoras), possuem tamanho variando entre poucas dezenas de nucleotídeos, suas rotas de biogênese e seus papéis funcionais.

Os pequenos RNAs fazem parte de um grupo de pequenas moléculas, sendo conhecidos há décadas, e inicial e erroneamente creditados como produtos de degradação de RNA, não possuindo um papel biológico específico. Com a identificação do fenômeno de silenciamento gênico (RNAi) foi observado que pequenos RNAs poderiam, de fato, desempenhar um papel funcional, regulando a expressão gênica em vários níveis. Devido ao papel de forte regulador da expressão gênica, muita atenção tem sido dada aos pequenos RNAs, com um número crescente de trabalhos sendo feitos relacionando estes com patologias e controlando processos básicos do desenvolvimento.

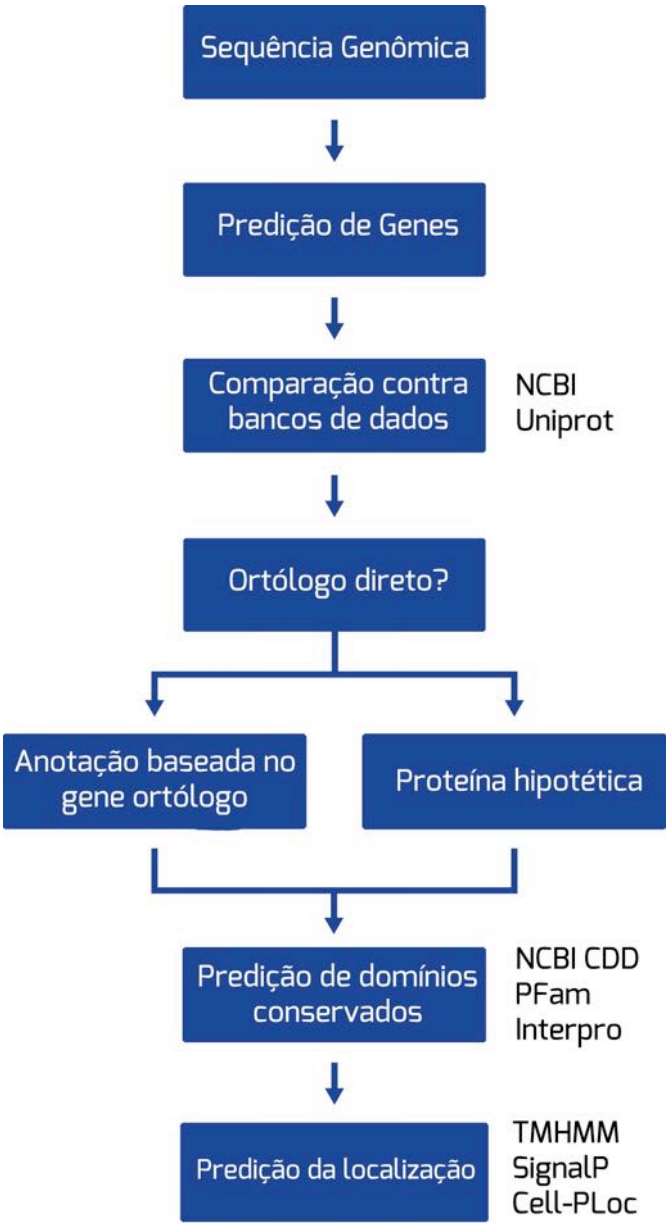


Figura 7-4: Um fluxo de trabalho genérico para anotação de genes.

O RNAi, algumas vezes denominado de "silenciamento gênico", é um mecanismo que induz a diminuição da expressão gênica de um transcrito alvo através da clivagem do transcrito alvo e sua posterior degradação, ou através da repressão da maquinaria de tradução. Estes mecanismos são denominados também de Silenciamento Gênico Pós-Transcricional (PTGS – no inglês) (Figura 8-4). Existem adicionalmente alguns pequenos RNAs que induzem silenciamento gênico em nível transcricional, ligando-se em regiões de DNA, impedindo sua transcrição. Este mecanismo é denominado de Silenciamento Gênico Transcricional (TGS – no inglês).

As metodologias de sequenciamento de alta eficiência tem auxiliado de maneira contundente na caracterização de pequenos RNAs, sendo que variações de protocolos também possibilitaram validar alvos (técnica de degradoma) e identificar pequenos RNAs associados com proteínas específicas (sequenciamento de ácidos nucleicos associados a proteínas imunoprecipitadas).

Existe uma grande diversidade de pequenos RNAs em células eucarióticas, sendo os principais listados na Tabela 4-4. Dentre estas, os microRNAs são a classe de pequenos RNAs melhor descrita. Caracterizam-se por serem transcritos a partir de genes MIR, geralmente intergênicos, por uma RNA polimerase II, resultando em um pri-miRNA, o qual recebe um 5'-CAP e um 3'-poli-A. Este pri-miRNA é processado por um complexo proteico, denominado D-*body*, o qual é orquestrado por uma enzima classicamente denominada DICER ou DROSHA (RNAses classe III), resultando na liberação do pré-miRNA. Este apresenta estrutura em forma de grampo devido à alta complementaridade que suas extremidades 5' e 3' possuem. O pré-miRNA é

novamente processado por uma enzima DICER, liberando o microRNA maduro, dupla-fita, de aproximadamente 20 nucleotídeos de comprimento, o qual é reconhecido por uma enzima ARGONAUTA e direcionado ao PTGS (Figura 9-4).

Outra classe bastante estudada se refere aos siRNA (*small interfering* RNAs), os quais tem a biogênese bastante variada, podendo ser derivados de regiões de sobreposição de genes em orientação inversa natsiRNAs (*natural anti-sense small interfering* RNAs). A transcrição de ambos transcritos resulta em uma região de dupla-fita complementar, a qual é reconhecida por uma enzima DICER que cliva o natsiRNA, resultando na sua forma madura (aproximandamente 24 nt).

Existem também os tasiRNA (*trans-acting small interfering* RNAs), derivados do processamento do transcrito alvo de um microRNAs. Para a síntese de tasiRNA, é necessário uma RNA polimerase dependente de RNA, a qual utiliza o microRNA como iniciador da transcrição e a sequência transcrito alvo como molde. O longo RNA dupla-fita resultante é reconhecido também por uma enzima DICER, a qual cliva o tasiRNA, resultando na sua forma madura (aproximadamente 20 nt).

Os siRNAs são reconhecidos por enzimas argonautas e podem tanto induzir o silenciamento gênico por PTGS, mas também o remodelamento de cromatina, controlando a expressão gênica em nível trancricional (TGS). A interação entre microRNAs e transcrito alvo é a melhor caracterizada, não sendo necessário uma complementariedade perfeita entre o microRNA e transcrito alvo, apesar disto ser mais comum em plantas. Em animais existe uma região de maior complementariedade denominada *seed* a qual se localiza entre a 2ª e 7ª bases no microRNA, e está relacionada à especificidade do microRNA com seu transcrito alvo. Outra característica é o fato de ha-

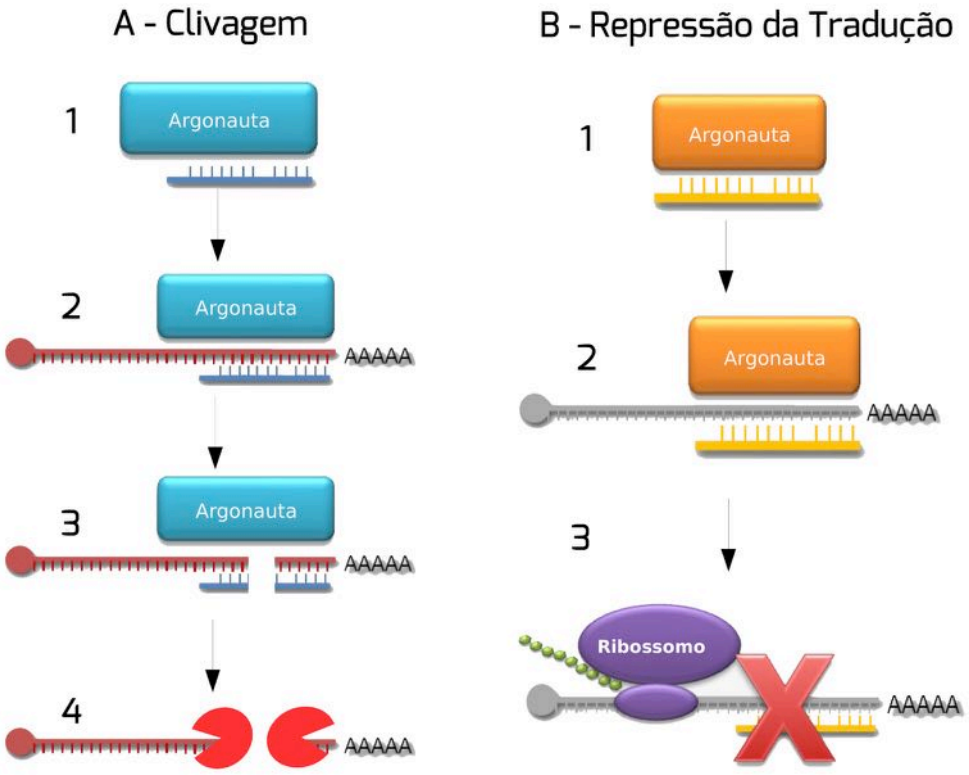


Figura 8-4: Mecanismo de PTGS. A) clivagem: 1, uma proteína argonauta reconhece uma fita do pequeno RNA; 2, O microRNA associado com uma argonauta reconhece um transcrito alvo; 3, ocorre a clivagem do transcrito alvo na posição medial do microRNA; 4, degradação do transcrito alvo clivado por nucleases. B) repressão da tradução: 1, uma proteína argonauta reconhece uma fita do pequeno RNA; 2, o microRNA associado com uma argonauta reconhece um transcrito alvo; 3, ocorre repressão da maquinaria de tradução.

Tabela 4-4: Principais classes de pequenos RNAs com função regulatória.

Classe	Tamanho (nt)	Função biológica	Mecanismo de ação	Origem	Organismos
microRNA ou miRNA	21-24	PTGS	Clivagem e repressão da maquinaria de tradução	Intergênica e íntrons	Plantas, animais, fungos e vírus
siRNA	21-24	PTGS, TGS	Clivagem, repressão da maquinaria de tradução e metilação de DNA	Intergênica, éxons e íntrons	Plantas, animais, fungos e vírus
tasiRNA	21-22	PTGS	Clivagem	Transcritos alvo de microRNAs	Plantas, animais e fungos
natsiRNA	21-22	PTGS	Clivagem	Transcritos convergentes parcialmente sobrepostos	Plantas

ver pareamento guanina – uracila (G-U), também denominado de *wobble* entre o transcrito alvo e o microRNA (Figura 9-4).

Existem dois desafios principais no emprego da bioinformática a pequenos RNAs. O primeiro é relativo à identificação da região, ou precursor, que dá origem ao pequeno RNA. O segundo envolve a identificação dos genes alvos regulados por estes. As metodologias de identificação da região que resulta no pequeno RNA variam com a classe de pequenos RNAs e estão intimamente relacionadas às suas biogêneses.

Os microRNAs são a classe melhor caracterizada, de forma que há uma maior disponibilidade de ferramentas para identificação destes, como os algoritmos miRTools, miRDeep, miRExpress, miRAnalyser e miRCat. A funcionalidade geral destes programas se baseia na análise de *reads* de sequenciamento de bibliotecas de pequenos RNAs e na delimitação das regiões de ancoramento com o genoma. Com base no conjunto de sequências ancoradas, são realizados cálculos para avaliação da estabilidade da possível estrutura em forma de grampo gerado pelo transcrito.

Para as demais classes, não existe uma metodologia padrão, sendo que variações da ferramenta BLAST são geralmente utilizadas. Para a identificar siRNAs, por exemplo, pode-se empregar a ferramenta SiLoCo. Mas é bastante comum laboratórios que pesquisam pequenos RNAs desenvolverem suas próprias ferramentas.

Já os programas de predição de alvos de microRNAs e siRNAs podem ser baseadas em ferramentas como o BLAST, procurando regiões complementares ao pequeno RNA. O problema é que esta técnica gera um número muito grande de falsos-positivos. Com isso, algumas ferramentas começaram a utilizar outros aspectos envolvidos na interação entre pequenos RNAs e transcritos alvos, tais como características energéticas, a presença da região *seed* (em humanos), o pareamento perfeito entre 10-11 pares de base do microRNA (válido somente para PTGS, por clivagem) e a conservação de microRNAs e transcritos alvo em organismos diferentes.

Mesmo assumindo estas regras, existem muitas interações entre microRNA e transcrito alvo que são excluídas, e muitas falsas que são incluídas, fazendo como que seja necessário a validação experimental desta interação. Especialmente para organismos modelo, existem bancos de dados próprios que disponibilizam, baseados em ferramentas de predição, os possíveis alvos para um determinado miRNA. Um importante banco de dados é o microRNA.org, cujas predições foram realizadas pelo algoritmo miRanda.

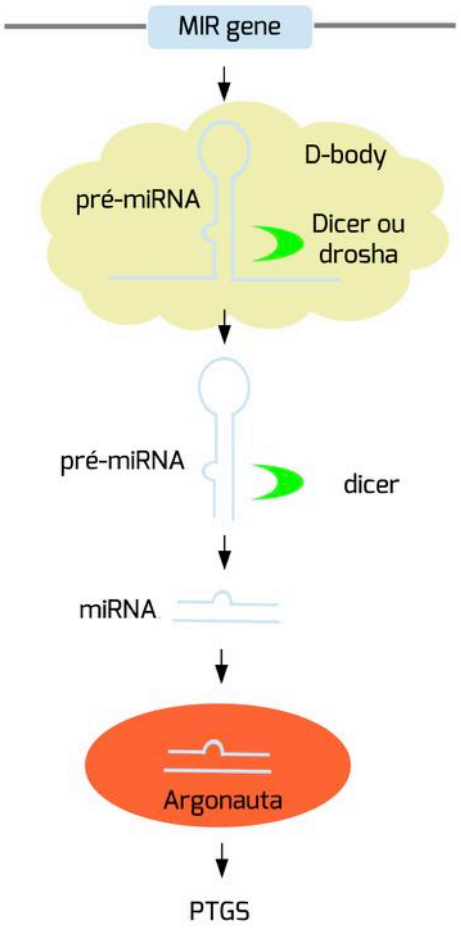


Figura 9-4: Modelo simplificado da biogênese de microRNAs. A partir de um gene MIR, um pré-miRNA é transcrito e processado num D-*body*, por uma enzima DICER, liberando o pré-miRNA, o qual é processado novamente por uma enzima DICER, liberando a forma madura do miRNA. Este é reconhecido por uma enzima argonauta e direcionado ao transcrito alvo, induzindo o silenciamento gênico.

4.6. Conceitos-chave

Anotação funcional: conjunto de abordagens que predizem a função e classificam uma proteína codificada por um genoma.

Contig: conjunto de segmentos de DNA com sobreposição de sequência que, conjuntamente, representam uma sequência consenso de DNA

Detectores de conteúdo: sistemas para delimitação de regiões codificantes baseados na classificação da sequência em codificante ou não codificantes, baseada em cálculos estatísticos ou em conservação de sequência. Compreendem detectores extrínsecos e intrínsecos.

Detectores de sinais: sistemas para delimitação de regiões codificantes baseados em caracteres funcionais de genes, como elementos canônicos necessários à transcrição ou tradução.

N50: índice associado à qualidade de montagem de um sequenciamento. Um valor de N50 igual a N significa que 50% dos *reads* estão montados em um *contig* de tamanho N ou maior.

ORF: *open reading frame* ou fase aberta de leitura. Refere-se a toda sequência nucleotídica delimitada por um códon de início e um códon de término de tradução.

Predição baseada em evidência: identificação de sequências codificantes baseada em experimentos prévios, como transcriptomas.

Predição *ab initio*: identificação de sequências codificantes baseada unicamente em cálculos estatísticos.

Reads: resultado obtido do sequenciamento de um determinado clone ou fragmento de DNA/cDNA.

Sequenciamento por *Shotgun*: metodologia de sequenciamento caracterizado por fragmentação aleatória de um grande segmento de DNA, determinação individual da sequência de cada um dos fragmentos e agrupamento dos *reads* obtidos em *contigs*.

Sinais transcricionais: sequências conservadas associadas ao processo de transcrição, como por exemplo TATA box, Sítios de clivagem e poliadenilação, etc.

Sinais traducionais: sequências conservadas associadas ao processo de tradução, como a sequência de Kozak, códon de início de

tradução, sítio de ligação de ribossomo, etc.

Transcriptoma: sequenciamento e avaliação geral de transcritos de uma célula/tecido com o intuito de descrever os RNAs presentes naquele momento. Além de trazer informações sobre a situação fisiológica daquele conjunto de células, permite construir modelos para procura de genes baseados em evidência.

4.7. Leitura recomendada

GARBER, M. et al. Computational methods for transcriptome annotation and quantification using RNA-seq. ***Nat. Methods***, 8, 469-477, 2011.

RICHARDSON, E. J.; WATSON, M. The automatic annotation of prokaryotic genomes. ***Brief. Bioinform***., 14, 36-45, 2013.

SLEATOR, R. D. An overview of the current status of eukaryotic prediction strategies. ***Gene***, 461, 1-10, 2010.

WILLIANSON, V. et al. Detecting miRNAs in deep-sequencing data: a software performance comparison and evaluation. ***Brief Bioinform***., 14, 36-45, 2013.

YANDELL, M.; ENCE, D. A beginner's guide to eukaryotic genome annotation. ***Nat. Rev. Genet***., 13, 329-342, 2012.