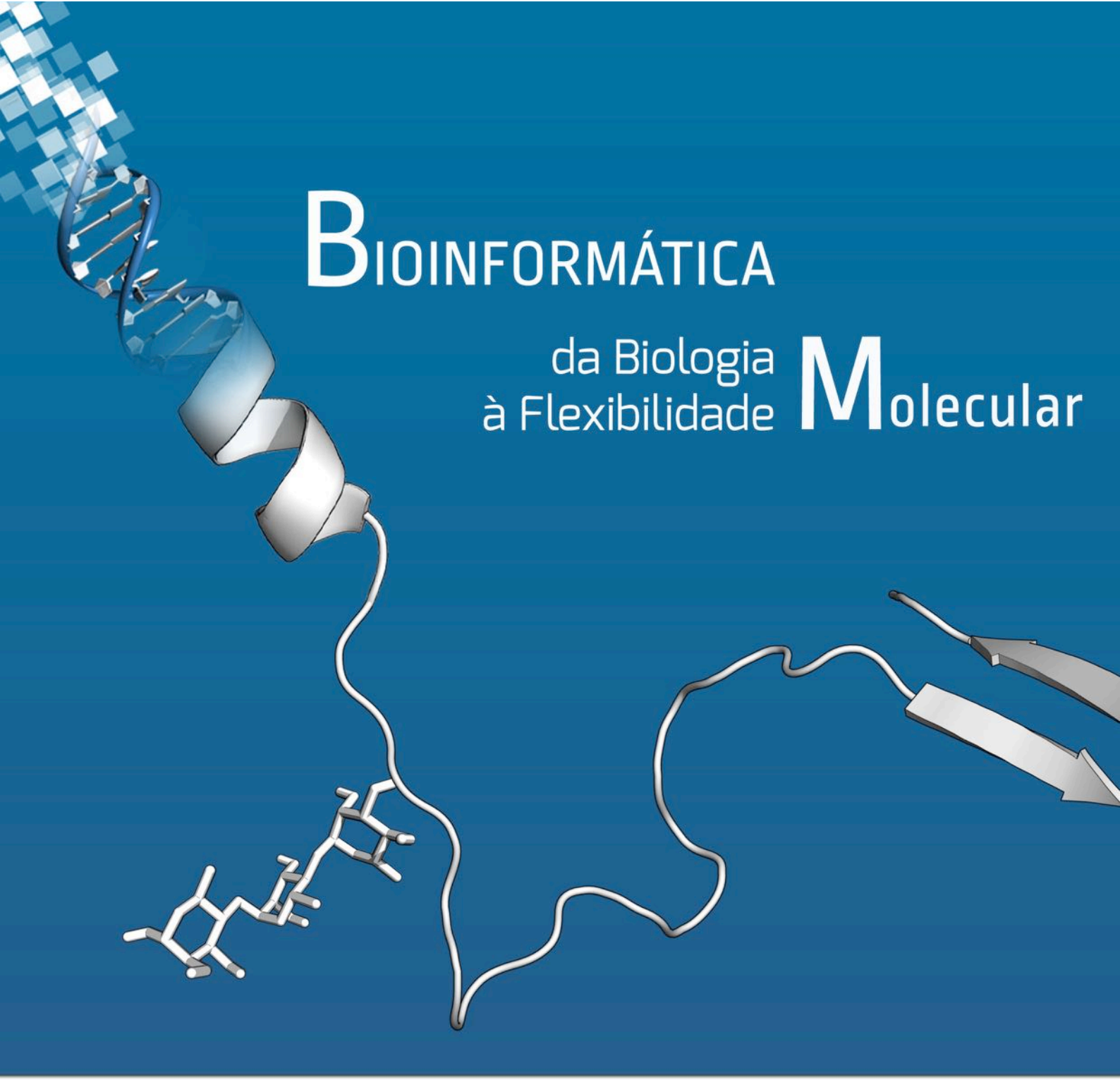


Bioinformática
da Biologia à Flexibilidade Molecular

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

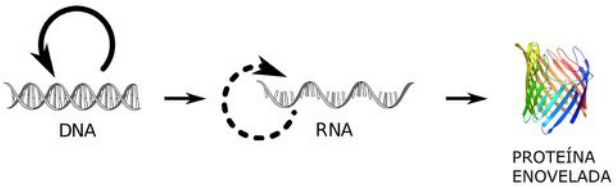
B615 Bioinformática da Biologia à flexibilidade molecular / organização de Hugo Verli. - 1. ed. - São Paulo : SBBq, 2014. 282 p. : il.

1. Bioinformática 2. Biologia Molecular

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

2. Níveis de Informação Biológica

Hugo Verli



Representação do fluxo de informação em sistemas biológicos.

2.1. Introdução

Por mais que possam apresentar enormes diferenças em suas características os seres vivos, desde bactérias a mamíferos, passando por plantas e fungos, são compostos aproximadamente pelos mesmos tipos de moléculas. Estes compostos incluem proteínas, ácidos nucleicos, lipídeos e carboidratos, moléculas nas quais a vida como conhecemos é baseada.

Cada uma destas classes de biomoléculas apresenta, contudo, enormes variações de forma, estrutura e função na natureza, o que possibilita a gigantesca variedade e complexidade de manifestações da vida em nosso planeta. Mesmo em estruturas que não são normalmente consideradas vivas, como é o caso dos vírus, estas biomoléculas são também encontradas e se mostram essenciais à execução de suas funções, sejam estas patológicas ou não.

Independentemente da forma pela qual a vida se manifesta, a informação que a rege está armazenada nas moléculas de DNA. Contudo, tais dados não são usados diretamente, mas através de uma molécula intermediária, o RNA (mais precisamente o RNAm), sintetizado por um processo denominado transcrição (uma molécula de ácido nucleico é transcrita em outra molécula de ácido nucleico). Esta molécula de RNAm irá servir como molde para a síntese de proteínas, em um processo chamado de tradução (uma molécula de ácido nucleico é traduzida em uma molécula de proteína). As proteínas, assim expressas, irão reger a maioria dos fenômenos relacionados à função dos organismos e à perpetuação da vida (embora diversos outros processos sejam modulados por outras biomoléculas). Esta informação segue um sentido tão conservado na natureza que foi convencionado denominá-lo como dogma central da biologia molecular (Figura 1-2).

A importância do dogma central no entendimento da informação e função biológicas pode ser exemplificada no fato de que ele aborda os três tipos mais comuns de moléculas estudadas por técnicas de bioinformática, o DNA, o RNA e as proteínas, estabelecendo um fluxo de informação universal à vida como conhecemos. Adicionalmente, a efetivação da informação genética, através das proteínas, acarreta na construção e manutenção de outras biomoléculas, igualmente essenciais ao desenvolvimento da vida, como carboidratos e lipídeos. Em decorrência de sua elevada massa molecular, proteínas, ácidos nucleicos, lipídeos agregados em membranas e carboidratos complexos são chamados de macromoléculas.

Embora carboidratos e lipídeos não estejam explici-

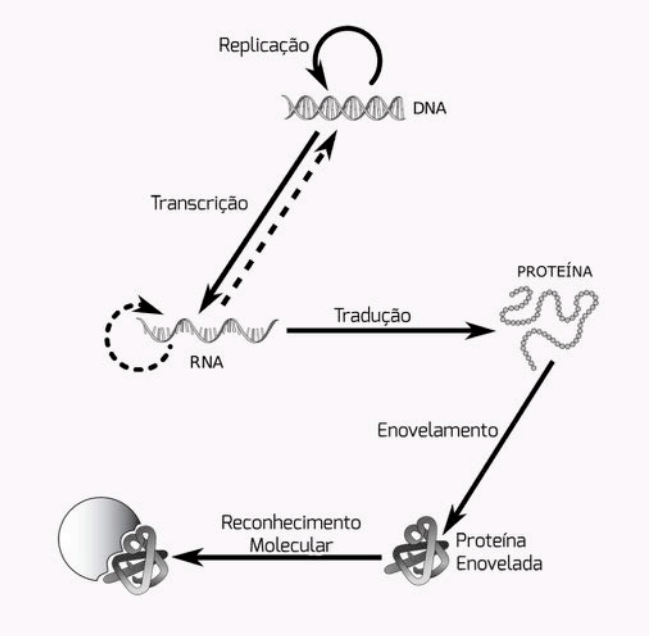


Figura 1-2: Representação do dogma central da biologia molecular, no qual o fluxo de informação em sistemas biológicos é descrito, desde seu armazenamento no DNA até a manifestação da função biológica. O esquema tradicional sofreu a adição do processo de enovelamento de de reconhecimento molecular devido ao seu caráter fundamental para a manifestação da função gênica. Adaptado de Hupé, 2012.

tamente inseridos no dogma central, não devemos minimizar sua importância. Apesar de por muito tempo estes compostos terem sido reconhecidos simplesmente por papéis energéticos e estruturais, ambos vêm sendo demonstrados como envolvidos em inúmeros fenômenos biológicos, como na glicosilação de proteínas e na formação de jangadas lipídicas. Estes, por sua vez, podem interferir diretamente na execução da função de proteínas e na homeostasia dos organismos.

Não somente macromoléculas são importantes biologicamente. Proteínas sintetizam uma infinidade de compostos de baixa massa molecular, ou micromoléculas, que atuam como neurotransmissores, sinalizadores e moduladores dos mais variados tipos representando, portanto, diferentes tipos de informação em sistemas biológicos. Por exemplo, a infecção do nosso organismo por bactérias desencadeia um processo inflamatório mediado por derivados lipídicos denominados prostaglandinas. Para combater micro-organismos competidores, fungos e bactérias produzem pequenos compostos com atividade antibiótica, muitos destes usados até hoje como fármacos. Desta forma, se a bioinformática se dedica ao estudo, por ferramentas computacionais, dos fenômenos relacionados à vida, o estudo de micromoléculas também torna-se foco da bioinformática ao abordar compostos relacionados à manutenção fisiológica ou terapêutica (neste caso, no planejamento de novos candidatos a agentes terapêuticos).

As técnicas modernas de bioinformática são capazes de lidar com todas estas biomoléculas que, contudo, possuem particularidades derivadas de suas diferenças químicas. Tais aspectos devem ser conhecidos de forma a permitir a construção de modelos computacionais mais precisos e adequados ao estudo dos mais diversos aspectos relacionados à vida.

Não há uma forma única de representar as diferentes moléculas biológicas. Cada estratégia de representação possui suas vantagens e desvantagens, que devem ser avaliadas de acordo com o estudo em andamento. Estratégias com menor volume de informação associado possuem menor custo computacional e, portanto, nos permitem avaliar rapidamente grandes quantidades de dados, por exemplo, genomas inteiros de diferentes organismos, cada um contendo dezenas de milhares de proteínas. Por outro lado, estratégias com maior volume de informação associado acarretam em custo computacional gigantesco nos limitando a, por exemplo, um punhado de proteínas, de dois ou três organismos. O trânsito por tal disparidade é um dos grandes desafios atuais para o profissional que trabalha com bioinformática.

2.2. Macromoléculas biológicas

As biomoléculas descritas no dogma central da biologia molecular, proteínas, DNA e RNA, são o que chamamos de biopolímeros, isto é, polímeros produzidos pelos seres vivos. Somam-se a este grupo de moléculas os carboidratos, que também podem ser encontrados como polímeros em meio biológico.

As propriedades de um polímero tornam-se consequência das propriedades de suas unidades monoméricas constituintes. No

caso dos biopolímeros, os monômeros podem ser aminoácidos, nucleotídeos e monossacarídeos. Assim, o conhecimento destas unidades básicas irá auxiliar diretamente no estudo de suas formas poliméricas e, por conseguinte, das funções biológicas destes polímeros sintetizados na natureza.

Ácidos nucleicos

Os compostos denominados ácidos nucleicos são polímeros sintetizados a partir de unidades denominadas nucleotídeos. Os nucleotídeos são formados por três partes constituintes: uma base nitrogenada, um carboidrato e um grupo fosfato. A base nitrogenada pode ser adenina (A), guanina (G), citosina (C), uracila (U) ou timina (T), enquanto a parte sacarídica poderá ser β-D-ribose (frequentemente abreviada simplesmente como ribose, para o RNA) ou a 2-desoxi-β-D-ribose (usualmente abreviada como desoxirribose, para o DNA) (Figura 2-2). Nas moléculas de ácidos nucleicos, os nucleotídeos são ligados através da denominada ligação fosfodiéster (ver adiante).

Quando a base nitrogenada está ligada ao carboidrato, na ausência do grupo fosfato, os compostos gerados são denominados nucleosídeos. Formados por ligação de diferentes nucleotídeos à β-D-ribose temos a adenosina, a guanosina, a citidina, a uridina e a timidina. A estes compostos podem ainda se ligar diferentes números de grupos fosfato. Assim, a adenosina pode se apresentar monofosfatada (AMP, do inglês *adenosine monophosphate*), difosfatada (ADP, do inglês *adenosine diphosphate*) ou ainda trifosfatada (ATP, do inglês *adenosine triphosphate*).

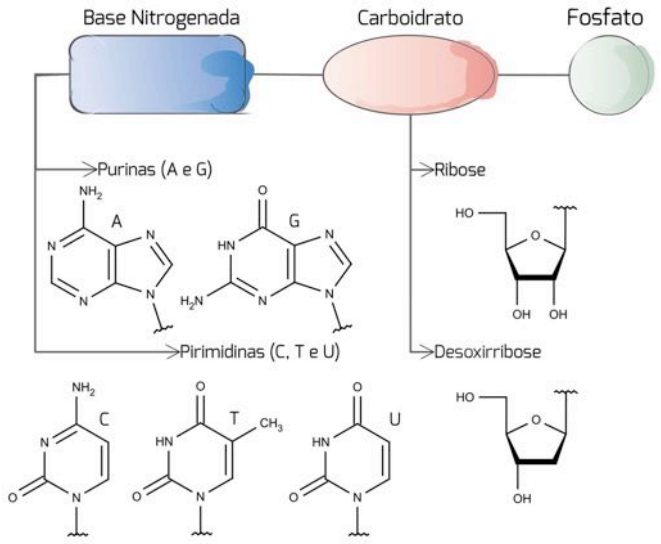


Figura 2-2: Representação esquemática de um nucleotídeo e suas variações na base nitrogenada e no carboidrato.

Conforme veremos adiante, carboidratos apresentam características conformacionais específicas, como sua capacidade de deformar seu anel em diferentes estados conformacionais. Esta característica se soma à grande flexibilidade da ligação fostodiéster na criação de um esqueleto bastante flexível para ácidos nucleicos. Em contrapartida a esta flexibilidade da parte sacarídica dos nucleotídeos, cada base nitrogenada é essencialmente planar, uma vez que constituem-se de anéis aromáticos, e portanto apresentam flexibilidade bastante reduzida.

Proteínas

As proteínas são polímeros sintetizados pelas células a partir de aminoácidos. São talvez as biomoléculas mais versáteis na natureza, sendo capazes de adotar uma gigantesca possibilidade de arranjos tridimensionais, não encontrada nos demais biopolímeros. Não por acaso, constituem-se no principal produto direto da informação genética, a partir da tradução do RNAm.

O genoma codifica diretamente 20 aminoácidos (22 contando selenocisteína e pirrolisina, que são codificadas por codons de parada) para composição de proteínas (Figura 3-2), embora outros resíduos de aminoácidos, não codificados no genoma (Figura 4-2), possam ser sintetizados a partir destes e exercer funções bastante específicas, como o ácido γ-amino butírico (GABA), um neurotransmissor inibitório no sistema nervoso central, ou como o resíduo ácido γ-carbóxi glutâmico (GLA), constituinte de diversas proteínas plasmáticas e fundamental na hemostasia.

Os aminoácidos codificados no genoma apresentam algumas características bem definidas e compartilhadas entre si. Todos os resíduos apresentam uma região comum, independente do resíduo. Esta região é denomi-

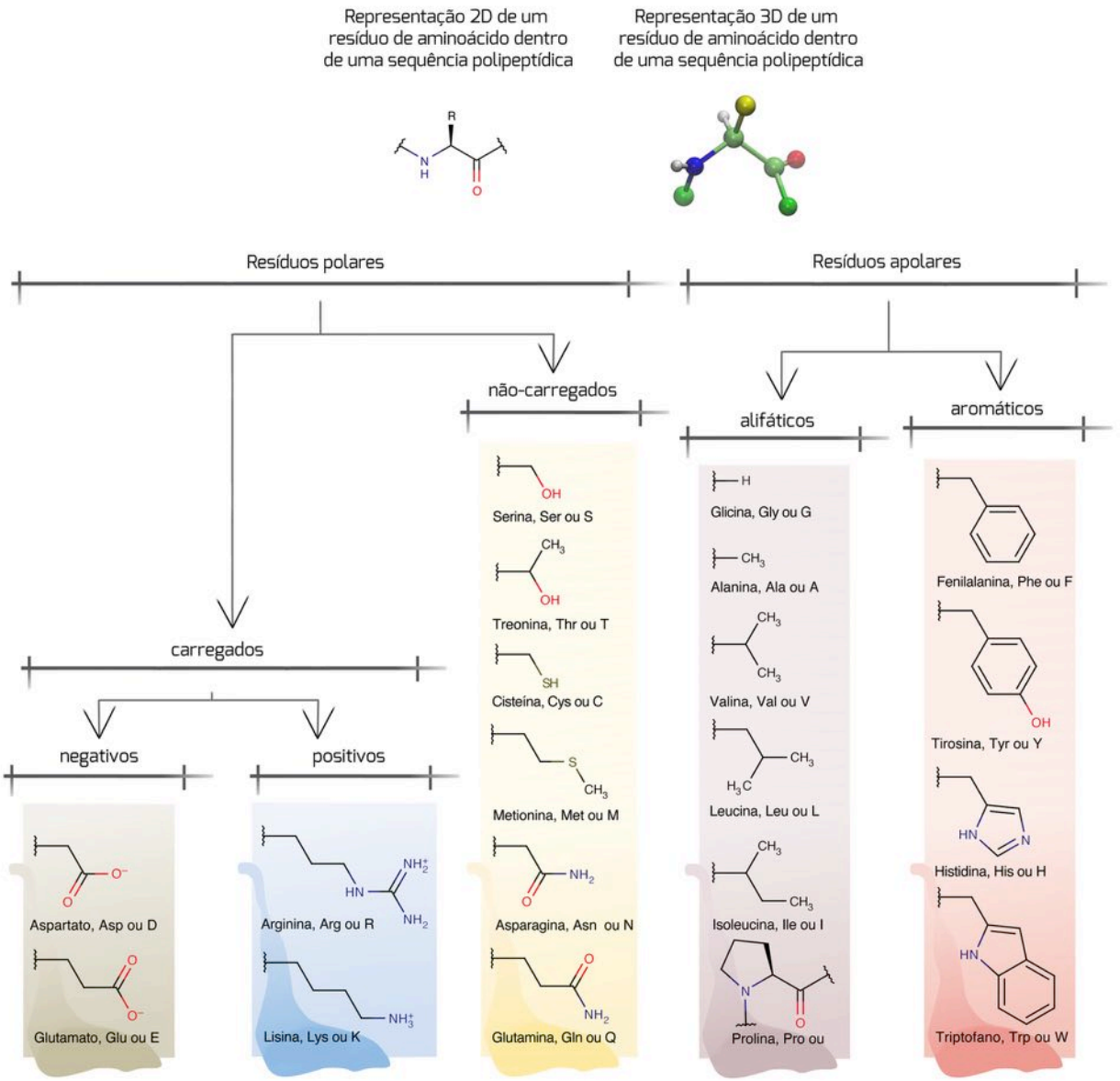


Figura 3-2: Estrutura dos aminoácidos codificados no genoma, organizados segundo as propriedades de suas cadeias laterais. No topo o esqueleto peptídico é representado como encontrado dentro de uma proteína, tanto em sua forma 2D quanto 3D. Nesta última, o grupo R (cadeia lateral) está apresentado como uma esfera amarela, enquanto a continuação da cadeia polipeptídica como esferas verde-escuras. As cadeias laterais estão apresentadas em sua ionização mais comum, plasmática.

nada esqueleto peptídico, e é composta pelo grupo amino, pelo grupo ácido carboxílico e pelo átomo de carbono que liga estes dois grupos, denominado carbono α (Cα). A diferença entre estes resíduos está no grupamento ligado ao Cα, chamado cadeia lateral (Figura 3-2).

Enantiômeros são compostos que, diferindo somente no arranjo de seus átomos no espaço (como no caso de L-Ser e D-Ser), correspondem um à imagem especular do outro (isto é, uma é o reflexo em um espelho da outra).

À exceção da glicina, todos os aminoácidos são quirais, em decorrência da presença de quatro substiuintes diferentes ligados ao Cα. Salvo casos específicos, todos os aminoácidos quirais são encontrados em somente uma forma enantiomérica, L. Como consequência, todas as proteínas são quirais, e isto tem implicações importantes em fenômenos bioquímicos e na prática terapêutica.

Dois enantiômeros interagem de forma idêntica com compostos que não sejam quirais. Por exemplo, a

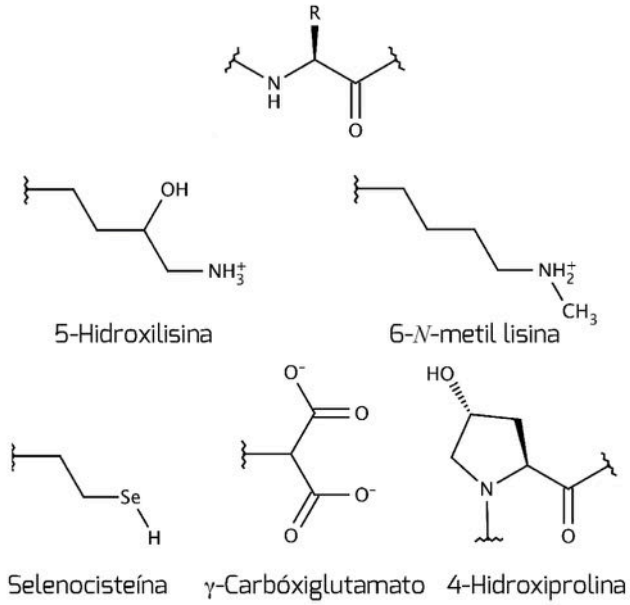


Figura 4-2: Exemplos de aminoácidos encontrados em nosso organismo mas não codificados no genoma humano.

interação de L-Ser e D-Ser com a água é idêntica. Em contrapartida, compostos quirais interagem diferentemente com cada enantiômero. Assim, a interação de L-Ser e D-Ser com uma dada proteína seria diferente. Assim, se tivermos um fármaco quiral, uma de suas formas enantioméricas será ativa e a outra provavelmente inativa, menos ativa ou mesmo tóxica.

O esqueleto peptídico de aminoácidos apresenta um grupo do tipo ácido carboxílico somente em aminoácidos livres, monoméricos, ou na posição terminal da proteína, denominada região C-terminal (o final da sequência polipeptídica). Da mesma forma, só encontramos o grupo amino na região demominada N-terminal (o início da sequência polipeptídica). À exceção destas extremidades, os grupos amino e carboxílico reagem, dando origem a um grupo amida. Assim, dentro de uma proteína, cada aminoácido contribui com um um átomo de nitrogênio e com uma carbonila para a formação de uma amida contida no esqueleto peptídico.

Os aminoácidos frequentemente são agrupados de acordo com as propriedades de suas cadeias laterais (Figura 3-2). Inicialmente, podem ser separados em resíduos polares e apolares. Os resíduos polares incluem aminoácidos não-carregados e carregados (com carga positiva ou negativa), enquanto os resíduos apolares incluem aminoácidos aromáticos e alifáticos (não aromáticos).

As propriedades dos aminoácidos são altamente influenciadas pelo pH do meio circundante. De acordo com sua acidez ou basicidade, a carga dos resíduos pode ser modificada e, por conseguinte, algumas propriedades da proteína. Assim, dependendo do compartimento celular, uma mesma proteína pode apresentar ionização distinta de seus resíduos de aminoácidos e, por conseguinte, propriedades eletrostáticas diferentes. Tais características destacam a importância de uma avaliação adequada do estado de ionização dos resíduos de aminoácidos das proteínas em estudo, principalmente o resíduo de histidina.

Durante a síntese proteica, os aminoácidos são conectados através da denominada ligação peptídica (ver adiante). Neste processo, o grupo carboxilato de um resíduo e o o grupo amino de outro resíduo de aminoácido reagem, dando origem a um grupo amida que compõe a ligação peptídica.

Carboidratos

Carboidratos compõem um terceiro grupo de biomoléculas. São compostos que, ao contrário das proteínas, não estão codificados diretamente no genoma. Enquanto a síntese de proteínas é guiada por um molde (a molécula de RNAm), a síntese de carboidratos não segue uma referência direta, mas um processo complexo e menos específico.

Embora o genoma não codifique a sequência oligossacarídica, ele determina a expressão de diversas enzimas que sintetizam carboidratos, ligam-os a outras estruturas polissacarídicas ou ainda modificam os resíduos monossacarídicos, adicionando ou removendo grupamentos substituintes nos anéis furanosídicos ou piranosídicos (Figura 5-2). Todo este processo é bastante específico, envolvendo tipos de monossacarídeos ou ainda posições específicas dentro destas moléculas. Uma das principais famílias de enzimas envolvidas neste processo são as denominadas glicosil transferases.

Esta família de biomoléculas apresenta uma grande variedade de formas (e, por conseguinte, funções), desde suas formas monoméricas até grandes polímeros com centenas de unidades monossacarídicas. São encontrados ligados a proteínas, formando as chamadas glicoproteínas; sulfatados, dando origem aos glicosaminoglicanos; ligados a lipídeos em membranas celulares (os glicolipí-

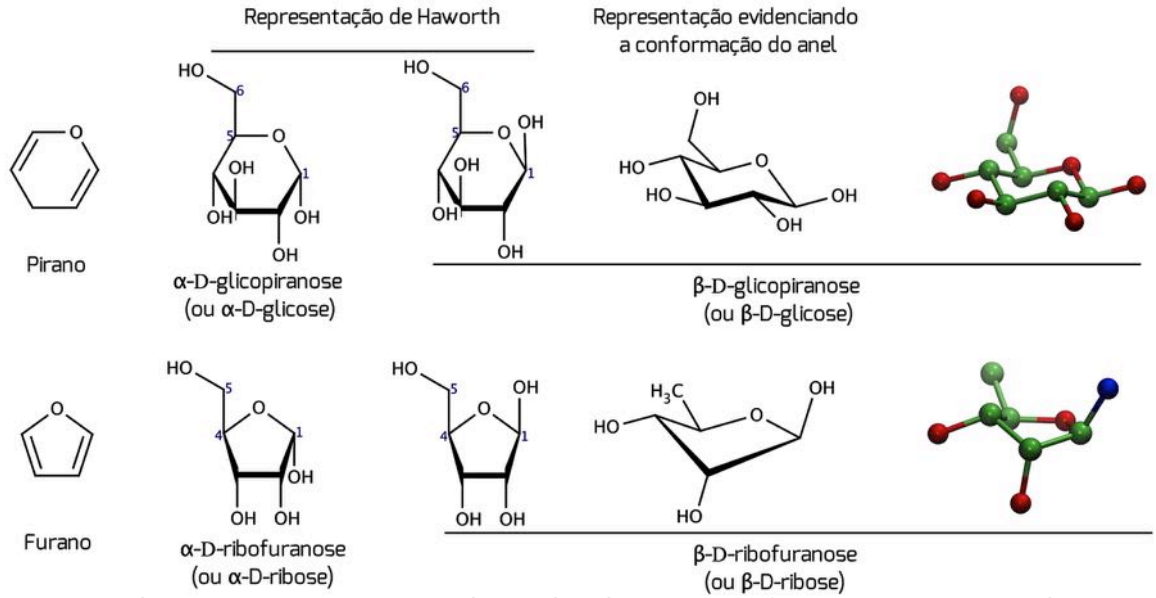


Figura 5-2: Os dois principais grupos de carboidratos envolvem monossacarídeos compostos por anéis de 5 (furanoses) e 6 membros (piranoses). São apresentados 3 tipos de visualização para estas moléculas, duas 2D e uma 3D.

deos) e como exopolissacarídeos da parede celular de fungos, dentro outros.

A forma majoritária de monossacarídeos biológicos em solução é um ciclo, mais comumente composto por 5 ou 6 átomos. Os carboidratos com anéis de 5 membros são denominados furanoses (como a ribose e a desoxirribose), por semelhança ao composto furano, enquanto os carboidratos com anéis de 6 membros são denominados piranoses (como a glicose, a manose e a galactose), pela sua similaridade com o composto pirano (Figura 5-2).

Estes anéis apresentam características conformacionais importantes. No caso das furanoses, podem ser as formas em envelope e torcida. No caso das piranoses, podem ser as formas em cadeira e bote torcido (Figura 6-2). Cada uma destas formas pode apresentar ainda variações, específicas para cada carboidrato em solução. Esta transição entre diversos estados conformacionais de monossacarídeos é denominada de equilíbrio pseudo-rotacional.

Os carboidratos possuem algumas diferenças importantes em relação aos aminoácidos. São, em geral, compostos mais polares, o que indica que irão interagir fortemente com a água. Outra diferença importante se refere à sua diversidade. Em comparação aos 20 aminoácidos codificados no genoma, mais de 100 possíveis unidades monossacarídicas já foram observadas como presentes em biomoléculas (Figura 7-2).

Em analogia à ligação peptídica, carboidratos são ligados entre si (ou a outras moléculas) através da denominada ligação glicosídica. Contudo, aminoácidos possuem somente um grupo amino e um grupo ácido carboxílico em seu esqueleto peptídico, de forma que somente um tipo de ligação peptídica é possível entre dois resíduos (o mesmo se dá com nucleotídeos). Como a ligação glicosídica entre dois monossacarídeos é formada pela reação entre dois grupos hidroximetileno (CHOH), e cada monossacarídeo possui vários destes grupos, múltiplas ligações entre dois monossacarídeos consecutivos tornam-se possíveis. Cria-se, assim, um complexo espectro de possíveis ligações entre os mesmos dois monossacarídeos.

O átomo de carbono na posição 1 (C1) de um monossacarídeo apresenta propriedades específicas, sen-

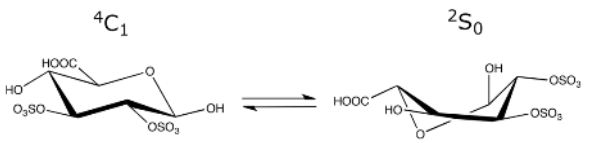


Figura 6-2: Equilíbrio conformacional entre a forma de cadeira e bote torcido para o resíduo de ácido idurônico, componente da heparina.

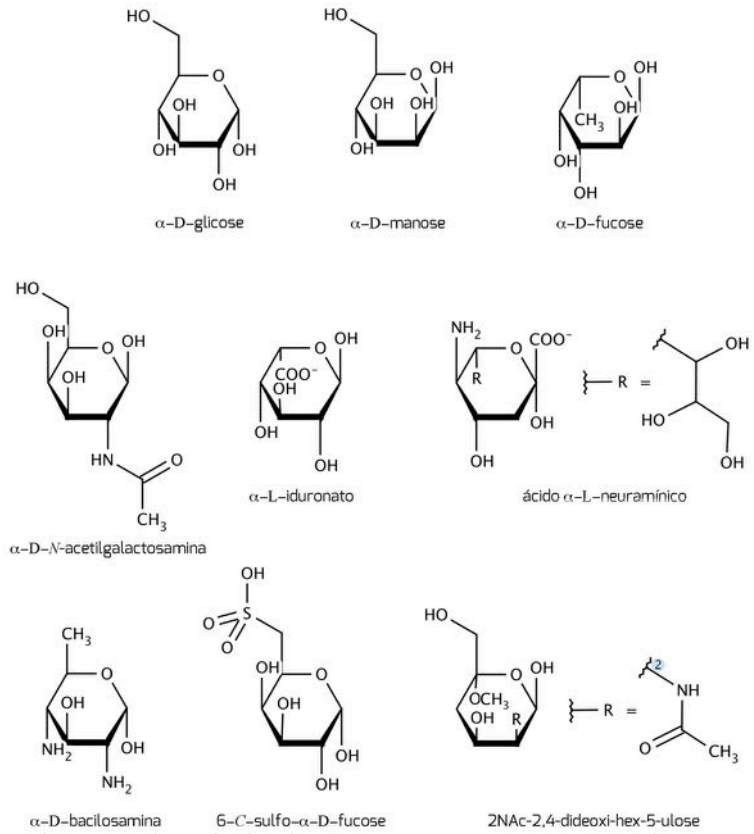


Figura 7-2: Exemplo da complexidade de possíveis monossacarídeos encontrados na natureza.

do denominado carbono anomérico. Para um mesmo monossacarídeo, o carbono anomérico pode ser encontrado em duas possíveis configurações, α e β (Figura 5-2). Assim, uma ligação glicosídica entre o carbono anomérico (C1) de uma manose e o átomo C3 de outra manose poderia ocorrer de duas formas, α-Man-(1→3)-Man ou β-Man-(1→3)-Man. No caso de glicoproteínas, contudo, a forma α é aquela usualmente encontrada para o resíduo de manose (para outros resíduos, a forma anomérica preferencial pode ser diferente).

Tomando como exemplo o tetrassacarídeo α-Man-(1→2)-α-Man-(1→2)-α-Man-(1→3)-Man, comumente encontrado em glicoproteínas do tipo oligomanose, o primeiro resíduo de manose (denominada extremidade não-redutora) possui seu carbono anomérico ocupado na ligação glicosídica, tendo sua configuração (neste exemplo α) fixa. Em contrapartida, o quarto resíduo de manose possui seu carbono anomérico livre. Esta porção é denominada redutora, e tem a configuração do carbono anomérico variável, isto é, pode estar tanto na forma α quanto β.

Membranas

Diferentemente dos ácidos nucleicos, proteínas e carboidratos, membranas não se constituem em polímeros biológicos, mas em agregados moleculares de lipídeos anfipáticos organizando uma bicamada (Figura 8-2). Apresentam papel fundamental à vida, compartimentalizando a célula, definindo seus limites, propriedades e organizando estruturas celulares.

É importante ter em mente que membranas são muito mais do que simples "paredes" delimitadoras da célula. Os componentes de membranas são variados, incluídos diferentes tipos de lipídeos, proteínas e carboidratos. A presença e localização destes componentes pode ser modulada de forma dinâmica em função de necessidades da célula, tecido ou organismo, sinalizando e modulando cadeias de eventos e definindo regiões da célula com propriedades específicas (a chamada polaridade celular).

Moléculas anfipáticas apresentam como característica a presença simultânea de uma região polar, também chamada de cabeça polar (hidrofílica ou lipofóbica) e de uma região apolar, também chamada de cauda hidrofóbica (hidrofóbica ou lipofílica). Assim, membranas celulares possuem superfícies polares e

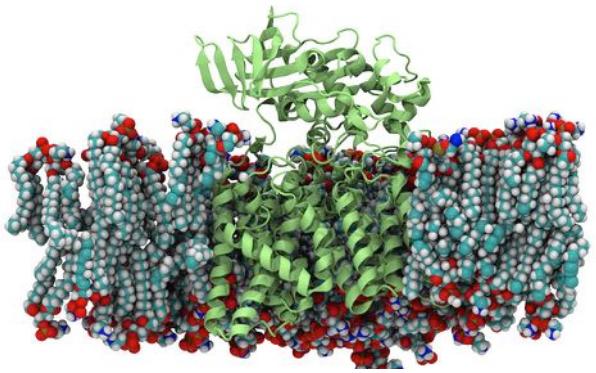


Figura 8-2: Representação de uma membrana POPE (palmitoil oleil fosfatidil etanolamina) contendo a enzima PglB (oligossacaril transferase) de *Campylobacter lari*. Os átomos de oxigênio estão representados em vermelho, os átomos de carbono em verde, os átomos de hidrogênio em branco e nitrogênios em azul. A enzima está representada como cartoon verde.

interiores apolares. As características destas duas regiões, contudo, podem variar bastante em função da composição dos lipídeos, interferindo na carga, espessura e fluidez da membrana (e, por conseguinte, na sua capacidade de modular fenômenos biológicos).

"Micromoléculas" biológicas

Quando pensamos nos efetores da informação genética é natural que a primeira família de biomoléculas que venha a nossa mente seja a das proteínas, codificadas diretamente no genoma. Contudo, como vimos anteriormente, outros tipos de biomoléculas são fundamentais ao funcionamento dos organismos, mesmo que estas não estejam codificadas diretamente no DNA.

Da mesma forma como não há um conjunto de bases nitrogenadas que codifique monossacarídeos ou lipídeos, diversos compostos de baixa massa molecular (por isso muitas vezes chamados de micromoléculas, em oposição às macromoléculas, compostos de elevada massa molecular) não possuem codificação direta no genoma, mas são produzidos a partir de enzimas que, estas sim, têm suas sequências de aminoácidos definidas pela molécula de DNA. Neurotransmissores, hormônios, metabólitos primários e secundários em plantas e uma infinidade de compostos, em decorrência de sua importância biológica (e terapêutica), são potenciais alvos de estudos computacionais. Contudo, justamente em decorrência de sua grande variedade química, torna-se difícil estabelecer padrões ou referências estruturais, como é o caso das biomacromoléculas vistas anteriormente. Frequentemente, esta característica cria uma série de dificuldades e desafios no emprego de ferramentas computacionais no estudo de micromoléculas. Dentre estas dificuldades destaca-se a necessidade de desenvolvimento de parâmetros específicos para cada molécula (como veremos no capítulo 8).

2.3. Níveis de organização

A classificação da estrutura de biomacromoléculas envolve, didaticamente, quatro diferentes níveis de complexidade. Esta separação facilita o nosso entendimento do como e do porquê macromoléculas adotarem determinadas formas em meio biológico e, a partir destas, desempenharem funções específicas. Adicionalmente, cada nível traz volume e tipos de informação diferentes, exigindo poder computacional e abordagens distintas, como veremos adiante.

Em princípio, estes níveis apresentam um componente hierárquico, ou seja, a informação de um nível é importante ou necessária para o nível de complexidade seguinte. Contudo, outros fatores podem participar neste processo.

Por exemplo, no caso das proteínas, embora normalmente consideremos que a informação contida na estrutura 1^{ária} (isto é, a sua sequência de aminoácidos) seja determinante para a sua estrutura 2^{ária}, ela não é o único determinante. Concessões podem ser realizadas para permitir uma estrutra 3^{ária} ou mesmo 4^{ária} mais estável.

Assim, uma determinada região em hélice pode ser parcialmente desestruturada para facilitar a formação de um determinado domínio (ver adiante). Este tipo de consideração é importante na validação de modelos teóricos para a estrutura de proteínas, como veremos no capítulo 7.

Adicionalmente, fatores externos à própria sequência proteica podem interferir nestes níveis de organização. Um dos fatores mais comuns é a glicosilação de proteínas, que frequentemente estabiliza partes da mesma e, assim como as chaperonas, pode interferir na forma proteica tridimensional existente em meio biológico.

Estrutura 1ária

O nível inicial de complexidade, a estrutura 1ária, consiste num padrão de letras (ou pequenos conjuntos de letras) que representa a composição do biopolímero. Esta sequência de letras representa uma informação de natureza unidimensional (1D), em que a única dimensão descrita é a ordem de aparecimento dos monômeros.

Para ácidos nucleicos, a estrutura 1ária consiste numa sequência de nucleotídeos, enquanto para proteínas em uma sequência de aminoácidos e, para carboidratos, em uma sequência de monossacarídeos (Figura 9-2). Este último caso é o único para o qual não há uma descrição de uma única letra para cada monômero, principalmente em face do elevado número de possíveis monômeros encontrados na natureza, maior que o número de letras no alfabeto.

Embora de menor complexidade, a estrutura 1ária nos oferece um grande volume de informações sobre a forma nativa da biomolécula e, por conseguinte, sobre suas funções. Tais informações advém principalmente da comparação de sequências de biomoléculas (aminoácidos ou nucleotídeos) em busca de padrões específicos associados a determinadas características ou funções. Uma vez identificados, esses padrões ou assinaturas podem ser usados na busca das mesmas características em outras proteínas, desconhecidas. Estas comparações ainda nos permitem estudar a evolução destas biomoléculas e de seus organismos, contribuindo no entendimento de como a vida se desenvolveu e atingiu o seu estágio atual de complexidade (ver capítulo 5).

DNA:
GGTATAGGCGCTGTTCTTAAGGTGCTAACAACGGGGT
TACCCGCGTTGATCTCGTGGATAAAACGCAAACGCCA
ACAG

RNA:
GGUAUAGGCGCUGUUCUUAAGGUGCUAACAACGGG
GUUACCCGCGUUGAUCUCGUGGAUAAAACGCAAAC
GCCAACAG

Aminoácidos:
GIGAVLKVLTTGLPALISWIKRKRQQ

Sequência sacarídica:
α-D-GlcNAc,6S-(1→3)-β-D-GlcA-(1→4)-α-D-
GlcNS,3S,6S-(1→4)-α-L-IdoA,2S-(1→4)-α-D-
GlcNS,6S

Figura 9-2: Representação da estrutura 1ária de diferentes biomacromoléculas: DNA, RNA, proteína (estas três representando o peptídeo melitina, componente do veneno da abelha *Apis mellifera*) e carboidratos (representando uma sequência repetitiva de heparina). A letra S na sequência oligossacarídica indica sulfatação.

Estrutura 2ária

A partir da sequência de monômeros descritos, em uma determinada ordem específica, na estutura 1ária surgem interações entre monômeros vizinhos e com as moléculas de solvente circundantes. Por exemplo, enquanto dois nucleotídeos vizinhos tendem a "empilhar" os anéis das bases, uma cadeia lateral de um aminoácido polar vai se expor à água, maximizando interações por ligação de hidrogênio com este solvente. De forma semelhante, uma cadeia apolar irá se expor aos lipídeos em uma membrana, maximizando interações hidrofóbicas com este outro solvente.

Estas interações entre monômeros acabam por dar origem a padrões repetitivos de organização espacial, denominados de estrutura 2ária (Figura 10-2). Estes padrões ou elementos aparecem em número relativa-

mente pequeno de tipos, de forma que a estrutura tridimensional de biomoléculas pode ser descrita como uma combinação de conjuntos destes elementos.

Diferentes composições de estrutura 1ária podem gerar um mesmo tipo de estrutura 2ária. Não por acaso, as propriedades destas estruturas 2árias, mesmo que formadas por sequências diferentes, apresentam semelhanças. Por exemplo, uma alça em proteínas é frequentemente uma estrutura 2ária bastante flexível, enquanto folhas e hélices tendem a ser mais rígidas.

As estuturas 2árias mais frequentemente lembradas são aquelas relacionadas a proteínas. Incluem três grupos de elementos principais: as alças, as hélices e as folhas β.

As alças ou voltas são elementos envolvidos na conexão entre hélices e folhas. Tendem a ser, portanto, estruturas flexíveis para acomodar as mais variadas orientações que estas hélices e fitas podem adotar entre si. Embora alças pequenas possam ser bastante rígidas, suas flexibilidades tendem a aumentar conforme o tamanho da alça aumenta (Tabela 1-2). Justamente em função desta elevada flexibilidade, alças são mais susceptíveis evolutivamente a sofrerem mutações (salvo se estiverem sob alguma pressão evolutiva, determinada por alguma função específica). Em outras palavras, a troca de um resíduo por outro de propriedades distintas pode ser mais facilmente acomodada nesta estrutra flexível do que nos outros tipos de estrutura 2ária, mais rígidos.

Enquanto hélices e folhas apresentam periodicidade ao longo de suas estruturas (semelhança nos pares de ângulos φ e ψ a cada aminoácido, ver adiante), alças se distinguem por não apresentarem periodicidade. Ainda, embora alças sejam frequentemente consideradas como elementos sem estrutura definida (as chamadas *random coils*), ou mesmo com estrutura aleatória, isto não é sempre verdade. Alças podem adotar formas mais definidas, dependendo de seu tamanho e composição.

De forma semelhante, é equivocado subestimar a importância das alças, considerando somente seu papel como elemento de conexão. Alças apresentam diversos impactos funcionais importantes em proteínas.

Tabela 1-2: Tipos de alças mais comuns encontrados em proteínas.

Tipo	Tamanho (nº de resíduos)
voltas γ	3
voltas β	4
voltas α	5
voltas π	6
alças Ω	6-16[a]
alças ζ	6-16[a]

[a] A despeito de tamanhos semelhantes, as formas destas alças se aproximam das letras que as denominam. Na volta Ω os resíduos das extremidades da alça estão próximos, e na volta ζ observa-se uma distorção na geometria.

Por exemplo, sua flexibilidade permite que atuem como tampas ou abas, cobrindo sítios ativos e regulando o acesso de moduladores ou substratos. De forma ainda mais direta, alças são frequentemente os elementos de estrutura 2ária mais expostos ao solvente. Assim, muitas vezes envolvem-se em contatos proteína-proteína (ou com outras biomoléculas), os quais podem ser determinantes para a função proteica. Assim, embora mais susceptíveis evolutivamente a mutações, não são incomuns alças com resíduos conservados, fundamentais para suas respectivas funções biológicas.

A hélice α e as folhas β foram inicialmente descritos por Linus Pauling e Robert B. Corey em 1951, embora as primeiras propostas para as estruturas em folhas datem de décadas mais cedo, em 1933, por Astbury e Bell. As folhas β são formadas por sequências de aminoácidos (cada sequência é denominada de fita) quase completamente extendidas. Estas fitas, quase lineares, interagem lado a lado ao longo de seus eixos longitudinais, através de uma série de ligações de hidrogênio entre o grupamento N-H de uma fita e o grupamento C=O da fita vizinha (Figura 10-2). Para que esta organização seja possível, os átomos de Cα adotam orientação intercalada, acima e abaixo do plano da folha. Esta organização se assemelha a uma série de dobraduras em uma folha de papel, de forma que este tipo de estrutura 2ária é tam-

bém denominado de folhas β pregueadas (Figura 10-2).

A forma pregueada de folhas β também é acompanhada pelas cadeias laterais dos resíduos de aminoácidos, ora acima do plano da folha, ora abaixo. Contudo, resíduos em fitas vizinhas orientam suas cadeias laterais para o mesmo lado, frequentemente de forma justaposta (Figura 10-2). Isto permite, por exemplo, que uma face da folha seja hidrofóbica e a outra hidrofílica.

A organização das fitas em folhas pode seguir duas orientações possíveis: *i*) a porção N-terminal de uma fita interagindo com a porção N-terminal da fita vizinha (e, consequentemente, o C-terminal interagindo com o C-terminal), ou *ii*) a porção N-terminal de uma fita interagindo com a porção C-terminal da fita vizinha. Estas duas possibilidades de interações de fitas dão origem a dois tipos de folhas β: as paralelas e as antiparalelas.

As folhas β paralelas e antiparalelas diferem em outras características. Esta organização diferenciada das fitas acarreta, por exemplo, em um padrão distinto de ligações de hidrogênio. Enquanto nas folhas antiparalelas as ligações de hidrogênio formam um ângulo de 90º com as fitas, nas folhas paralelas estes ângulos se tornam maiores (e as interações mais fracas) (Figura 10-2).

As folhas β podem ser encontradas em formas puras, paralelas ou antiparalelas, ou mistas, em que folhas paralelas pareiam com folhas antiparalelas. Contudo, folhas β paralelas tendem a ser menos estáveis conformacionalmente que folhas β antiparalelas. Esta diferença pode ser bastante significativa, suficiente para acarretar na desnaturação de proteínas por seus inibidores, como foi proposto na ação de serpinas sob suas proteses alvo.

O trabalho pioneiro de Pauling e Corey no início dos anos 50 do século XX identificou não somente as folhas, mas também hélices em sequências polipeptídicas. A formação da hélice, de forma similar às folhas, também envolve a realização de ligações de hidrogênio entre grupos N-H e C=O vizinhos no espaço (mas não na sequência) (Figura 10-2). Contudo, enquanto nas folhas β estas interações se dão com resíduos em fitas vizinhas, nas hélices estas interações acontecem com resíduos mais próximos na sequência, entre as voltas da hélice.

Diversos tipos de hélices podem ser encontrados em proteínas (Tabela 2-2). A hélice mais comum, denominada de hélice α, apresenta 3,6 resíduos de aminoácidos por volta da hélice, e cada aminoácido (n) realiza ligação de hidrogênio com o quarto resíduo seguinte (n + 4), que perfaz (aproximadamente) uma volta completa da hélice. Outro tipo de hélice comum em alguns tipos de proteína é a hélice de poli-prolina II encontrada, por exemplo, em proteínas de parede celular de plantas e no colágeno. Neste tipo de hélice, contudo, como o átomo de nitrogênio da prolina está ligado a três átomos de carbono, não há formação de ligação de hidrogênio durante a organização da hélice.

Existem, ainda, outros tipos de hélice, menos comuns, como a hélice π e a hélice 3_{10} (Tabela 2-2). Quanto à nomenclatura, a hélice 3_{10} foge ao padrão de uso de letras gregas das hélices α e π. O número 3 representa o número de resíduos por volta da hélice, enquanto o número 10 reflete o número de átomos entre duas ligações de hidrogênio vizinhas dentro da hélice. Assim, segundo esta nomenclatura, a hélice α seria chamada de $3{,}6_{13}$ e a hélice π de $4{,}4_{16}$. Tais nomenclaturas, contudo, não são normalmente empregadas.

Não são só as proteínas que que apresentam estruturas 2árias. Ácidos nucleicos e carboidratos também podem apresentar padrões repetitivos de organização espacial, definidos pela sequência de monômeros que os constituem.

A molécula de DNA pode adotar três tipos de estrutura 2ária, denominados A, B e Z (Figura 11-2), embora a forma B seja a estrutura mais comum e a partir dela sejam definidas as fendas maior e menor do DNA (Tabela 3-2). A transição entre estas formas é determinada pela hidratação, tipos de cátions e da própria sequência de nucleotídeos. Contudo, a dificuldade em mimetizar as interações biológicas, envolvidas no DNA e em complexos DNA-proteínas, durante a determinação de estruturas 3D dificulta associações mais claras de cada tipo de estrutura 2ária a fenômenos específicos *in vivo*.

Diferentes tipos de estrutura 2ária acarretam em diferentes propriedades estruturais

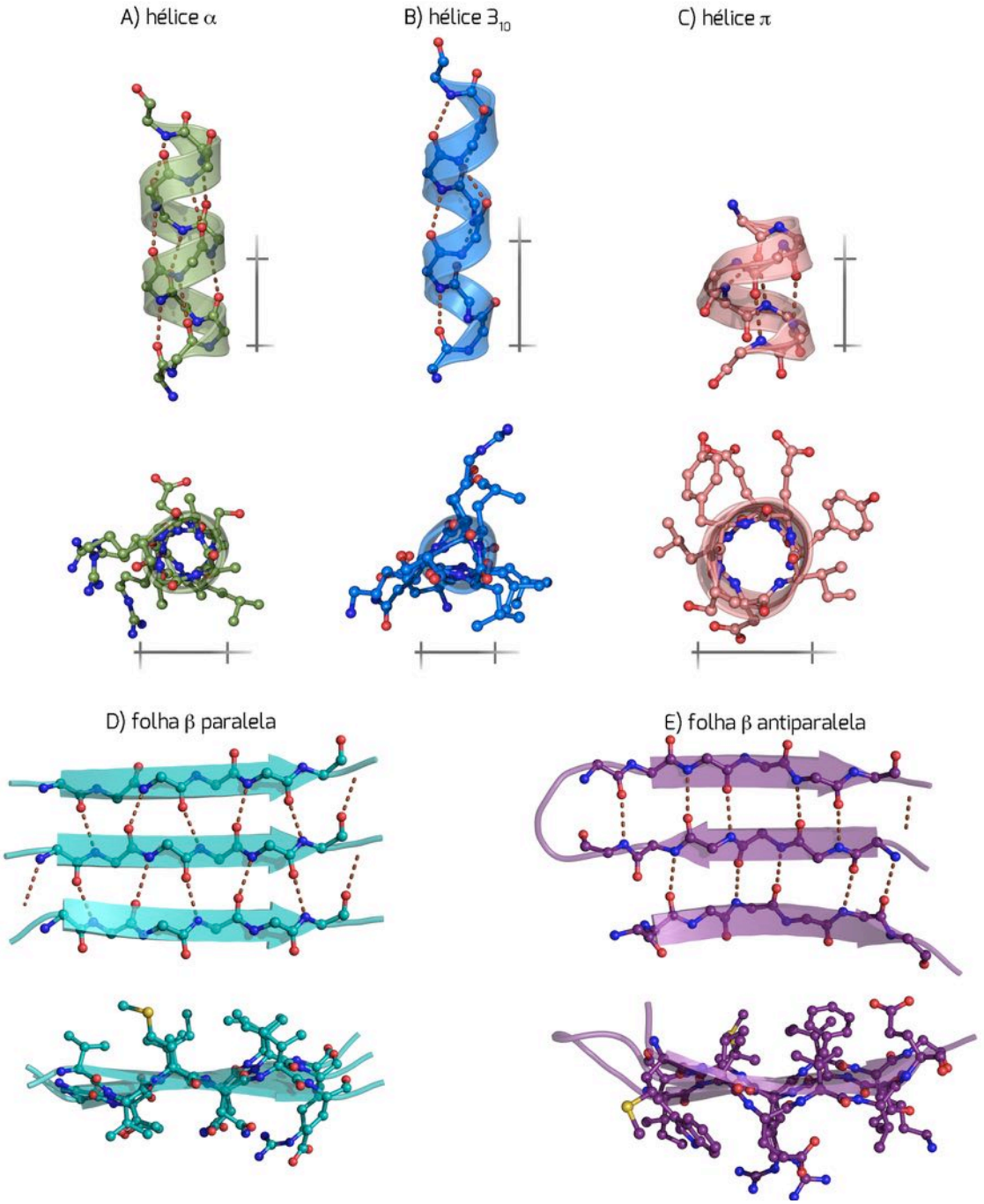


Figura 10-2: Representação dos tipos mais comuns de estrutura 2^{ária} encontrados em proteínas. Em verde estão as hélices α (A), em azul as hélices 3$_{10}$ (B), em salmão as hélices π (C), em ciano as folhas β paralelas (D) e roxo as antiparalelas (E). As ligações de hidrogênio entre átomos do esqueleto peptídico estão apresentadas como linhas tracejadas em marrom. As estruturas são partes que compõe as proteínas descritas pelos códigos PDB 18D8, 1ABB, 2QD1, 1EE6 e 1PCO, e para cada uma duas diferentes orientações são apresentadas. Note que as cadeias laterais apontam para fora do eixo das hélices e, para as folhas, para cima e para baixo do plano definido pelas fitas.

na molécula de DNA, como na largura e profundidade das fendas maior e menor e na disposição e orientação dos grupos fosfato, propriedades estas que, por sua vez, estão diretamente relacionadas à especificidade da interação do DNA com proteínas e fármacos.

A forma B do DNA pode assumir dois sub-estados, denominados BI e BII, definidos por diferenças em tor-

Tabela 2-2: Tipos de hélices encontrados em proteínas.

Tipo de hélice	Resíduos / volta	Ligação de hidrogênio	Elevação / resíduo (Å)	Elevação / volta (Å)	Direção mais comum
hélice α	3,6	$n + 4$	1,5	5,4	direita
hélice 3_{10}	3	$n + 3$	2,0	6,0	direita
hélice π	4,4	$n + 5$	1,2	5,3	direita
poli-Pro I	3,3	-	1,7	5,6	direita
poli-Pro II	3	-	3,1	9,3	esquerda

ções na parte sacarídica e no grupo fosfato (ver adiante). Essa região, formada por carboidrato e fosfato, é também denominada de esqueleto do DNA, em analogia ao esqueleto peptídico. A lógica é a mesma: o esqueleto é composto pela região comum a todos os monômeros formadores do biopolímero. Adicionalmente, outras formas de DNA já foram identificadas (alguns autores afirmam inclusive que poucas letras do alfabeto sobram para nomear novas formas de DNA que por ventura venham a ser identificadas), embora muitas ainda não tenham papel biológico claro.

A maioria dos genomas eucarióticos está sujeita a um fenômeno de metilação do DNA, que consiste na adição de um grupo metila no átomo de carbono na posição 5 dos resíduos de citosina. Como uma modificação estrutural epigenética envolvida na regulação do potencial regulatório e transcricional do DNA, deve-se estar atento à necessidade de incluir tal modificação na descrição deste ácido nucleico.

Não somente o DNA, mas também o RNA possui estrutura 2ária. Contudo, ao contrário do DNA, que é uma molécula contendo duas fítas de ácidos nucleicos, na maioria das situações o RNA é uma molécula composta por uma única fita. Assim, enquanto no DNA os pareamentos entre bases que dão origem à estrutura 2ária surgem da interação de moléculas (fitas) diferentes e complementares, no RNA a estutura 2ária surge de interações na própria fita, que dobra-se sobre si mesma.

As estruturas 2árias de RNA incluem regiões de bases pareadas, alças de grampos, alças internas, bojos (do inglês *bulge*) e junções. Quando o RNA se dobra sobre si, ele forma pareamentos entre bases complementares de forma análoga àquelas vistas no DNA. Quando uma das fitas no RNA pareado apresenta bases que não possuem uma contrapartida para formar um par A-U ou C-G, forma-se uma protuberância ou bojo.

Estes bojos, isto é, bases não pareadas em uma dupla-fita, também podem ser encontradas em folhas β. Neste caso, resíduos de aminoácidos de uma fita deixam de interagir com a fita vizinha, dando origem a este outro tipo de estrutura 2ária de proteínas.

As alças de grampos em moléculas de RNA são análogas às voltas observadas em proteínas, conectando duas fitas β por um pequeno segmento de poucos resíduos. No RNA, quando a fita dobra-se sobre si mesma, deixa alguns resíduos (no mínimo 4) projetados para fora, formando uma alça. Neste tipo de estrutura 2ária, a alça está vizinha a somente uma região de pareamento de bases, enquanto que há duas regiões, a cada lado do bojo, de bases pareadas.

As alças internas podem ser entendidas como uma dupla fita de DNA em que, no seu meio, as bases não são complementares e, por isso, não pareiam. Assim, ambas as fitas apresentam bases que não estão pareadas, o que a diferencia do bojo. Por fim, as junções conectam 3 ou mais regiões de bases pareadas.

O terceiro tipo de biopolímero constituinte de biomacromoléculas, os carboidratos podem, similarmente a proteínas e ácidos nucleicos, adotar padrões repetitivos de organização de suas unidades formadoras, monossacarídeos, isto é, em elementos de estrutura 2ária.

Polissacarídeos lineares desenvolvem estruturas de hélices, similarmente à proteínas e ácidos nucleicos. No caso destas moléculas, contudo, a variabilidade de organizações possíveis é muito maior, de for-

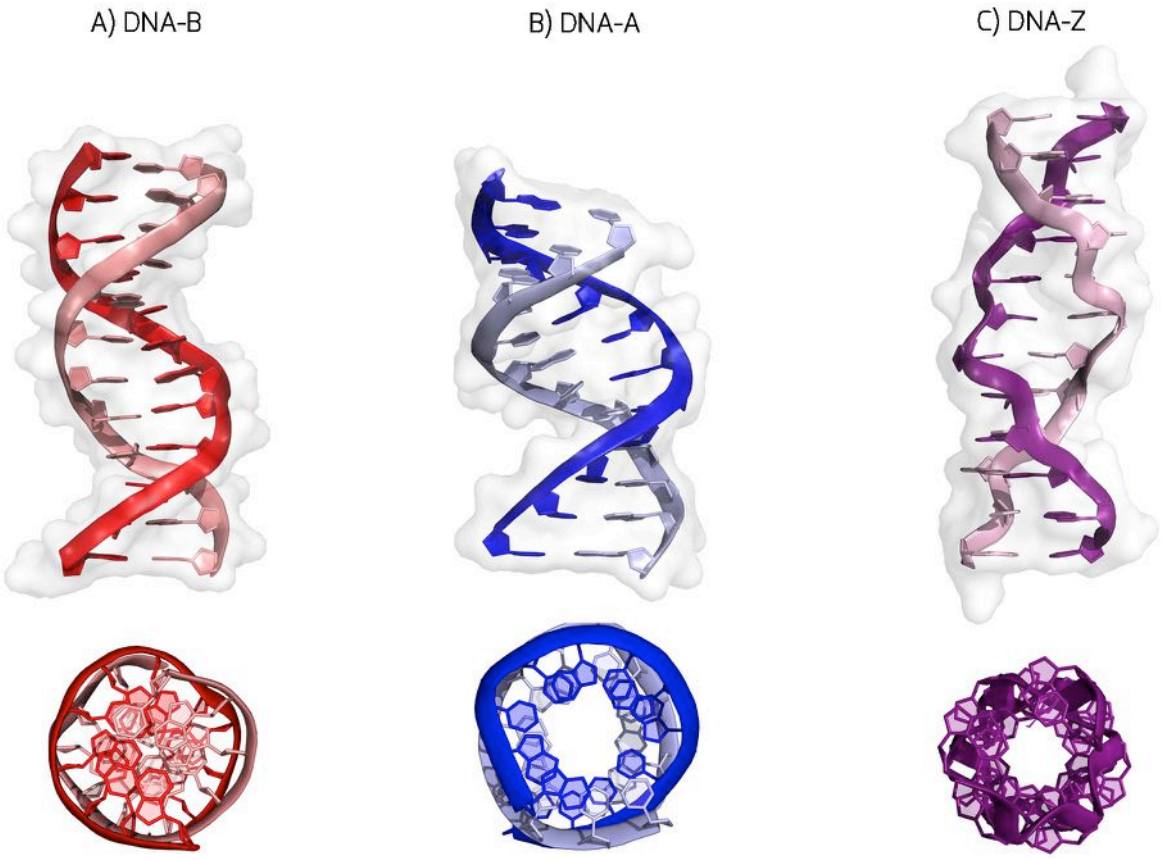


Figura 11-2: Representação dos tipos mais comuns de estrutura 2ária encontrados no DNA, ilustradas para sequências de 12 nucleotídeos. Em vermelho estão as hélices B (A), em azul as hélices A (B) e em magenta as hélices Z (C). As estruturas pelos códigos PDB 3BSE, 3V9D e 279D. Para cada uma duas diferentes orientações são apresentadas, e o esqueleto das moléculas de DNA está representado como fitas.

ma que não há definição específica para um ou alguns tipos de hélices, como vimos anteriormente. Ao invés disto, cada tipo de polissacarídeo apresentará um número de resíduos por volta, elevação por resíduo e elevação por volta, assim como seu sentido para a direita ou para a esquerda (vide tabela 2-3).

Estas características, contudo, são normalmente determinadas experimentalmente através de difração de raios-X, na qual a amostra está na fase cristalina. Esta é uma condição adequada à descrição, por exemplo, da quitina, polissacarídeo encontrado na natureza em condições semelhantes. Contudo, quando estes polissacarídeos são transpostos para soluções biológicas, estas moléculas adotam uma elevada flexibilidade e, por conseguinte, grande variação conformacional. Não raramente, perdemos a capacidade de identificar formas repetitivas, e a denominação de alças desordenadas pode também ser aplicada a polissacarídeos.

Adicionalmente, carboidratos não se apresentam somente como polissacarídeos lineares, mas como oligo- ou polissacarídeos ramificados. Esta ramificação agrega um grau adicional de complexidade na descrição da forma destes compostos. Mesmo assim, ainda é possível descrever a forma destes compostos, caso a caso, como veremos adiante.

Estrutura 3ária

A importância do conhecimento da estrutura 2ária de biomoléculas reside, principalmente, no fato de que estes elementos se organizam no espaço tridimensional, dando

Tabela 2-3: Tipos de hélices encontrados em ácidos nucleicos.

Tipo de hélice	pb / volta	Elevação / pb (Å)	Elevação / volta (Å)	Fenda maior (Å)		Fenda menor (Å)		Direção
				Largura	Profundidade	Largura	Profundidade	
DNA A	11	2,9	32	2,7	13,5	11,0	2,8	direita
DNA B	10	3,4	34	11,7	8,5	5,7	7,5	direita
DNA Z	12	3,8	45	-	convexa	4	9	esquerda

origem ao que chamamos de estrutura 3ária. Em outras palavras, a estrutura 3ária de uma dada biomolécula corresponde à montagem dos seus elementos de estrutura 2ária. Por outro lado, é a estrutura 3ária (ou a 4ária, que veremos a seguir) que irá exercer a função biológica da molécula em questão.

Os diversos elementos de estrutura 2ária de uma dada molécula se organizam em uma estrutura 3ária através de um fenômeno denominado enovelamento (também chamado em português de dobramento, do termo em inglês *folding*). Neste processo, uma combinação de forças converge para que a biomolécula adote uma conformação mais estável no meio biológico alvo.

O termo conformação é usado para descrever a forma de uma dada molécula, como já empregado neste capítulo. Contudo, deve-se adotar uma distinção entre conformação e estrutura, importante para o entendimento de propriedades moleculares. Estrutura se refere a uma única forma, bem definida e conhecida. Conformação se refere a uma forma dentre múltiplas possíveis, em um determinado meio ou ambiente molecular. Assim, é comum nos referirmos a estrutura cristalina de uma dada proteína, pois no cristal temos uma única forma 3D, como uma foto única que compõe um filme. Em solução, contudo, há diversas formas simultaneamente co-existindo. Neste caso, cada forma pode ser denominada de conformação. Podemos, de forma mais precisa, dizer que a forma de uma biomolécula, determinada por cristalografia de raios-X, é uma conformação cristalográfica.

O processo de enovelamento é mais estudado para proteínas, biopolímeros que apresentam uma versatilidade de estrutura 3ária que nenhuma outra biomolécula possui. Isso faz todo o sentido, tendo em vista que são as proteínas os principais efetores da informação gênica. Em proteínas, o enovelamento envolve a aproximação mútua de resíduos hidrofóbicos, que buscam se esconder da água (também chamado de colapso hidrofóbico), ocasionando a expulsão deste solvente da região central da proteína.

Simultaneamente, os resíduos polares são expostos ao solvente, e interações inter-resíduo são estabelecidas. Assim, a estrutura enovelada, nativa, terá uma quantidade mínima de moléculas de água em seu interior e um número máximo de contatos inter-resíduo (Figura 12-2).

A ideia de ambiente molecular para o enovelamento ou para que uma dada biomolécula exerça sua função é mais complexa do que parece à primeira vista. Embora a ideia usual seja de que o meio aquoso seja predominante, diversos tipos de ambientes aquosos podem ser encontrados dentro de um organismo, tecido ou célula. Por exemplo, o pH pode apresentar grandes variações entre vacúolos lisossomais, citoplasma, plasma, secreção gástrica ou duodenal. Por outro lado, a força iônica da solução pode mudar drasticamente na proximidade de membranas com diferentes cargas.

Outro tipo de ambiente molecular que deve ser destacado é definido pelas membranas biológicas. Membranas são fluidos, e moléculas inseridas em membranas estão solvatados pelas moléculas de fosfolipídeos. Assim, sendo o interior de membranas apolar (ou seja, lipofílico), o colapso hidrofóbico pode acontecer ao inverso, com a exposição de resíduos apolares para o solvente (neste caso, a membrana). Ambientes mais específicos para o enovelamento de proteínas podem ainda ser criados por outras proteínas, denominadas chaperonas. Como um barril, chaperonas podem isolar uma proteína do meio aquoso, levando a formação de interações inter-resíduo que não seriam observáveis de forma significativa em sua ausência. Por conseguinte, podem contribuir diretamente na formação de estruturas 3árias.

Além de interações não covalentes entre os resíduos de aminoácidos de uma dada proteína (ou as bases de um ácido nucleico e os monossacarídeos de um polissacarídeo) e destes com o solvente, o enovelamento de

proteínas também é influenciado por intera-

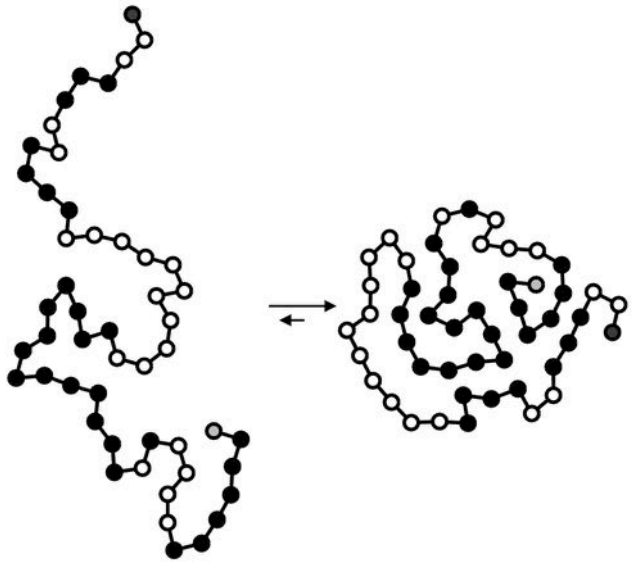


Figura 12-2: Representação 2D do enovelamento de uma proteína hipotética, com o direcionamento de resíduos hidrofóbicos (círculos pretos) para o interior da proteína e dos resíduos hidrofílicos para sua superfície (círculos brancos). Reproduzida de Tomixdf, 2008 (*Creative Commons*).

ções covalentes, associadas a modificações co- ou pós-traducionais.

Durante ou após a síntese proteica (tradução), podem ser formadas ligações dissulfeto entre grupamentos sulfidrila (SH) de resíduos de cisteína, cofatores como o grupamento heme podem ser adicionados ou mesmo processos reversíveis podem ocorrer, nos quais reações como N-acetilação ou fosforilação podem ser observadas de forma transiente. Mas o tipo mais abundante de modificação co- ou pós-traducional na natureza é a glicosilação de proteínas, ou seja, a adição de uma estrutura oligossacarídica a um determinado aminoácido. Assim, a adição destas ligações covalentes e grupamentos altera não somente a forma 3D da proteína, mas sua flexibilidade e múltiplas propriedades fisico-químicas, enzimáticas e, por fim, pode também exercer papel importante em suas funções biológicas.

A glicosilação de proteínas ocorre em mais de 70% das proteínas de eucariotos. Diversos aminoácidos podem estar envolvidos na ligação a carboidratos, mais comumente resíduos de asparagina ou serina, embora também possam participar resíduos de treonina, hidroxiprolina, tirosina, arginina, triptofano e cisteína. Dependendo do aminoácido, a parte sacarídica pode estar ligada a átomos de nitrogênio, oxigênio, carbono ou enxofre, dando origem às glicosilações chamadas de N-, O-, P-, C- ou S-ligadas.

Estrutura 4ária

A despeito da função de um gene ser exercida por uma proteína com estrutura 3D, envolvendo a transmissão de informação de uma estrutura 1ária para uma estrutura 3ária, ainda há um quarto e último nível de organização de biomacromoléculas, denominado de estrutura 4ária. Nem todas as biomoléculas, contudo, apresentam este grau de organização.

A estrutura 4ária é constituída por agregados macromoleculares, principalmente de proteínas. Estas biomoléculas podem adotar estados oligoméricos, sejam estes compostos por 2 (dímeros), 3 (trímeros), 4 (tetrâmeros), 5 (pentâmeros), 6 (hexâmeros) ou mais subunidades necessárias à realização de determinada função em condições nativas. No caso de ácidos nucleicos, a estrutura 4ária também pode ser observada, por exemplo, em complexos entre DNA e proteínas, como histonas.

Não é porque uma proteína se mostra como um oligômero em ambiente cristalino que em solução a mesma organização, necessariamente, será observada. Mesmo *in vivo*, diferentes ambientes fisiológicos podem acarretar em mudanças no estado oligomérico de uma proteína. Por exemplo, um peptídeo que se mostra como monômero no plasma pode formar tetrâmeros quando inserido em membranas.

Portanto, assim como no caso da estrutura 3ária, a estrutura 4ária frequentemente se constitui em uma complexa combinação de múltiplas possibilidades que podem ser modificadas ou reguladas em função de inúmeras variáveis químicas e biológicas. Reproduzir com precisão este comportamento dinâmico é um dos principais desafios para a bioinformática.

2.4. Descritores de forma

O uso dos conceitos de níveis hierár-

quicos nos permite entender as organizações básicas da estrutura 3D de macromoléculas. Estes níveis, contudo, nos oferecem definições qualitativas, gerais, que não abordam nuances ou variações dentro dos níveis. Por exemplo, definir uma região da proteína como uma hélice α não nos informa se esta hélice apresenta ou não algum grau de deformação. Similarmente, podemos saber que uma determinada sequência de nucleotídeos de DNA assume uma hélice do tipo B, mas esta classificação simplemsente não avalia a deformação provocada nesta hélice por um fármaco intercalador do DNA.

Portanto, em acréscimo aos níveis hierárquicos de classificação da estrutura de macromoléculas, há a necessidade de introduzir medidas quantitativas da forma destes compostos. Podemos, assim, calcular precisamente formas associadas a determinados eventos biológicos (como a regulação da expressão de um gene) e, por conseguinte, interferir nestes processos de forma racional (como no desenho de novos fármacos capazes de inibirem a expressão deste gene).

Considerando que proteínas, carboidratos e ácidos nucleicos são biopolímeros, suas formas tridimensionais são definidas, basicamente, pelas conectividades entre seus monômeros constituintes (isto é, aminoácidos, monossacarídeos e bases nitrogenadas, respectivamente).

Esta forma de compreender a estrutura de biomacromoléculas foi proposta inicialmente em 1963 por Gopalasamudram Narayan Ramachandran. Neste trabalho, G. N. Ramachandran descreve a forma de dois aminoácidos vizinhos como fruto dos ângulos de torção ao redor do Cα (Figura 13-2), denominados φ e ψ. Assim, em função das cadeias laterais de cada aminoácido, algumas combinações de ângulos φ e ψ seriam favorecidas, enquanto outras proibidas. As combinações favorecidas correspondem às estruturas 2^{árias} de proteínas que nós conhecemos e oferecem, assim, uma medida quantitativa para definir hélices, fitas, alças e voltas. O gráfico que combina os valores de ângulos φ e ψ para um determinado dipeptídeo ficou assim sendo conhecido como mapa de Ramachandran (Figura 13-2).

O uso de ângulos de torção para descrever a estrutura e a conformação molecular não se limita somente a proteínas, mas também pode ser aplicado a ácidos nucleicos e carboidratos. Em cada caso, o número de ângulos de torção é definido pelas características das ligações entre os monômeros, isto é, se é uma ligação peptídica, glicosídica ou fosfodiéster.

Para a descrição da forma de uma ligação peptídica em uma proteína são empregados três ângulos: ω, ψ e φ. Os ângulos ψ e φ são aqueles descritos no mapa de Ramachandran, localizando-se antes e depois do Cα (porções N- e C- terminais da ligação, respectivamente). O ângulo ω, por sua vez, corresponde ao grupamento amida, ou seja, a ligação entre os grupamentos N-H e C=O (Figura 14-2).

A ligação glicosídica pode ser descrita por dois ou três ângulos torcionais. Em analogia à ligação peptídica, podem ser empregados os ângulos φ e ψ (porção não-redutora e porção redutora, respectivamente). A exceção é quando descrevem-se ligações envolvendo o átomo de carbono na posição 6 de piranoses (como glicose, manose, fucose e etc.) e na posição 5 de furanoses (como na ribose e na desoxirribose). Nestes casos, há a necessidade de se considerar um terceiro ângulo torsional, denominado ω.

O terceiro caso de biopolímeros usualmente descritos por ângulos torcionais, os ácidos nucleicos, consistem em um caso à parte. Como podemos observar na Figura 14-2, o grupamento fosfato agrega grande flexibilidade à cadeia, exigindo assim sete ângulos torsionais para sua adequada caracterização, a saber: α, β, γ (na região 5'), δ (entre os átomos 3' e 4' da pentose), ε e ζ (na porção 3'). Há, ainda, o ângulo χ, formado entre o carbono 1' da pentose e a base nitrogenada.

Ângulos torsionais não são, contudo, a única forma de descrever e avaliar a forma de biomacromoléculas. A despeito de serem biopolímeros, proteínas, carboidratos e ácidos nucleicos apresentam suas particularidades, exigindo assim descritores específicos, capazes de lidar com as propriedades fisico-químicas particulares de cada tipo de monômero (e, por conseguinte, em lidar com as diferentes propriedades biológicas resultantes).

Como mencionado anteriormente, biomoléculas em condições biológicas apresentam não somente uma, mas múltiplas conformações que coexistem, simulta-

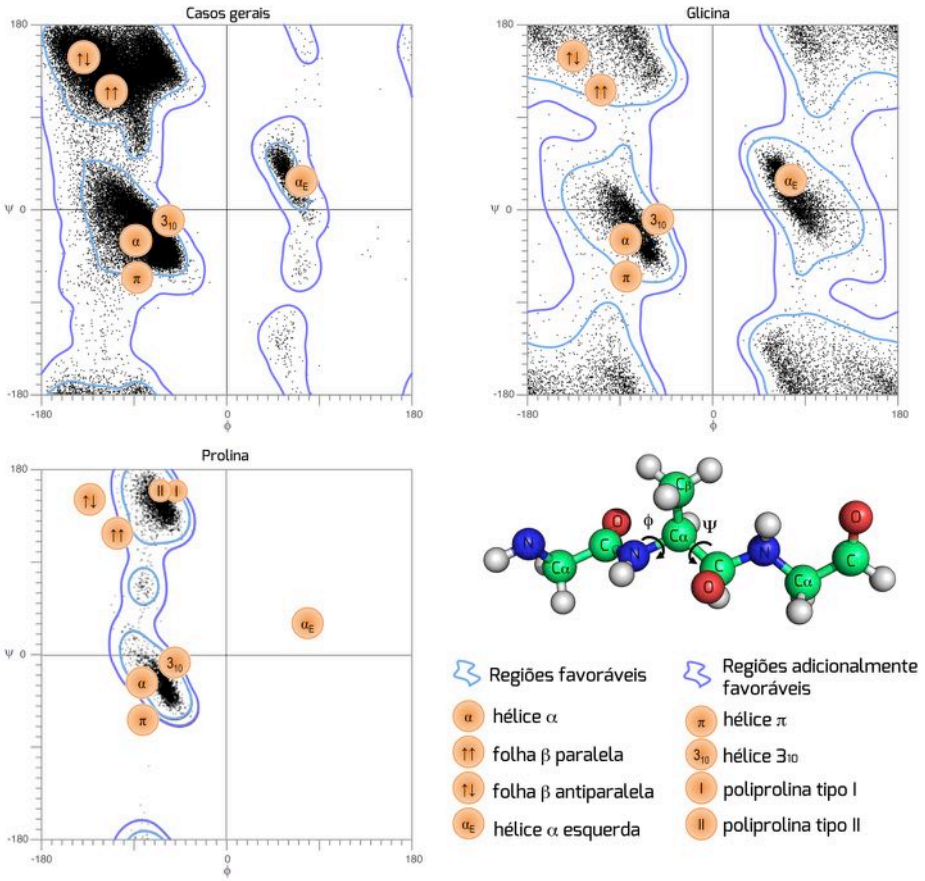


Figura 13-2: Mapas de Ramachandran para casos gerais (resíduos que não sejam prolina ou glicina), para resíduos de glicina e para resíduos de prolina. Os pontos correspondem às distribuições de ângulos φ e ψ de cerca de 100 mil resíduos componentes de 500 estruturas proteicas obtidas em alta resolução. As regiões onde se localizam as estruturas secundárias típicas estão destacadas nos mapas. [Figura baseada em LOVELL, Simon C. *et al*. Structure Validation by Cα Geometry: φ, ψ and Cβ Deviation. Proteins, 50, 437-450, 2003; e Hollingsworth, Scott A. & Karplus, P. Andrew. *A fresh look at the Ramachandran plot and the occurrence of standard structures in proteins. Biomol. Concepts*, 1, 271–283, 2010].

neamente. Assim, os valores de ângulos torsionais devem ser considerados como médias, referências geométricas em torno das quais o comportamento da molécula em questão irá variar em solução.

Ácidos nucleicos

Em acréscimo aos ângulos torcionais os ácidos nucleicos, ao formarem pares de bases, definem quase duas dezenas de parâmetros geométricos distintos, importantes para uma cartacterização precisa da estrutura destas biomoléculas (Figura 15-2). Isto ocorre em decorrência de movimentos de translação ou rotação que cada base ou par de bases pode sofrer dentro da região pareada. Assim, moléculas ou regiões de ácidos nucleicos não pareadas não são descritas por estes parâmetros.

Considerando um espaço cartesiano definido pelos eixos x, y e z, sendo z o eixo maior da região de pareamento e bases (Figura 15-2), os parâmetros geométricos oriundos da translação de bases em uma dupla fita envolvem: *i*) o deslocamento do par de bases ao longo do eixo x ou do eixo y; *ii*) o deslocamento de uma base em relação à outra, seja como uma distensão ao longo do eixo y (do inglês *stretch*), seja como cisalhamento ao longo do eixo x (do inglês *shear*), ou ainda um escalonamento acima ou abaixo do plano xy (do inglês *stagger*); *iii*) o deslocamento de um par de base em relação a outro par de base, seja como uma elevação ao longo do eixo z (do inglês *rise*), seja como um deslizamento ao longo do eixo y (do inglês *slide*) ou ao longo do eixo x (chamada em inglês de *shift*).

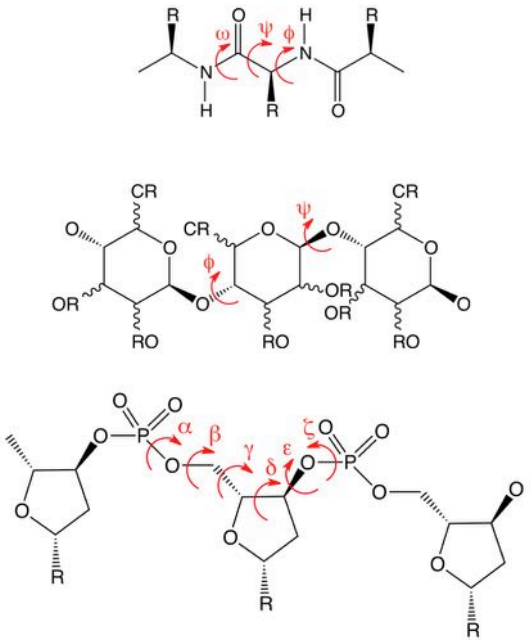


Figura 14-2: Ângulos torsionais para proteínas, carboidratos e ácidos nucleicos ilustrados para, respectivamente, um tripeptídeo, um trissacarídeo e um trinucleotídeo.

Os parâmetros originados da rotação de bases ou pares de bases entre si produzem diferentes tipos de inclinação (definidas em inglês como *tip*, *inclination*, *roll* e *tilt*), dependendo do vértice e do eixo ao longo dos quais ocorre o movimento do par de bases. Pares de bases podem ainda sofrer modificações caracterizando-os como: *i*) torcidos (chamadas em inglês de *twist*, *propeller twist* ou *buckle*), e *ii*) abertos (definida em inglês como *opening*).

Proteínas

Considerando os 20 aminoácidos codificados no genoma, poderíamos imaginar que teríamos 20^n possíveis proteínas diferentes, sendo n o número de aminoácidos. A situação, felizmente, não é tão complexa por uma série de motivos.

Um primeiro aspecto a ser observado é que, quando uma sequência de aminoácidos se enovela para adotar uma determinada estrutura 3^{ária}, alguns aminoácidos se localizam em pontos chave para a estabilização da estrutura 3D. Assim, sua modificação poderia desestabilizar total ou parcialmente a conformação nativa da proteína. Como consequência, algumas posições na sequência de aminoácidos tornam-se conservadas evolutivamente como decorrência de determinantes estruturais. Ao mesmo tempo, podem haver determinantes funcionais para a conservação de posições na sequência ao longo da evolução.

Em contrapartida, como os aminoácidos podem ser agrupados de acordo com a semelhança em suas propriedades fisico-químicas, diferentes combinações de resíduos podem levar a uma mesma estrutura 3D. De fato, sabe-se que a estrutura 3^{ária} de proteínas é mais conservada ao longo da evolução que a estrutura 1^{ária}. Em outras palavras, proteínas com identidade muito baixa entre suas sequências podem possuir estruturas 3^{árias} muito semelhantes.

Conclui-se, assim, que sequências de aminoácidos podem arranjar-se em um conjunto de formas 3D mais ou menos definidos e finitos. Estas formas são denominadas motivos (ou no inglês *fold*), e possuem diversas classificações a partir de suas características (Figura 16-2). Dada a relação entre forma e função, o conhecimento do motivo de uma dada proteína (diretamente por métodos experimentais como cristalografia de raios-X, ver capítulo 13, ou por inferência a partir de similaridade de sequência, ver capítulo 3) é um passo importante para a elucidação de seu mecanismo de ação em nível molecular.

Por exemplo, um barril-β é um motivo que se assemelha a um barri, onde as tiras de madeira correspondem a fitas β (Figura 16-2). Define, assim, uma cavidade central que pode tanto servir como carreador de substâncias, como no caso das nitroforinas, ou como poro, como no caso das porinas. Embora o número de fitas β possa mudar (8 no caso das nitroforinas e 16 no caso das porinas), a característica geral do motivo se mantém. Essas relações são ilustradas visualmente de forma muito elegante na "tabela periódica" de proteínas, desenvolvida pelos professores Richard Garratt e Christine Orengo. Para acessar as classificações dos diferentes motivos já identificados, os bancos de dados CATH e SCOP são as fontes mais completas

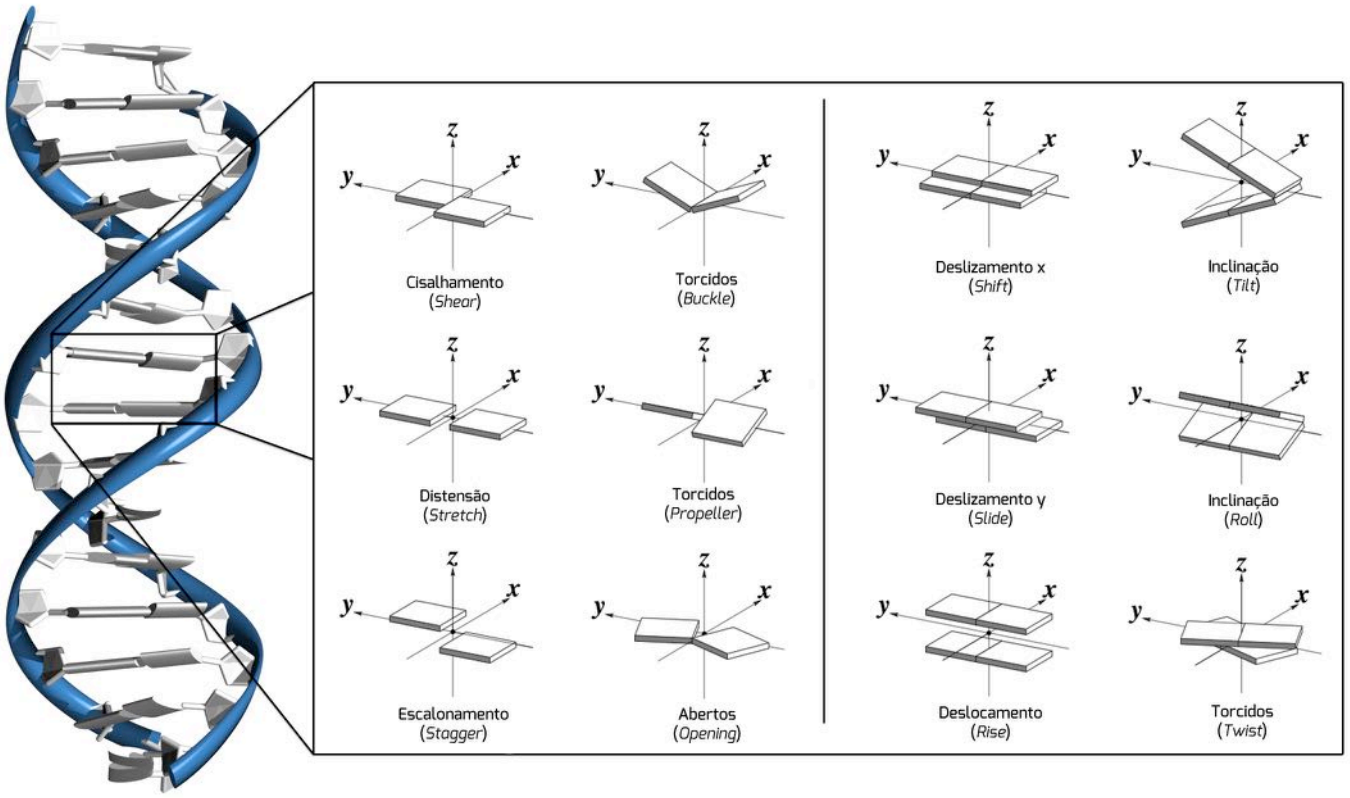


Figura 15-2: Parâmetros geométricos empregados como descritores da geometria de ácidos nucleicos.

de informações.

Um outro conceito, que se confunde e em vários momentos é usado como sinônimo de motivo, é o de domínio proteico. Um domínio é uma parte da sequência polipeptídica de enovelamento independente (e, potencialmente, de função também independente). Assim, se um domínio for recortado de um gene e expresso separadamente ele deve, em princípio, manter suas características estruturais.

Um domínio proteico pode ser composto por mais de um motivo intrinsecamente associado. Por outro lado, um mesmo motivo pode ser encontrado e mais de um domínio de uma mesma proteína.

Membranas

Não temos falado muito de membranas até este momento por alguns motivos. Primeiramente, membranas não são biopolímeros, mas agregados de múltiplas moléculas, o que tira de cena a ideia de análise de uma molécula a partir de suas sub-unidades formadoras. Segundo, estes agregados apresentam-se como um fluido, diferentemente das outras biomoléculas que vimos. Assim, não faz sentido analisar cada molécula de lipídeo individualmente em uma membrana, mas o seu comportamento como um todo ou como uma média ao longo de múltiplos lipídeos.

Contudo, a despeito da natureza fluida de membranas e da sua capacidade de adotar múltiplas formas, os lipídeos (e também proteínas) não se distribuem homogeneamente ao longo das membranas, podendo formar regiões ou domínios enriquecidos em um determinado componente. Assim, para o estudo das propriedades de membranas biológicas torna-se necessário caracterizá-las estruturalmente. Isto pode ser feito através de diversas medidas, tais como a área por lipídeo, espessura da membrana e coeficientes de difusão lateral de lipídeos ou proteínas embebidas na membrana, dentre outros (Figura

8-2).

A área por lipídeo nos oferece informações acerca do grau de compactação das moléculas que constituem uma membrana, ou seja, uma área menor indica uma membrana mais compacta. Isto, por sua vez, sugere uma interação mais intensa entre os componentes da membrana.

Embora proteínas inseridas em membranas adaptem-se a este meio, são as membranas que fazem a maior parte do ajuste em sua estrutura para receber as proteínas (esse processo está relacionado às diferenças de compressibilidade entre estas biomoléculas). Como consequência, a inserção de proteínas em membranas biológicas promove uma perturbação na organização da bicamada lipídica, podendo tanto aumentar quanto reduzir a espessura desta na região ao redor da

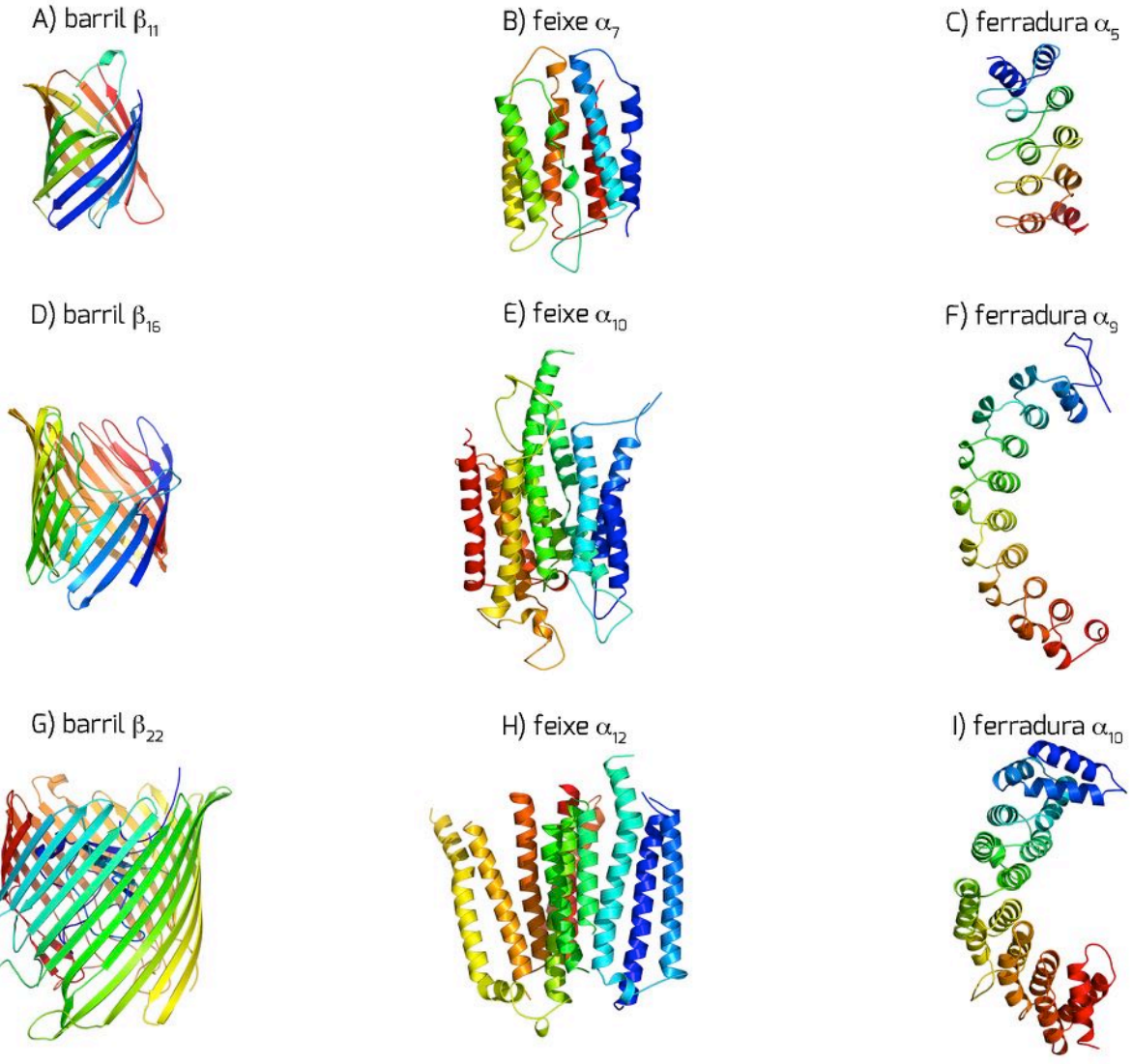


Figura 16-2: Exemplos de motivos proteicos, coloridos por cada elemento de estrutura 2ária. São apresentados barris compostos por fitas-β, em A a proteína verde fluorescente (do inglês *green fluorescent protein*, GFP, código PDB 1EMG), em D a porina OMP32 (código PDB 2FGQ) e em G o transportador FECA (código PDB 1KMO); feixes de hélices α, em B a bacteriorodopsina (código PDB 1AP9), em E a proteína SERCA1 (código PDB 1WPG) e em H parte do sistema fotossintético de uma cianobactéria (código PDB 1JB0); e ferraduras compostas por hélices α, em C um inibidor de crescimento tumoral (código PDB 1BD8), em F uma repetição rica em resíduos de leucina, associada à fixação de nitrogênio (código PDB 1LRV) e em H a lipovitelina (código PDB 1LSH). Partes das estruturas foram omitidas buscando maior clareza na imagem. Imagem construída usando o programa Pymol, a partir de organização proposta em "The Protein Chart", de Richard C. Garratt e Christine A. Orengo, 2008, Wiley-VCH.

proteína.

2.5. Formas de visualização

O corolário *uma imagem fala mais do que mil palavras* também se aplica ao estudo de moléculas. E, de fato, o desafio de representar graficamente proteínas vem acompanhando os pesquisadores desde o início dos estudos da estrutura destas moléculas. Os primeiros relatos do uso de representações em cartoon para proteínas datam da década de 1960. Atualmente, múltiplas representações estão à nossa disposição, com qualidade gráfica a cada momento superior, e gerados através de ferramentas gratuitas (Figura 17-2).

Podemos definir hélices de proteínas por suas características geométricas, nomes ou pelos pares de ângulos φ e ψ. Mas visualizar uma hélice proteica, tridimensionalmente, não deixa dúvidas quanto ao seu significado. Portanto, o cuidado com a maneira pela qual iremos apresentar, visualmente, os aspectos estruturais que estudamos e tenhamos relacionados a alguma função biológica, é uma parte fundamental no trabalho do bioinformata.

Formas de visualização, contudo, são representações muitas vezes incapazes de descreverem detalhes sobre a molécula em estudo. É difícil distinguir visualmente uma hélice α de uma hélice 3_{10} ou de uma hélice π. Por outro lado, estas hélices podem apresentar deformações importantes, também de difícil visualização. Assim, a combinação de representações visuais, qualitativas, com medidas precisas, quantitativas, da estrutra molecular é uma estratégia bastante útil no estudo de macromoléculas.

A ideia de combinar múltiplas estratégias na apresentação de um determinado aspecto molecular não se limita somente às formas de descrever visualmente ou numericamente a estrutura molecular. Embora a visualização de estruturas 1árias, isto é, de sequências de nucleotídeos, aminoácidos ou monossacarídeos não nos ofereça muitos artifícios visuais, devemos nos lembrar que as formas apresentadas na Figura 17-2 não informam o leitor facilmente sobre quais resíduos compõe a nossa macromolécula. É difícil distinguir, em representações de arames, bastões ou esferas, uma Ile de uma Leu, e mesmo impossível em cartoon ou superfície. Portanto, pode ser muito útil combinar estas representações tridimensionais a alinhamentos de sequências da região de interesse.

O mesmo vale para a apresentação de sequências isoladas de estruturas. Enquanto uma mutação em um único nucleotídeo pode interferir na função proteica, isso não é feito pela troca de uma letra por outra na sequência, mas por mudanças que esta troca acarretam na estrutura da proteína. O entendimento deste processo pode depender simplesmente da nossa imaginação ou da visualização da respectiva mudança na proteína.

Existem diversas formas de apresentar estruturas tridimensionais de macromoléculas, e escolher entre estas formas envolve tanto escolhas metodológicas quanto pessoais. Algumas propriedades são mais facilmente observadas em alguns tipos de visualização. Por exemplo, o volume da cadeia lateral de um resíduo de Val é muito mais facilmente observável enquanto seus átomos são apresentados como esferas do que como bastões ou arames (Figura 17-2). Diferentes tipos de moléculas, similarmente, se beneficiam de algumas formas de visualização. Por exemplo, a forma de cartoon é a mais comum para descrever proteínas, mas é pouco útil na

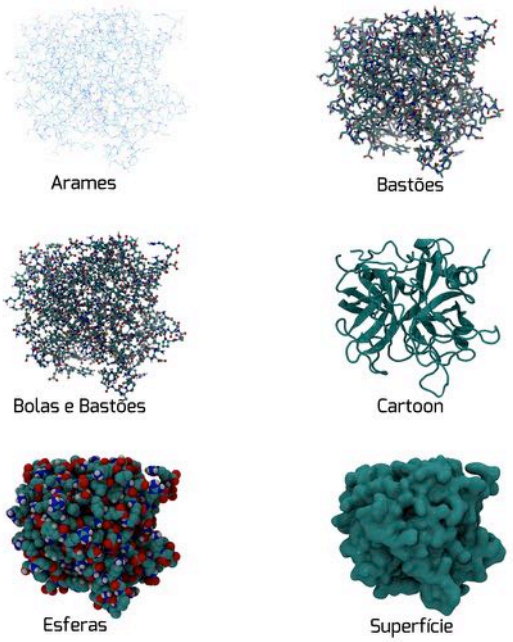


Figura 17-2: Exemplo das formas de visualização mais comumente empregadas na descrição de biomoléculas, aplicadas a uma proteína.

descrição de carboidratos ou membranas.

Em muitos casos poderemos empregar combinações destas formas, como na descrição por cartoon de uma proteína e de sua estrutura de glicosilação como bastões.

2.6. Conceitos-chave

Anfipatia: propriedade de moléculas que possuem tanto regiões hidrofílicas quanto hidrofóbicas.

Cadeia lateral: região variável dos aminoácidos codificados no genoma, responsável pela variação de suas propriedades.

Carbono anomérico: átomo de carbono numerado como 1 em carboidratos. A mudança em sua estereoquímica dá origem às formas anoméricas α e β em carboidratos.

Carbono α: átomo de carbono do esqueleto peptídico no qual a cadeia lateral de cada aminoácido está ligada (referindo-se aos 20 aminoácidos codificados no genoma para síntese proteica). É o primeiro átomo de carbono vizinho ao grupo carbonila.

Conformação em bote torcido: forma adotada pelo anel de alguns monossacarídeos.

Conformação em cadeira: forma adotada pelo anel de alguns monossacarídeos, semelhante a uma cadeira quanto vista de lado.

Conformação em envelope: forma adotada pelo anel de alguns monossacarídeos, destacadamente as furanoses.

Dogma central da biologia molecular: representação do fluxo de informação em sistemas biológicos, começando na molécula de DNA e culminando na síntese proteica - mas não no sentido oposto. Envolve principalmente os fenômenos de replicação, transcrição e tradução.

Enovelamento: processo segundo o qual uma sequência polipeptídica adquire sua estrutura tridimensional nativa, isto é, equivalente àquela observada em seu local biológico de ação e funcional. Também chamado por alguns autores de dobramento.

Equilíbrio pseudo-rotacional: processo de interconversão entre as diferentes conformações adotadas por carboidratos.

Esqueleto do DNA: parte da molécula de DNA composta pelas partes comuns a todos os nucleotídeos, isto é, o carboidrato e o grupo fosfato (ou seja, são excluídas as regiões das bases nitrogenadas).

Esqueleto peptídico: estrutura de peptídeos ou proteínas sem as cadeias laterais dos aminoácidos (ou seja, somente as regiões comuns aos aminoácidos).

Estrutura 1$^{\text{ária}}$: sequência de letras que compõe biomoléculas (principalmente DNA, RNA e proteínas, mas também carboidratos).

Estrutura 2$^{\text{ária}}$: padrões estruturais definidos pela organização das unidades monoméricas (isto é, nucleotídeos, aminoácidos e monossacarídeos) de cada biomolécula em formas tridimensionais. Estes padrões podem classificados segundo suas diferentes formas.

Estrutura 3$^{\text{ária}}$: estrutura 3D completamente enovelada.

Estrutura 4$^{\text{ária}}$: organização definida pela agregação de múltiplas estruturas 3$^{\text{árias}}$.

Furanoses: monossacarídeos cujo anel é composto por 5 átomos, quatro de carbono e um de oxigênio. O nome vem da semelhança deste anel com o composto furano.

Ligação fosfodiéster: ligação formada entre dois nucleotídeos, através de seus grupos fosfato.

Ligação glicosídica: ligação formada entre dois

monossacarídeos.

Ligação peptídica: ligação formada entre dois aminoácidos, através do grupo amino de um resíduo e do grupo carboxila do outro, dando origem a uma função amida.

Mapa de Ramachandran: um gráfico que descreve a variação da energia em função da rotação dos ângulos de diedro φ e ψ, ao redor do Cα.

Nucleosídeo: molécula formada por uma base nitrogenada ligada a um carboidrato (ribose ou desoxirribose), sem o grupo fosfato.

Nucleotídeo: molécula formada por uma base nitrogenada ligada a um carboidrato (ribose ou desoxirribose) e a um grupo fosfato.

Piranoses: monossacarídeos cujo anel é composto por 6 átomos, cinco de carbono e um de oxigênio. O nome vem da semelhança deste anel com o composto pirano.

2.7. Leitura recomendada

ALBERTS, Bruce; et al. ***Biologia Molecular da Célula***. 5.ed. Porto Alegre: Artmed, 2010.

BLOOMFIELD, Victor A.; CROTHERS, Donald M.; TINOCO, JR., Ignacio. ***Nucleic Acids Structure, Properties, and Functions***. Sausalito: University Science Books, 2000.

GARRATT, Richard C., ORENGO, Christine A. ***The Protein Chart***. Nova Iorque: Wiley-VCH, 2008.

PETSKO, Gregory A.; RINGE, D. ***Protein Structure and Function***. New York: Oxford University Press, 2009.