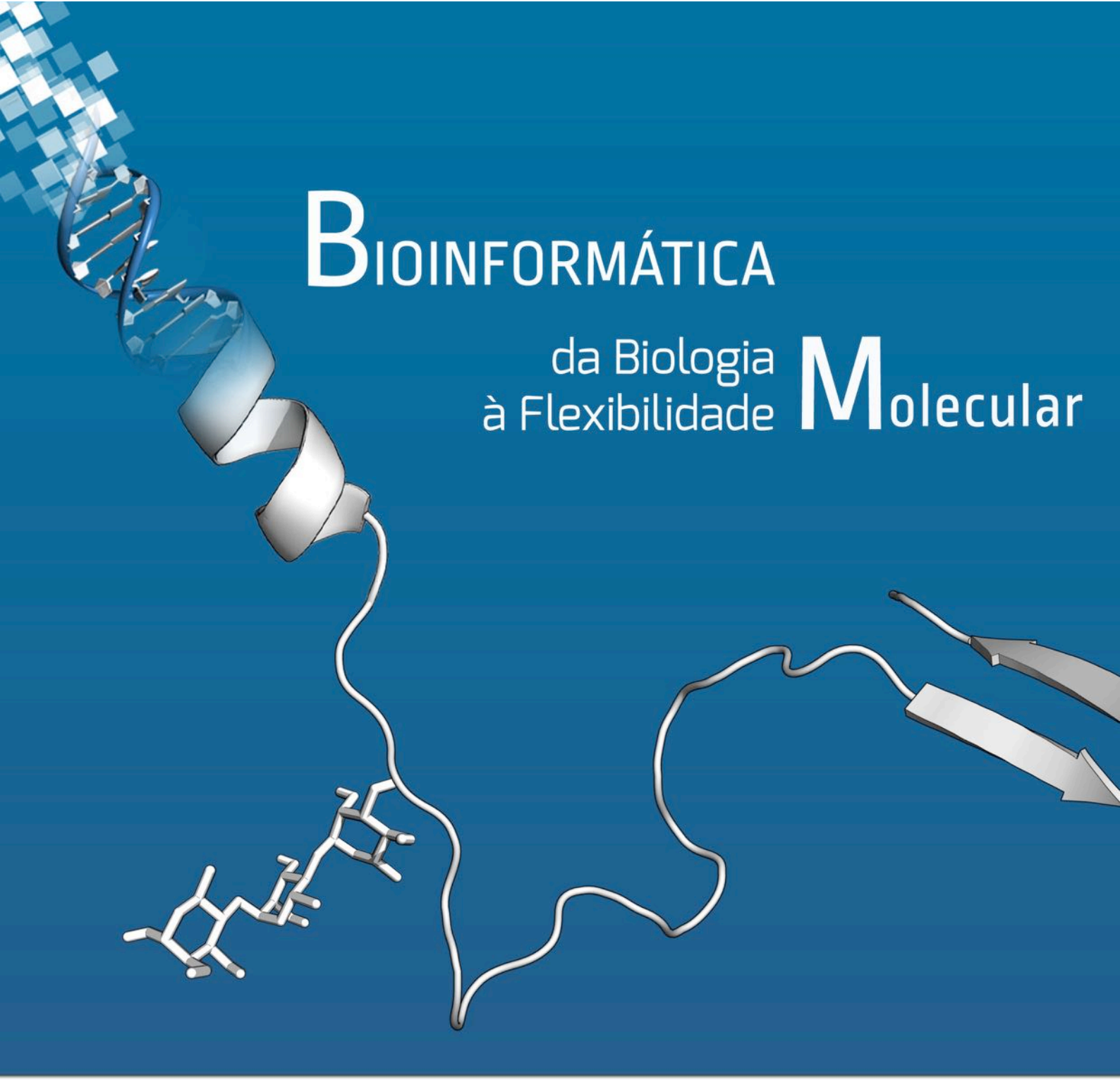


Bioinformática
da Biologia à Flexibilidade Molecular

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

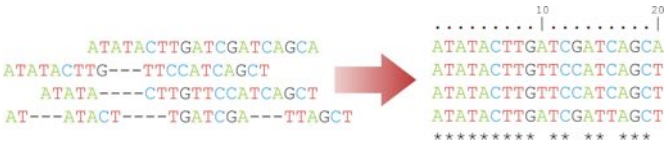
Ficha catalográfica elaborada por Rosalia Pomar Camargo CRB 856/10

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

3. Alinhamentos



Alinhamento de múltiplas sequências.

Dennis Maletich Junqueira
Rodrigo Ligabue Braun
Hugo Verli

3.1. Introdução

O avanço nas técnicas de sequenciamento do DNA tem permitido um crescente aumento no número de genomas disponíveis em bancos de dados públicos. Esta maior disponibilidade exigiu um grande aumento na capacidade computacional de armazenamento e no investimento em desenvolvimento de técnicas de processamento adequadas para a análise destes dados. Algoritmos de análise tiveram de ser criados e aperfeiçoados e, dentre estes, as técnicas de alinhamento de sequências tornaram-se ferramentas essenciais e primordiais na análise de sequências biológicas. Atualmente, diversos programas *online*, ou mesmo de instalação local, são capazes de alinhar centenas de sequências em poucos minutos.

Devido à extensão de suas aplicações, o alinhamento de sequências biológicas é um processo de fundamental importância para a bioinformática. Conceitualmente, os alinhamentos são técnicas de comparação entre duas ou mais sequências biológicas, que buscam séries de caracteres individuais que se encontram na mesma ordem nas sequências analisadas.

Em geral, as moléculas consideradas por estes programas, sejam elas formadas por nucleotídeos (DNA ou RNA) ou aminoácidos (peptídeos e proteínas), são polímeros representados por uma série de caracteres, e a comparação entre as moléculas depende apenas da comparação entre as respectivas letras. Apesar da facilidade e da aparente simplicidade do processo, a análise de similaridade das sequências é uma tarefa complexa e uma etapa decisiva para grande parte dos métodos de bioinformática que fazem uso de sequências biológicas.

Durante o alinhamento, as sequências são organizadas em linhas e os caracteres biológicos integram as colunas do alinhamento (Figura 1-3). Seguido à organização inicial, algoritmos específicos buscarão a melhor correspondência para as sequências em questão, permitindo a criação de espaços entre estes caracteres para que, ao final, todas as sequências tenham o mesmo comprimento. Isto possibilita uma fácil visualização da similaridade, permitindo que caracteres

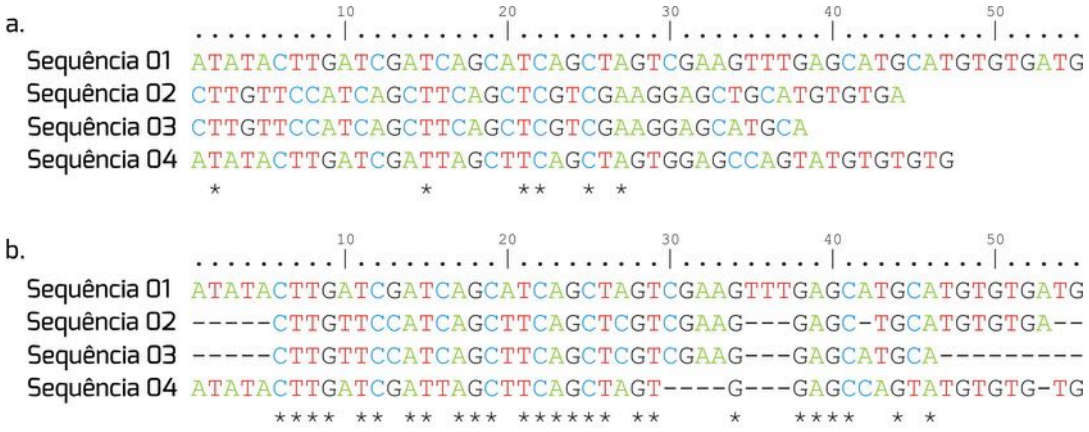


Figura 1-3: Alinhamento de quatro sequências de nucleotídeos envolvendo 55 caracteres. *a*) Grupo de sequências não alinhadas, cada sequência ocupando uma linha individual. *b*) Grupo de sequências alinhadas, onde caracteres idênticos são dispostos em uma mesma coluna e estas são identificadas por asteriscos (dispostos na parte inferior do alinhamento). Nucleotídeos ausentes em determinadas sequências são substituídos por hifens para identificar eventos de inserção/deleção.

idênticos ou similares em cada uma das sequências integrem a mesma coluna. A ideia central destes algoritmos é minimizar as diferenças entre as sequências, buscando um alinhamento ótimo. Comumente, a similaridade entre as sequências envolvidas é expressa pelo termo identidade, que quantifica a porcentagem de caracteres idênticos entre duas sequências.

A relevância e abrangência do uso do método tornam os procedimentos de alinhamento o cerne para diferentes campos dentro da grande área da bioinformática. Além de fundamentais em pesquisas de filogenética e análise evolutiva, os alinhamentos são exigidos em estudos de inferência estrutural e funcional de proteínas, análises de similaridade e identificação de sequências e em estudos aplicados ao campo da genômica.

Através dos métodos de alinhamento, é possível obter informações a respeito da relação evolutiva entre organismos, indivíduos, genes ou entre sequências diversas (Figura 2a-3). Se duas sequências distintas podem ser alinhadas com certo grau de similaridade, é possível inicialmente assumir que elas compartilharam, em algum momento do tempo passado, um ancestral comum e, por isso, são evolutivamente relacionadas. A partir da separação destas sequências de seu ancestral comum, individualmente cada uma delas acumulou diferentes variações ao longo do processo evolutivo. O termo homologia é utilizado frequentemente para definir estes eventos onde, através da relação de ancestralidade, dois indivíduos distintos possuem regiões em seu DNA (incluindo regiões codificantes) herdadas de um ancestral comum. Neste caso, a similaridade deve-se à descendência comum e, portanto, as sequências envolvidas na análise são ditas homólogas.

Cabe ressaltar que a homologia não requer necessariamente alta identidade de caracteres entre as sequências, uma vez que a maior ou menor identidade entre elas dependerá da taxa de evolução do organismo ou da espécie (consultar capítulo 5). Ainda, a similaridade entre sequências pode ser gerada não somente por descendência, mas por pressão seletiva de um determinado ambiente. Nestes casos, teremos regiões similares na sequência de nucleotídeos (ou aminoácidos) que surgiram de maneira independente, sem qualquer relação de descendência, e evoluíram por convergência, não sendo portanto homólogas. Assim, não é possível quantificar a homologia entre as sequências envolvidas, somente dizer se há ou não. Quando identificamos quantos caracteres se repetem nas mesmas posições entre duas ou mais sequências estamos, de fato, verificando a identidade entre estas, e não a homologia.

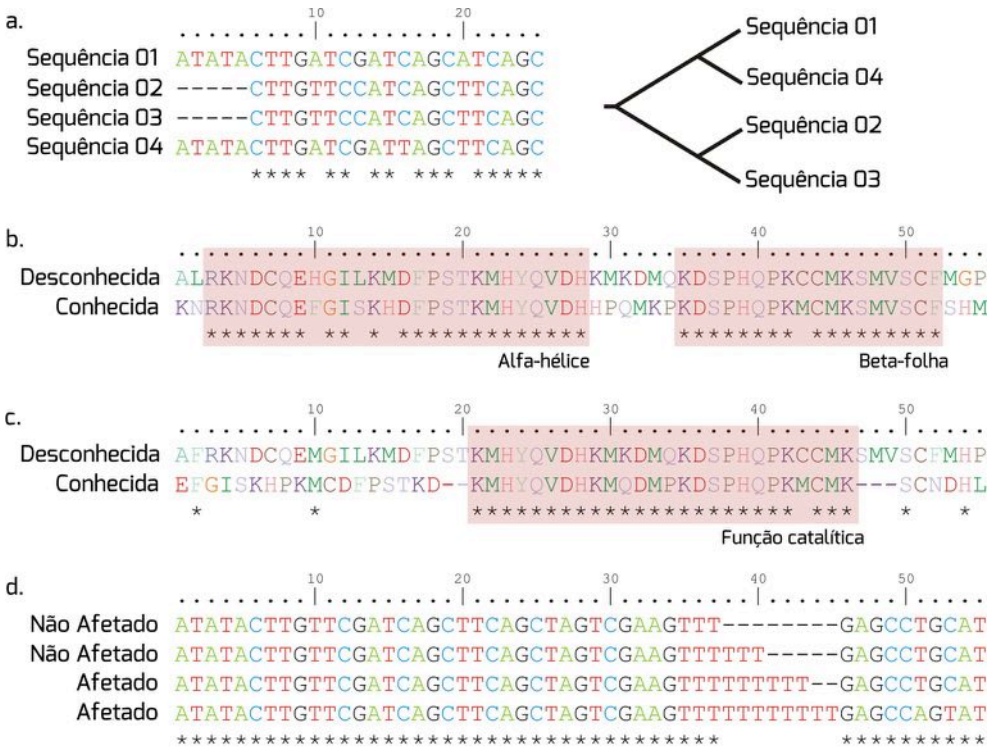


Figura 2-3: Aplicações dos métodos de alinhamento de sequências biológicas. *a*) Inferência filogenética a partir do alinhamento de quatro sequências de nucleotídeos. *b*) Inferência da estrutura de uma proteína alvo (Desconhecida) a partir do alinhamento com uma sequência de aminoácidos cuja estrutura tridimensional é conhecida (Conhecida). *c*) Inferência da função de um domínio proteico a partir da comparação de sequências de aminoácidos. *d*) Comparação de sequências de uma porção de determinado gene de indivíduos afetados e não afetados por uma doença genética. Os asteriscos identificam colunas com total similaridade dos caracteres.

As técnicas de alinhamento vêm se mostrando fundamentais na construção de algoritmos que visam comparar a informação de diversas sequências biológicas. À exemplo do programa BLAST, estes algoritmos permitem comparar uma sequência alvo com milhares de dados disponíveis em grandes bancos de armazenamento, fornecendo um valor de significância estatística associada a esta comparação de similaridade. Devido à facilidade de acesso e rapidez no processamento de dados, estes programas vêm cada vez mais ampliando as possibilidades e opções para o tipo de comparação ou pesquisa a ser realizada.

Os métodos de alinhamento podem ainda ser necessários para fornecer informações a respeito da função e da estrutura de sequências biológicas, particularmente nos alinhamentos de ribonucleotídeos e aminoácidos (Figura 2-3). Nestes casos, a similaridade entre duas ou mais sequências (dada em porcentagem) revela padrões referentes à composição química e podem fornecer embasamento para a definição de um arranjo tridimensional semelhante, principalmente no caso de proteínas (Figura 2b-3). A mesma relação é feita para inferir a função de domínios de uma proteína recém-descoberta, ainda sem função definida. Sabendo que sua forma está diretamente relacionada à sua função, através da comparação com outras proteínas com estrutura e função já estabelecidas, é possível inferir a função realizada por determinado domínio da proteína sob investigação (Figura 2c-3). Nestes casos, as sequências envolvidas no alinhamento não são necessariamente homólogas. Através do fenômeno da evolução convergente, diferentes regiões codificantes do DNA podem gerar produtos proteicos com funções similares, sem obrigatoriamente compartilharem um ancestral comum.

Finalmente, as técnicas de alinhamento

têm grande importância para a análise de genes e genomas. Com o aumento da disponibilidade de sequências nucleotídicas de genomas completos, e mesmo com o surgimento de modernas técnicas de biologia molecular, como o *microarray* e *deep sequencing*, os métodos de comparação permitiram o entendimento a respeito da variabilidade genética de indivíduos e populações.

A comparação entre genomas de diferentes espécies, ou até mesmo de indivíduos da mesma espécie, possibilita a análise de variações (mutações ou polimorfismos) nas sequências e, em alguns casos, permite a identificação de relações entre variações no DNA e susceptibilidade a determinadas doenças, beneficiando o campo da genética e áreas relacionadas. Adicionalmente, como um recurso para a caracterização de eventos evolutivos, os alinhamentos permitem análises comparativas entre genomas. A abrangência e importância evolutiva dos eventos de quebra e reparo de DNA, ou mesmo dos eventos de recombinação, inversões e translocações, tem sido desvendados, primariamente, através dos métodos de alinhamento.

Além do alinhamento de sequências, o alinhamento de estruturas constitui outra importante ferramenta em estudos de bioinformática. A metodologia é bastante diferente daquela empregada em alinhamentos de sequências, pois passamos de um problema unidimensional para um problema tridimensional. Sua utilização passou a ser difundida a partir de 1978, com o trabalho de Rossmann e Argos, comparando os sítios ativos de enzimas cujas estruturas eram conhecidas até aquele momento. Os métodos de sobreposição simples de estruturas estão disponíveis há mais tempo, tendo sido propostos a partir da década de 1970, enquanto os métodos de comparação e alinhamento se desenvolveram posteriormente, principalmente a partir da década de 1990.

A comparação de estruturas se refere à análise de similaridades e diferenças entre duas ou mais estruturas, enquanto o alinhamento de estruturas se refere à determinação de quais aminoácidos seriam equivalentes entre tais estruturas. É importante destacar também a diferença entre alinhamento e sobreposição de estruturas. Apesar desses termos ainda serem empregados na literatura como sinônimos, eles se referem a procedimentos diferentes. Conforme mencionado acima, enquanto o alinhamento de estruturas busca identificar equivalências entre pares de aminoácidos nas estruturas a serem sobrepostas, a sobreposição necessita desse conhecimento prévio sobre as equivalências.

Sendo assim, a sobreposição estrutural busca solucionar um problema muito mais simples, ou seja, minimizar a distância entre dois resíduos já reconhecidos como equivalentes. Isso se dá por encontrar transformações que satisfazem o menor desvio médio quadrático (RMSD) ou as equivalências máximas dentro de um valor limite para o RMSD.

Considerando que a estrutura das proteínas é mais conservada que a sequência, o alinhamento de estruturas confere maior especificidade ao alinhamento de sequências quando comparado ao alinhamento de sequências independente de estrutura. A maioria dos métodos de sobreposição de estruturas é adequado para identificar similaridades entre estruturas proteicas. O alinhamento de duas ou mais estruturas, porém, constitui uma tarefa mais difícil, e sua precisão depende tanto do método usado quanto do objetivo do usuário.

3.2. Alinhando sequências

À primeira vista, o processo de alinhamento entre diferentes sequências parece simples e não sujeito a qualquer tipo de erro. No entanto, esta afirmativa só é verdadeira em casos onde os organismos envolvidos possuem uma baixa taxa evolutiva (Figura 3a-3). Quando consideramos sequências homólogas amostradas de organismos com alta taxa evolutiva, ou até mesmo sequências similares, porém não homólogas, nos deparamos com casos particulares que tornam o processo de alinhamento complexo e, muitas vezes, sujeito a uma interpretação especialmente subjetiva por parte do usuário (Figura 3b-3).

A comparação de sequências homólogas de organismos evolutivamente distantes é um desafio para os programas de alinhamento. As diferentes pressões seletivas moldam os genomas de maneira imprevisível e, muitas vezes, acarretam a perda ou ganho de nucleotídeos ao longo do processo evolutivo. Para estes casos, a adição de lacunas (*gaps*) em matrizes de alinhamento, representadas por "-", é possível e muitas vezes necessária. As lacunas representam um ou mais eventos de inserção ou deleção de nucleotídeos. Estes eventos, comumente chamados de "indels" (*in* para inserção, e *del* para deleção), são fruto de processos mutagênicos (espontâneos ou induzidos) e, dependendo da região atingida, podem ser expressos nas moléculas de RNA e nas proteínas, onde poderão gerar consequências moleculares. Erros de replicação gerados pela DNA-polimerase durante a replicação do DNA, ou mesmo os eventos de recombinação, são os principais fatores atrelados à geração destes *indels* nos genomas. Em regiões codificadoras, estes eventos podem acarretar mudanças no quadro de leitura da proteína e torná-la não funcional.

Em termos analíticos, a inserção de lacunas dificulta o processo de alinhamento e exige interpretações cautelosas. Para determinados casos, especialmente em análises evolutivas e filogeográficas, é comum que regiões do alinhamento com determinado nível de incerteza, especialmente regiões com grande número de lacunas, sejam eliminadas

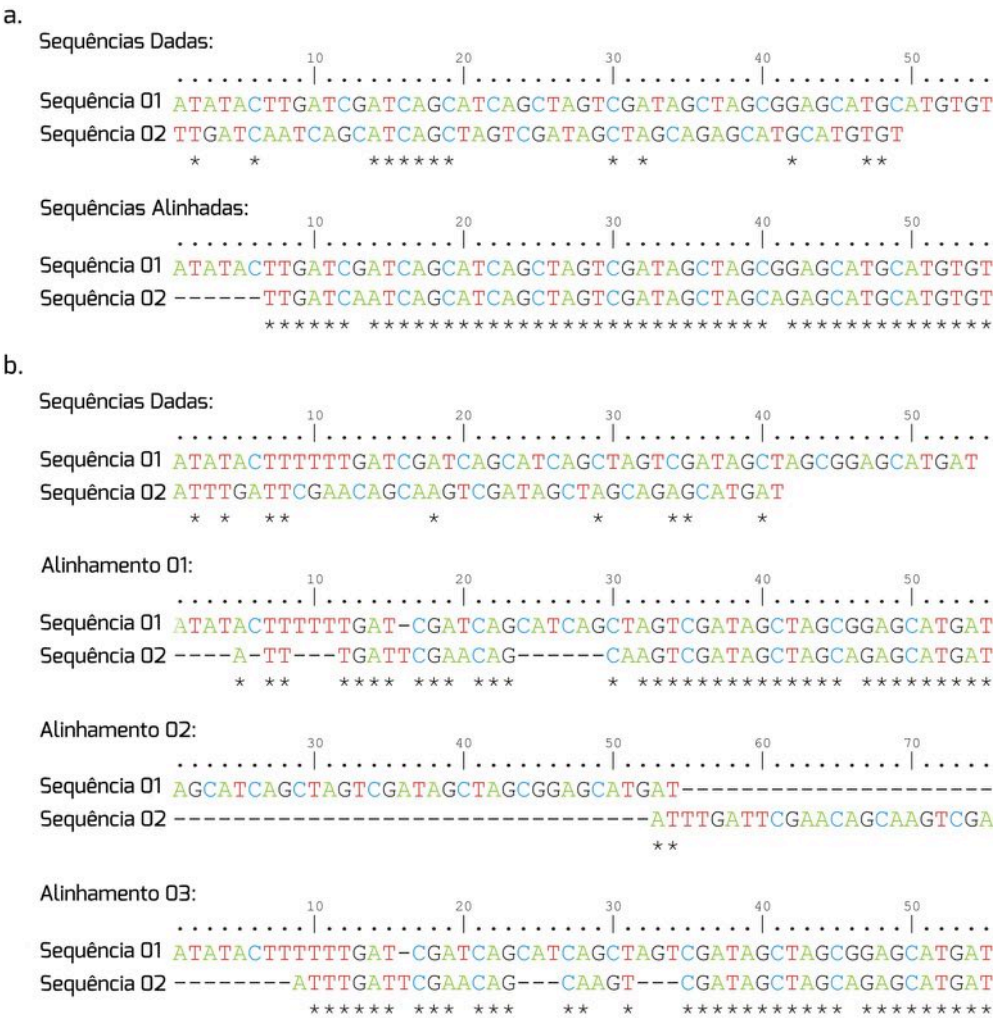


Figura 3-3: Alinhamentos de nucleotídeos. *a*) Duas sequências homólogas originadas de organismos com baixa taxa de evolução são dadas e seu alinhamento é proposto. *b*) Duas sequências homólogas amostradas de organismos com alta taxa de evolução são dadas e diferentes alinhamentos são propostos. Os hifens representam eventos de inserção ou deleção únicos na sequência. Os asteriscos identificam colunas com total similaridade dos caracteres.

da análise. Contudo, até o momento não existem programas capazes de lidar com as lacunas de forma coerentemente biológica. Apesar de sabermos que se tratam de eventos evolutivos comuns e bem caracterizados, as incertezas sobre o número de eventos e sua intensidade tornam as lacunas, em grande parte dos casos, um fator de confusão para análises de alinhamento.

Conforme mostrado na Figura 3-3, diferentes alinhamentos são possíveis para um mesmo grupo de sequências. A pergunta que se segue é: como reconhecer o melhor resultado quando nos deparamos com diversos alinhamentos possíveis para um mesmo conjunto de dados? Buscou-se resolver este problema através da criação de um sistema de pontuação para comparar os resultados de diferentes alinhamentos. Caracteres idênticos em sequências diferentes representam igualdades ou correspondências (*matches*) e, por serem resultados preferenciais durante o processo de alinhamento, são pontuados positivamente. Pelo contrário, caracteres não idênticos que ocupam a mesma coluna são chamados de desigualdades, ou *mismatches*, e recebem atribuições negativas. Como resultado, o melhor alinhamento possível para duas sequências é aquele que maximiza a pontuação total, somando os valores de *matches* e debitando os valores de *mismatches*.

Do ponto de vista biológico, as mudanças entre as bases nitrogenadas nas sequências de nucleotídeos não ocorrem com a mesma probabilidade (Figura 4a-3). Sendo assim, podemos atribuir valores de *mismatches* diferentes às transições (trocas de purinas por purinas ou pirimidinas por pirimidinas) e às transversões (trocas de purinas por pirimidinas ou pirimidinas por purinas). Para sequências de aminoácidos, é necessário escolher ativamente uma matriz de pontuação específica. Essas matrizes são resultados diretos de estudos de variação proteica e estão diretamente relacionadas à probabilidade de substituição de um aminoácido por outro (matrizes BLOSUM e PAM). Atualmente, as matrizes BLOSUM são as mais disseminadas e aplicadas para os mais diversos casos de comparação entre sequências de aminoácidos (Figura 4b-3).

a.

	A	C	G	T
A	1	-2	-2	-2
C		1	-2	-2
G			1	-2
T				1

b.

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	5	-2	-2	-2	0	0	0	0	-2	-2	-3	-2	-1	-2	0	0	0	-2	-3	0
R		5	-2	-3	-3	0	-1	-2	0	-3	-4	1	-3	-3	-2	-2	0	0	-3	-4
N			5	0	0	0	-2	0	0	-4	-5	-2	-3	-3	-2	0	0	-2	-2	-5
D				5	-4	0	1	-1	0	-5	-6	-3	-4	-4	0	-2	-2	-2	-2	-5
C					8	-2	-3	-1	-1	0	-2	-3	0	-1	-1	1	0	0	-2	0
Q						5	2	0	0	-2	-4	0	-2	-3	0	0	0	0	-2	-3
E							5	0	0	-3	-4	0	-3	-3	0	0	0	-2	-3	-3
G								6	0	-4	-5	-2	-3	-2	-2	0	0	0	-2	-3
H									6	-3	-4	0	-2	0	0	0	0	0	2	-2
I										4	0	-3	2	0	-2	-3	0	0	-3	2
L											4	-4	0	0	-3	-4	-3	0	-4	0
K												4	-2	-4	-1	-2	0	0	-3	-4
M													6	0	-3	-3	-2	0	-3	2
F														6	-3	-2	-2	2	2	0
P															7	0	0	-2	-3	0
S																4	2	-2	-2	-3
T																	5	-1	-3	0
W																		9	2	-1
Y																			7	-3
V																				4

Figura 4-3: Matrizes de custo utilizadas no cálculo de pontuação dos alinhamentos. *a*) Matriz de custo exemplo utilizada para cálculos de pontuação em alinhamentos de nucleotídeos. *b*) Matriz de custo BLOSUM62 utilizada para cálculo da pontuação em alinhamentos de aminoácidos.

Ainda, é necessário que as lacunas de alinhamentos recebam determinadas pontuações, pois são frequentemente encontradas em alinhamentos de dados biológicos. Se lacunas podem ser adicionadas em qualquer posição sem qualquer restrição, tanto nas extremidades quanto no interior das sequências, é possível gerar alinhamentos com mais lacunas do que propriamente caracteres a serem comparados (Figura 3b-3, alinhamento 2). Com o intuito de prevenir inserção excessiva, a adição de lacunas é penalizada durante a atribuição da pontuação de uma sequência, conforme um conjunto de parâmetros, chamado de penalidades por lacuna (*gap penalties*, PL). A abrangência da lacuna é pontuada pelo respectivo número de *indels* presentes no alinhamento. A fórmula mais comum para cálculo destas penalizações segue abaixo:

$$PL = g + e\,(L - 1)$$

onde L é o tamanho da lacuna (número de *indel*s presentes na lacuna), g é a penalidade pela abertura da lacuna (necessária para evitar que os alinhamentos contenham lacunas desnecessárias) e e é a penalidade atribuída a

cada *indel* (novamente para evitar grandes lacunas sem necessidade). Os valores de penalidade por lacuna são desenhados para reduzir a pontuação de um alinhamento quando este possui uma quantidade de *indels* desnecessária. Apesar da disseminação deste conceito, não há qualquer relação matemática ou biológica sustentando este cálculo. É importante destacar que, através da propriedade de "alinhamento livre de colunas em branco" (ou seja, *gaps* não são alinhados), as penalizações ainda impedem o alinhamento de *indels* entre as sequências envolvidas na análise. Assim, o melhor alinhamento entre as sequências será dado por um valor que resulta da soma dos valores associados a cada um dos *matches*, *mismatches* e lacunas, de acordo com um critério pré-definido (Figura 5-3).

O método de pontuação foi a solução encontrada para avaliar e classificar diferentes alinhamentos em busca da melhor explicação para a relação evolutiva entre as sequências. O próximo problema encontrado foi enumerar todas as possibilidades de alinhamentos para um grupo de dados. Assumindo-se duas sequências com tamanho de 100 caracteres cada, poderíamos enumerar até 10^{77} possíveis alinhamentos, diferentes entre si. A extensão de possibilidades inviabiliza a enumeração de todos os casos devido ao tempo e ao requerimento de enorme processamento destes dados. Apesar da exigência computacional, alguns algoritmos são capazes de realizar tal tarefa e ainda aplicar o método de pontuação para cada um dos casos, em busca do melhor resultado. No entanto, estes algoritmos não são capazes de lidar com sequências que contenham mais que algumas dezenas de caracteres. Em virtude da capacidade de explorar todas as soluções do problema, o processo realizado por estes algoritmos é chamado de "alinhamento ótimo".

Contudo, em virtude da inerente demora do processo, foi necessário desenvolver algoritmos que acelerassem a busca de um alinhamento capaz de explicar de maneira ótima os processos evolutivos para um determinado grupo de sequências sem, no entanto, enumerar todas as possibilidades. Os alinhamentos gerados por estes programas são chamados heurísticos, e compreendem métodos aproximados de busca pelo resultado ótimo. Diferentes métodos foram criados para diferentes tipos de alinhamento (Figura 6-3). Entre estes, devido à eficiência e à rapidez de processamento das informações de um alinhamento, incluindo o cálculo de pontuação, os algoritmos de programação dinâmica são, atualmente, os mais utilizados para este fim, tanto em alinhamentos simples como integrado aos algoritmos de alinhamentos múltiplos.

É fundamental assumirmos, para a maior parte dos problemas em bioinformática, o alinhamento como um modelo de relação evolutiva entre as sequências envolvidas. E como modelo, está sujeito à presença de certos problemas na explicação dos eventos evolutivos reais. Portanto, os alinhamentos devem ser avaliados com extrema cautela. A facilidade e a aparente simplicidade na análise dos programas tornam o processo mecânico e desvinculado de análises críticas pela maior parte dos usuários. A associação dos métodos de alinhamento a outras análises de bioinformática tende a desvincular a real importância desta técnica e a coloca apenas como um procedimento, e não formalmente como uma técnica sujeita à análise crítica. Isto pode ocasionar na obtenção de modelos incorretos ou mesmo de falsos positivos.

3.3. Tipos de alinhamento

Em estudos de bioinformática, é comum compararmos moléculas de dois ou mais indivíduos, sejam eles da mesma espécie ou de espécies diferentes. Quanto maior o número de sequências comparadas, maior o tempo exigido para conclusão do alinhamento e, dependendo das sequências envolvidas, maior a dificuldade dos algoritmos em encontrar o melhor resultado. Conforme a quantidade de sequências envolvidas, podemos dividir os alinhamentos em dois tipos: alinhamentos simples, ou par-a-par, e alinhamentos múltiplos, ou de múltiplas sequências (Figura 7-3).

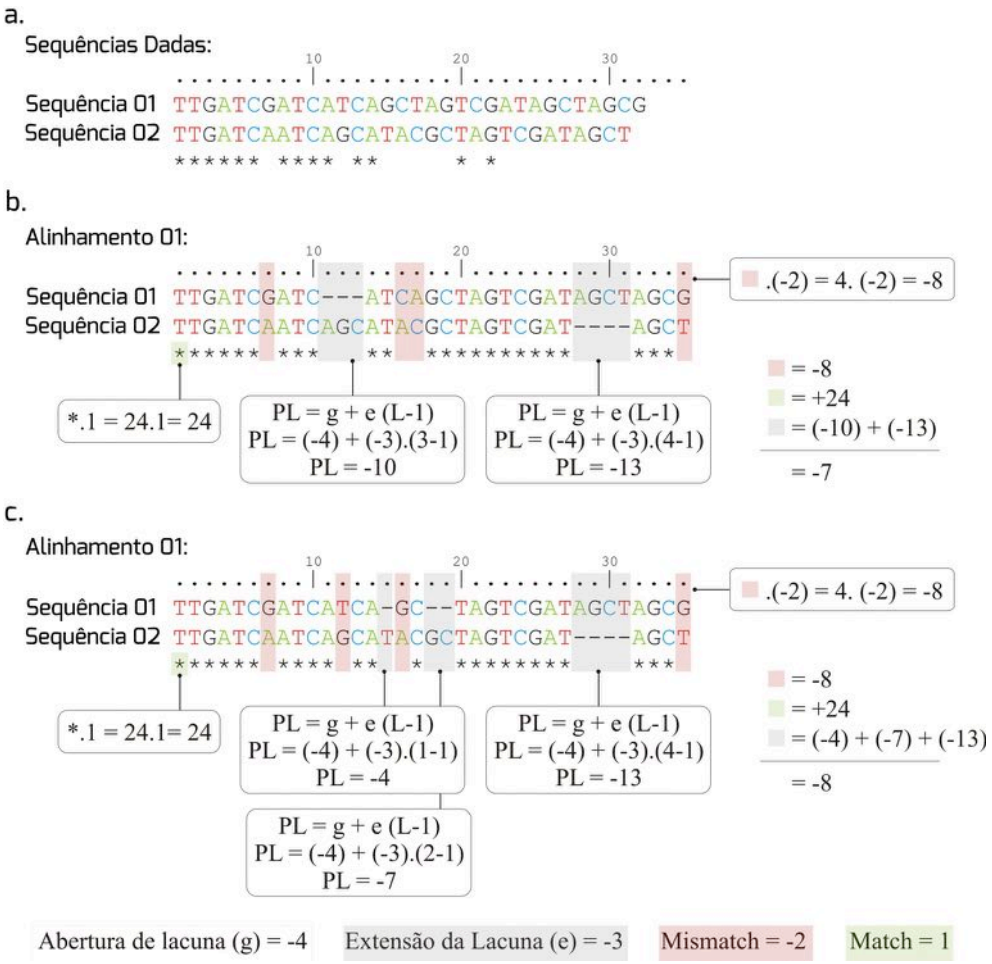


Figura 5-3: Esquema de pontuação para avaliação de alinhamentos. *a*) Duas sequências de desoxirribonucleotídeos não alinhadas. *b*) Proposição de um alinhamento para as sequências dadas em *a*. O alinhamento possui 24 colunas de *matches*, 4 colunas de *mismatches* e duas lacunas com 3 e 4 *indels*. A pontuação total para o alinhamento desta sequência é -7. *c*) Proposição de um segundo alinhamento para as sequências dadas em *a*. O alinhamento possui 24 colunas de *matches*, 4 colunas de *mismatches* e três lacunas com 1, 2 e 4 *indels*. A pontuação total para o alinhamento desta sequência é -8. A partir deste exemplo, o alinhamento com a maior pontuação é o mostrado em *b*. Os valores de pontuação utilizados neste exemplo são especificados na parte inferior da figura.

Os alinhamentos simples descrevem especificamente a relação de similaridade entre duas sequências quaisquer. Já os alinhamentos múltiplos incluem três ou mais sequências na análise de similaridade e, dependendo do objetivo do usuário, podem envolver até centenas de sequências.

Conceitualmente, ainda podemos dividir os alinhamentos, tanto simples, como múltiplos, em dois grandes tipos. Os alinhamentos que levam em consideração toda a extensão das sequências são conhecidos como globais, enquanto aqueles que buscam pequenas regiões de similaridade são chamados de locais (Figura 7-3). Em algoritmos que buscam o alinhamento global de duas sequências, reforça-se a busca do alinhamento completo das sequências envolvidas, procurando incluir o maior número de *matches* do início ao final das sequências. Quando necessário, estes algoritmos permitem a inserção de lacunas para que as sequências tenham o mesmo tamanho no resultado do alinhamento (Figura 7b-3).

Graficamente, os sítios com caracteres idênticos são representados ligados por barras verticais, enquanto os sítios que possuem caracteres diferentes nas duas sequências, ou

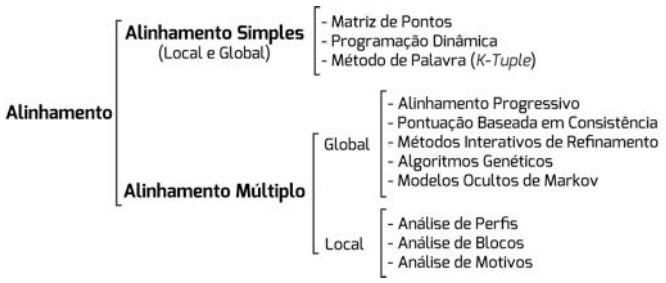


Figura 6-3: Tipos de alinhamento e os algoritmos aplicados à bioinformática.

mesmo a presença de uma lacuna em uma delas, permanecem sem qualquer notação (Figura 7-3). O principal algoritmo envolvido no processamento de alinhamentos globais é aquele desenvolvido por Needleman e Wunsch durante a década de 1970. Além de ter uma notável importância metodológica, este algoritmo tem grande importância na história do alinhamento, pois foi o primeiro algoritmo a aplicar o método de programação dinâmica para a comparação de sequências biológicas.

Em seu início, os métodos de alinhamento eram utilizados especialmente para a comparação par-a-par de sequências de proteínas inteiras. No entanto, com a ampliação da disponibilidade de sequências completas de proteínas, foi necessário buscar métodos de alinhamento que privilegiassem a busca de similaridade, não entre sequências completas, mas apenas entre porções isoladas destas sequências. Durante a década de 1980 iniciou-se o desenvolvimento de novos algoritmos de alinhamento, já que os desenvolvidos até aquele momento não eram aplicáveis para esta particularidade. Entre estes novos algoritmos, o desenvolvido por Smith e Waterman, em 1981, ganhou maior destaque e atualmente é o principal algoritmo utilizado por programas para realização de alinhamentos locais. Nestes casos, privilegia-se o alinhamento de partes da sequência, buscando apenas as regiões com a maior similaridade (Figura 7c-3). Em algoritmos para busca local, o alinhamento pára no final das regiões de alta similaridade e substitui as regiões excluídas por hifens (lacunas) no resultado final (Figura 7c-3).

3.4. Alinhamento simples

Para entender como se processa um alinhamento par-a-par e como o grau de si-

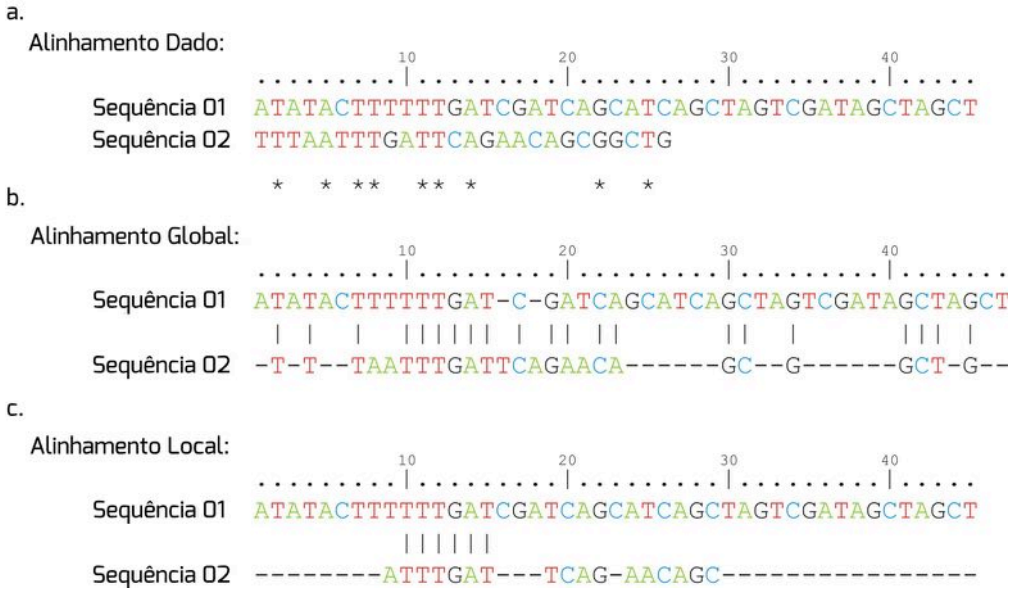


Figura 7-3: Diferenças entre alinhamento local e global. *a*) Duas sequências de nucleotídeos de tamanhos diversos são amostradas e alinhadas por algoritmos diferentes. *b*) No alinhamento local, a prioridade é encontrar as regiões altamente similares, independentemente do tamanho desta região. Neste caso, porções da sequência que não foram alinhadas com alta similaridade foram excluídas do resultado final. *c*) No alinhamento global, as duas sequências são alinhadas por completo, independentemente do número de lacunas que tenham que ser inseridas.

milaridade entre elas pode ser computado, apresentamos três dos principais algoritmos desenvolvidos para este fim: algoritmos de programação dinâmica, análise de matriz de pontos (*dot matrix*) e método de palavra ou *k-tuple*.

A programação dinâmica é, atualmente, o método mais utilizado por programas para realizar o alinhamento de sequências. Em casos simples (par-a-par), é capaz de encontrar o melhor alinhamento para duas sequências através da aplicação da pontuação de similaridades. É, portanto, um método de execução relativamente rápida nos computadores modernos, requerendo um tempo e memória de processamento proporcional ao produto do tamanho das duas sequências envolvidas.

O método é baseado no princípio de otimização de Bellmann, e propõe a solução de problemas complexos através da resolução dos seus diversos subproblemas. Os subproblemas são resolvidos e seus resultados são armazenados pelo algoritmo. A vantagem funcional da resolução em partes é que, geralmente, problemas complexos combinam uma série de subproblemas. Como o algoritmo acumula os resultados dos diferentes subproblemas, acelera a resolução do problema complexo. Assim, a designação "programação" nada tem a ver com programação de computadores, mas com a organização dos resultados já solucionados para resolução de um problema maior.

Conforme discutimos anteriormente, em determinados casos, duas sequências podem apresentar diferentes alinhamentos. Se não há *indels* e as sequências são similares, o alinhamento é rápido e não deixa dúvidas. No entanto, quando existe certa diversidade entre as sequências envolvidas e uma quantidade de suficiente de *indels*, a solução para o alinhamento é menos óbvia visualmente. Nestes casos, os algoritmos de programação dinâmica buscarão solucionar os subproblemas envolvidos e fornecerão o melhor resultado.

Para cálculo do melhor alinhamento entre duas sequências, o algoritmo de programação dinâmica necessita da especificação de um esquema de pontuação, seja ele referente a nucleotídeos ou aminoácidos. Da mesma forma, é necessário fornecer um valor de penalidade para a abertura e extensão das lacunas. A partir destas informações, o algoritmo calculará uma relação entre todos os caracteres das sequências e fornecerá o melhor alinhamento como resultado final.

Como exemplo, consideraremos a Figura 8-3. São dadas duas sequências, sequência 1 e sequência 2, um esquema de pontuação e, para facilitar o entendimento do cálculo, um valor único de penalidade por lacuna de -8. O algoritmo toma as sequências e transforma a relação entre elas em uma tabela, onde as linhas são definidas pelos caracteres da sequência 01, e as colunas pelos caracteres da sequência 02. A fim de permitir lacunas no início do alinhamento, o algoritmo impõe a inserção de uma coluna e de uma linha iniciais contendo o símbolo de *indel*. A partir deste ponto, para cada um dos elementos da matriz, o algoritmo calculará a melhor pontuação dos subcaminhos associados ao alinhamento: uma substituição, uma inserção na sequência 01 ou uma inserção na sequência 2. Assim, o melhor subcaminho será calculado segundo uma função de pontuação, conforme abaixo:

$$F(i,j) = max \begin{cases} \text{valor da célula na diagonal superior esquerda + pontuação da similaridade} \\ \text{valor da célula acima + valor da penalidade por lacuna} \\ \text{valor da célula à esquerda + valor da penalidade por lacuna} \end{cases}$$

A partir do elemento (1,1) da matriz e ao longo da primeira linha, apenas a terceira condição é satisfeita (valor da célula à esquerda + valor da penalidade por lacuna). Na primeira coluna, apenas a segunda condição é satisfeita. Para outros elementos, as três condições devem ser calculadas e aquela que resultar no maior valor é escolhida para formar a matriz. Além disso, os procedimentos dos algoritmos de programação dinâmica podem ser representados por pequenas setas para indicar qual subcaminho obteve o melhor valor (Figura 8-3).

Outro método importante na área de alinhamento de sequências é a análise de matriz de pontos ou matriz *dot*. É um método simples e bastante eficiente em análises de

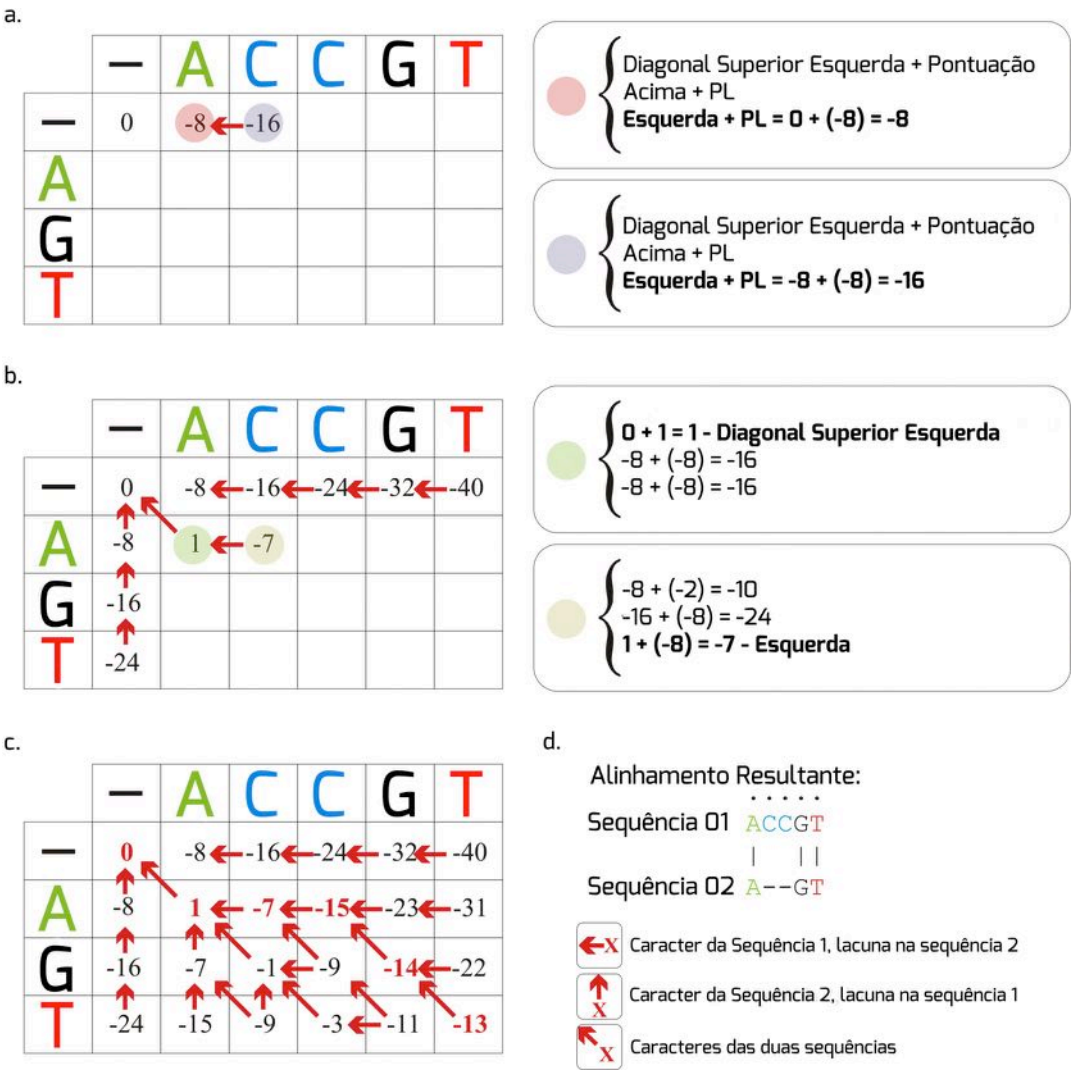


Figura 8-3: Alinhamento de duas sequências de nucleotídeos através do método de programação dinâmica. *a*) As sequências a serem alinhadas são dispostas em uma tabela onde o número de colunas corresponde ao número de caracteres da sequência 1 mais um (devido à adição de uma coluna para uma lacuna) e o número de linhas corresponde ao número de caracteres da sequência 2 mais um. O caractere atribuído à primeira linha e à primeira coluna é, por definição, o símbolo "-", atribuído a uma lacuna. Através da matriz de penalidades calculam-se os valores para as três possibilidades $F(i,j)$, buscando a equação que resulte no maior valor. O valor arbitrário de penalidade por lacuna (PL) é de -8. Em virtude de a primeira linha não possuir valores de comparação na diagonal superior esquerda e acima, considera-se apenas a terceira equação. *b*) O valor demarcado em verde é o primeiro a ser calculado após o preenchimento da primeira linha e primeira coluna, representando o menor valor encontrado no cálculo para $F(i,j)$. Além do cálculo, o algoritmo de programação dinâmica insere informações a respeito da direção da informação. Como o valor "1" foi o maior valor encontrado e representa o cálculo utilizando a informação situada na diagonal superior esquerda, demarcada em verde, insere-se uma seta nesta direção. *c*) O preenchimento completo da tabela e as respectivas setas ilustrando a direção da informação. Algumas casas estão demarcadas com duas setas, pois apresentaram dois valores máximos idênticos na resolução das equações. Ao final dos cálculos, iniciando pelo canto inferior direito, seguem-se as setas em busca dos maiores valores. *d*) Relacionando os dados da tabela com a simbologia apresentada, chega-se ao alinhamento final entre as sequências 1 e 2.

deleções/inserções e para detectar repetições diretas ou inversas, especialmente em sequências de nucleotídeos. Além disso, vem sendo utilizado para buscar regiões de pareamentos intra-cadeia capazes de formar estruturas 2árias em moléculas de RNA. Este método permite a visualização gráfica das regiões de similaridade entre sequências através da construção de uma matriz de identidade. O número de linhas desta matriz é definido pelo número de caracteres de uma das sequências, e o número de colunas é definido pelo número de caracteres da outra sequência a ser comparada (Figura 9-3). É primariamente um método visual, e não fornece o alinhamento propriamente dito como resultado final, embora seja frequentemente utilizado quando se deseja visualizar as regiões de similaridade entre duas sequências.

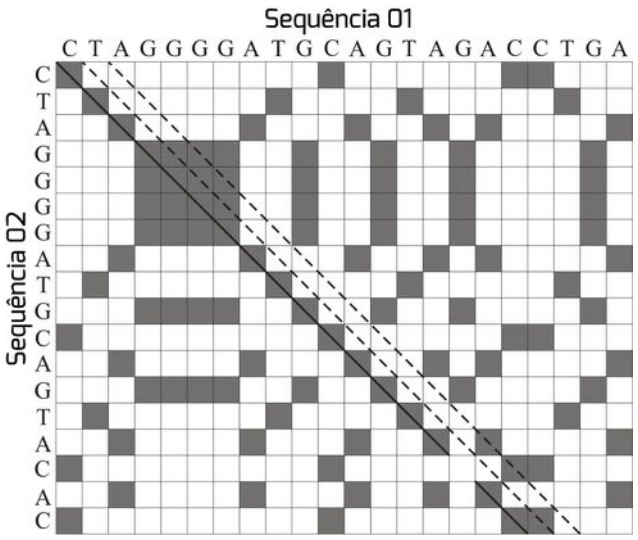


Figura 9-3: Análise de matriz de pontos de duas sequências de DNA. Os pontos assinalados em cinza representam a concordância de caracteres entre a sequência 1 e a sequência 2. A partir da diagonal direita inferior, são traçadas diferentes retas. Aquela que atingir o maior número de pontos assinalados deve ser escolhida como resultado para o alinhamento entre as duas sequências. A linha contínua representa a possibilidade mais adequada a esta análise e as linhas tracejadas representam possibilidades de insucesso.

Neste método, inicialmente, uma das sequências é disposta na vertical e a outra na horizontal (Figura 9-3). Regiões do gráfico que possuam o mesmo caractere tanto na sequência disposta na horizontal, quanto na sequência disposta na vertical, serão assinalados. Esta marcação representa os possíveis correspondências (*matches*) entre uma sequência e outra.

Qualquer região de similaridade entre as duas sequências será evidenciada por uma linha diagonal de assinalações. Pontos não dispostos na diagonal representam correspondências aleatórias que não estão relacionadas com a similaridade entre as sequências. A detecção de regiões de alta similaridade pode ser beneficiada, em alguns casos, através da comparação de dois ou mais caracteres ao mesmo tempo. Nestes casos, é necessário escolher um número de caracteres como janela.

Além disso, arbitrariamente, um número de correspondências deve ser escolhido. Por exemplo, para comparar duas sequências com 100.000 caracteres, podemos escolher uma janela de 15 caracteres e 10 correspondências requeridas. O algoritmo varrerá a matriz de 15 em 15 caracteres e, quando, entre estes quinze caracteres, existirem 10 formando correspondências entre as duas sequências, o algoritmo inserirá uma marcação de similaridade. Geralmente, esta variação do método é utilizada para a comparação de longas sequências de DNA.

Por último, outro algoritmo bastante comum no alinhamento par-a-par de dados biológicos é o *k-tuple*, ou método de palavras. Este método é geralmente mais rápido que o método de programação dinâmica, embora não garanta o melhor alinhamento como resultado. Este tipo de algoritmo é especialmente útil em casos onde se busca similaridade de uma única sequência contra um grande conjunto de dados. Para isso, o algoritmo dividirá uma sequência alvo em pequenas sequências, geralmente conjuntos de dois a seis caracteres, chamados de palavras. Da mesma forma, o conjunto total de sequências do banco de dados terá cada uma das sequências subdivida em pequenas pala-

vras. As palavras da sequência alvo serão comparadas às palavras oriundas do banco de dados. Após a busca de identidade, o algoritmo alinhará as duas sequências completas (sequência oriunda do banco de dados que teve uma palavra similar com umas das palavras da sequência alvo e a própria sequência alvo) a partir das palavras similares e estenderá a análise de similaridade para as regiões vizinhas, antes e depois da palavra similar. Através de uma matriz de penalidade, o algoritmo calculará o alinhamento que teve o maior valor de pontuação. É comum, para esta segunda etapa dos cálculos de similaridade, a utilização de algoritmos de programação dinâmica.

3.5. Alinhamento múltiplo global

Da mesma forma que no caso dos alinhamentos simples, o método de programação dinâmica é usualmente utilizado para lidar com múltiplas sequências. Nestes casos, utiliza-se o conceito de soma ponderada dos pares (*weighted sum of pairs*, WSP). Através deste conceito, para qualquer alinhamento múltiplo de sequências, uma pontuação para cada par possível formado por estas sequências será calculada (Figura 8-3) e, ao final, os valores de similaridade para cada um dos pares serão somados. Apesar de conceitualmente simples, este método exige grande capacidade computacional e, dependendo da quantidade de sequências envolvidas, pode requerer longo tempo para processamento.

Métodos alternativos tiveram que ser criados para acelerar os cálculos para alinhamento de sequências, incluindo-se: alinhamento progressivo, pontuação baseada em consistência (*consistency-based scoring*), métodos iterativos de refinamento, algoritmos genéticos e modelos ocultos de Markov. Cabe ressaltar que todos estes métodos realizam buscas aproximadas pelo resultado ótimo e, portanto, se tratam de métodos heurísticos.

Alinhamento progressivo

Leva em consideração a relação evolutiva entre as sequências. Os algoritmos utilizam as relações filogenéticas para gerar o resultado de alinhamento. Inicialmente, são realizados alinhamentos par-a-par de todos os possíveis pares. Nesta comparação, verifica-se apenas o número de caracteres diferentes entre as duas sequências (verificar o conceito de distância evolutiva observada no capítulo 6). Estas distâncias serão utilizadas para a construção de uma filogenia (geralmente através do método de *neighbor-joining*). A partir desta filogenia o alinhamento será construído progressivamente, dependendo da relação entre as sequências sendo, por isso, chamado de alinhamento progressivo.

Tomemos como exemplo um ramo de uma dada filogenia que inclui duas sequências. O algoritmo construirá um alinhamento através de programação dinâmica para estas duas sequências. A partir deste primeiro alinhamento, estas duas sequências serão agora tratadas como uma, e serão alinhadas à próxima sequência filogeneticamente relacionada. Devemos notar que todo o restante das sequências será alinhado baseando-se neste primeiro par. É um método rápido e amplamente utilizado para alinhar um grande número de sequências. Atualmente, os programas mais populares de alinhamento progressivo são o CLUSTALW e CLUSTALX.

Pontuação baseada em consistência

Baseado no algoritmo de alinhamento progressivo, não leva em consideração apenas o primeiro par de sequências alinhadas. Durante a realização do cálculo, realiza outros alinhamentos par-a-par para aperfeiçoar as comparações entre as sequências. O principal programa a utilizar este algoritmo é o T-COFFEE.

Métodos iterativos de refinamento

Funcionam como os algoritmos de ali-

nhamento progressivo, mas os grupos de sequências são realinhados constantemente ao longo das análises, garantindo que o alinhamento inicial não defina o resultado final. O principal programa a utilizar este algoritmo como base para os cálculos de alinhamento é o MUSCLE.

Algoritmos genéticos

Estes algoritmos buscam simular o processo evolutivo no conjunto de sequências a serem alinhadas, aplicando conceito de seleção e recombinação. É ainda um método lento e, devido à aleatoriedade do processo, não garante o mesmo resultado para diferentes alinhamentos do mesmo conjunto de dados. O programa SAGA é um dos poucos a implementar algoritmos genéticos.

Modelos ocultos de Markov

Modelo baseado em probabilidades estatísticas, destacando os eventos de substituição e inserção ou deleção de caracteres.

3.6. Alinhamento múltiplo local

Na busca por regiões localizadas de similaridade entre diferentes sequências, são aplicados principalmente os seguintes algoritmos: análise de perfis, análise de blocos e análise de motivos.

Análise de perfis

A partir de um alinhamento primário de todas as sequências envolvidas na análise e utilizando uma matriz de custo padrão, o algoritmo seleciona as regiões altamente conservadas e produz uma nova matriz de pontuação (matriz de custo), chamada de perfil. A construção deste perfil pode ser realizada através de dois métodos diferentes (método das médias e método evolutivo) e inclui pontuações para *matches*, *mismatches* e lacunas. Assim que produzido, este perfil pode ser utilizado para alinhar sequências entre si utilizando as pontuações calculadas para avaliar a probabilidade em cada posição ou para buscar sequências com o mesmo padrão em um banco de dados.

A desvantagem do método de perfis está na especificidade da nova matriz de custo obtida. Se o alinhamento inicial contiver poucas sequências, pode não representar adequadamente a variabilidade de caracteres em uma determinada posição e prejudicar o algoritmo na busca por similaridade com outras sequências. Este método é principalmente utilizado para alinhamentos de aminoácidos.

Análise de blocos

Assim como a análise de perfis este método requer, inicialmente, a seleção da região de maior similaridade de um alinhamento múltiplo. Estas regiões podem ser chamadas de blocos e diferem dos perfis por não acomodarem *indels*, que serão automaticamente eliminados das análises. Este método é também capaz de realizar a busca de pequenas regiões de similaridade entre sequências, de maneira semelhante ao método de palavras.

Análise de motivos

Este método é especialmente utilizado na busca por motivos proteicos em sequências de aminoácidos. O método foi desenvolvido através do alinhamento de milhares de sequências de aminoácidos extraídas de grandes bancos de dados de proteínas. A partir deste alinhamento, analisou-se cada uma das colunas para buscar um padrão de substituição entre os aminoácidos. Estes padrões de mudança refletem uma maior probabilidade de substituição. Para proceder ao alinhamento, os algoritmos que aplicam a análise de motivos iniciam o processo por uma análise de blocos. As regiões de alta similaridade são então analisadas para buscar os padrões de substituição descritos inicialmente. O conjunto de padrões resultante da análise das colunas é chamado de motivo. A probabilidade de existência de cada motivo em uma sequência de proteína é estimada através do banco de dados do SwissProt.

3.7. BLAST

O BLAST, ou Ferramenta de Busca por Alinhamento Local Básico (*Basic Local Alignment Search Tool*) é um algoritmo capaz de realizar buscas baseadas em alinhamento que, apesar de não serem exatas, são confiáveis e muito rápidas, sendo estas suas vantagens em relação a outros métodos. Ele é um dos programas mais usados em Bioinformática devido à velocidade em que consegue responder a um problema fundamental em biologia celular e molecular: comparar uma sequência desconhecida com aquelas depositadas em bancos de dados.

O algoritmo do BLAST aumenta a velocidade do alinhamento de sequências ao buscar primeiro por palavras comuns (ou *k-tuples*) na sequência de busca e em cada sequência do banco de dados. Em vez de buscar todas as palavras de mesmo tamanho, o BLAST limita a busca àquelas palavras que são mais significantes. O tamanho de palavra é fixado em 3 caracteres para sequências de aminoácidos e em 11 para sequências de nucleotídeos (3 se as sequências forem traduzidas nos 6 quadros de leitura possíveis). Esses são os tamanhos mínimos para obter uma pontuação por palavras que seja alta o suficiente para ser significativa sem perder fragmentos menores, mas importantes, de sequência.

Funcionamento do algoritmo BLAST

Para funcionar, o BLAST necessita de uma sequência de busca (*query*) e de sequências alvo. Comumente, as sequências alvos são o conjunto de sequências depositadas em um banco de dados, local ou na *web*. Um dos conceitos principais empregados pelo BLAST é de que alinhamentos estatisticamente significantes contêm pares de segmentos de alta pontuação (HSP, *high-scoring segment pairs*), e são esses HSPs que o algoritmo busca entre a sequência sendo analisada e aquelas depositadas no banco de dados.

As principais etapas do funcionamento do algoritmo BLAST, para uma sequência proteica genérica incluem:

i. Remoção de repetições ou regiões de baixa complexidade na sequência de busca.

Uma região de baixa complexidade é definida como uma região composta por poucos tipos de elementos. Essas regiões normalmente apresentam pontuações altas que podem confundir o programa em sua busca por sequências com similaridade significativa. Por esse motivo, tais regiões são identificadas antes da próxima etapa e ignoradas.

ii. Estabelecer uma lista de palavras com k-letras.

Sendo este um caso envolvendo sequências proteicas, k = 3, ou seja, cada palavra tem tamanho 3. Como mostrado na Figura 10-3, são listadas palavras com comprimento de 3 caracteres, sequencialmente, até que a última letra da sequência de busca seja incluída.

Figura 10-3: Exemplo de lista de palavras geradas pelo BLAST.

iii. Listar as possíveis palavras correspondentes.

Diferente de outros algoritmos (como o FASTA), o BLAST considera apenas as palavras de maior pontuação. As pontuações são estabelecidas por comparação das palavras listadas na etapa *ii* com todas as outras palavras de 3 letras. Uma matriz de substituição (BLOSUM62) é usada para pontuar as comparações entre pares de resíduos. Existem 20^3 possíveis pontuações de correspondência considerando uma palavra de 3 letras. Como exemplo, a comparação das palavras PQG e PEG tem pontuação de 15, enquanto a comparação de PQG com PQA pontua como 12. A seguir, um limiar T para pontuação de palavras vizinhas é usado para reduzir o número de possíveis palavras correspondentes. As palavras cujas pontuações forem maiores que o limiar T serão mantidas na lista de possíveis correspondências, enquanto aquelas cujas pontuações

forem menores serão descartadas. Considerando o exemplo anterior, se T = 13, PEG será mantida, enquanto PQA será abandonada.

iv. Organizar as palavras de alta pontuação.

As palavras remanescentes, com alta pontuação, são organizadas em uma árvore de busca. Isso permite que o programa compare as palavras com as sequências do banco de dados de maneira rápida.

v. Repetir os passos *iii* e *iv* para cada palavra de k-letras originadas da sequência de busca.

vi. Varrer as sequências do banco de dados em busca de correspondências com as palavras remanescentes.

O BLAST realiza uma varredura das sequências depositadas no banco de dados, buscando pelas palavras de alta pontuação (como PEG, no exemplo anterior). Se uma correspondência exata for encontrada, ela será empregada para nuclear um possível alinhamento sem lacunas (*gaps*) entre a sequência de busca e a depositada no banco de dados.

vii. Estender as correspondências exatas entre pares de segmentos de alta pontuação.

A versão original do BLAST estende o alinhamento para a esquerda e para a direita de onde ocorre uma correspondência exata. A extensão é parada apenas quando a pontuação acumulada pelo HSP começa a diminuir (um exemplo pode ser visto na Figura 11-3).

Figura 11-3: Exemplo do esquema de pontuação empregado pelo BLAST.

Para acelerar o processo, a versão atual do BLAST (BLAST2 ou *Gapped* BLAST) emprega um limiar mais baixo para a vizinhança das palavras, mantendo a sensibilidade na detecção de similaridade de sequências. Assim, a lista de possíveis correspondências obtidas na etapa *iii* é maior. Como observado na Figura 12-3, as regiões de correspondência exata com distância menor que A na mesma diagonal serão unidas como uma nova região, mais extensa. Posteriormente, essas regiões são estendidas da mesma maneira como ocorre no BLAST original, com os HSPs sendo pontuados com base em uma matriz de substituição.

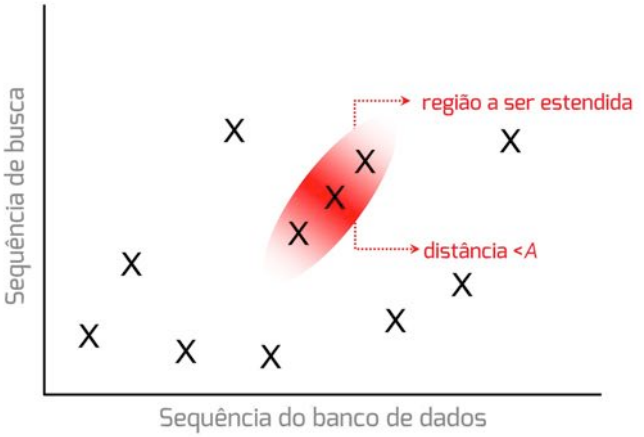


Figura 12-3: Esquema da extensão de zonas de correspondência entre sequências identificadas pelo BLAST.

viii. Listar todos os HSPs do banco de dados cuja pontuação seja alta o suficiente.

Nessa etapa são listados todos os pares de segmentos cuja pontuação seja maior que um determinado ponto de corte S. A distribuição de pontuações obtidas por alinhamento de sequências aleatórias é a base para determinação desse ponto de corte.

ix. Avaliar a significância da pontuação dos HSPs.

A avaliação estatística de cada par de segmentos de alta pontuação explora a Distribuição de Valores Extremos de Gumbel. O valor de confiança estatística e apresentado pelo BLAST, chamado de valor de expectativa, reflete o número de vezes que uma sequência não relacionada presente no banco de dados pode obter, ao acaso, um valor maior que S (ponto de corte). Ou seja, o e reflete o número de falsos positivos entre os resultados de similaridade encontrados. Para $p < 0{,}1$, o valor e se aproxima da distribuição de Poisson (ver item 4.8).

x. Transformar duas ou mais regiões de HSP em um alinhamento maior.

Em alguns casos, duas ou mais regiões de HSP podem ser combinadas em um trecho maior de alinhamento (uma evidência adicional da relação entre a

sequência de busca e a encontrada no banco de dados). Existem dois métodos para comparar a significância das novas regiões ligadas. Se, por exemplo, forem encontradas duas regiões de HSP combinadas com pares de pontuação (67 e 41) e (53 e 45), cada método se comportará de maneira diferente. O método de Poisson conferirá maior significância ao conjunto com valor mínimo maior (45 em vez de 41). O método de soma dos pontos, ao contrário, dará preferência ao primeiro conjunto, pois 108 (67+41) é maior que 98 (53+45). O BLAST original usa o primeiro método, enquanto o BLAST2 emprega o segundo.

xi. Exibir os alinhamentos locais entre a sequência de busca e cada uma das correspondências no banco de dados.

O BLAST original produz apenas alinhamentos sem lacunas (*gaps*), incluindo cada um dos HSPs encontrados inicialmente, mesmo que mais de uma região de correspondência seja encontrada numa mesma sequência do banco de dados. O BLAST2 produz um único alinhamento com lacunas, podendo incluir todas as regiões de HSP encontradas. É importante destacar que o cálculo da pontuação e do valor e leva em conta as penalidades por abertura de lacunas no alinhamento.

xii. Registrar as correspondências encontradas.

Quando o valor *e* dos alinhamentos encontrados entre a sequência de busca e as do banco de dados satisfazem o ponto de corte estabelecido pelo usuário, a correspondência é registrada. Os resultados da busca são apresentados de forma gráfica, seguidos por uma lista de correspondências organizada pela pontuação e pelo valor *e*, e finalizam com os alinhamentos. A Figura 13-3 traz um exemplo de resultado obtido pelo BLAST.

Diferentes tipos de BLAST

O BLAST constitui uma família de programas, que podem ser usados para diferentes fins, dependendo das necessidades do usuário. Esses programas variam quanto ao tipo de sequência de busca, o banco de dados a ser empregado, e o tipo de comparação a ser realizada. As diferentes aplicações disponíveis pelo BLAST incluem:

i. *blastn*: BLAST nucleotídeo-nucleotídeo. Usando uma sequência de DNA como entrada, dá como resultado as sequências de DNA mais similares presentes no banco de dados especificado pelo usuário.

ii. *blastp*: BLAST proteína-proteína. Usando uma sequência proteica como entrada, dá como resultado as sequências proteicas mais similares presentes no banco de dados especificado pelo usuário.

iii. *blastpgp*: BLAST iterativo com especificidade de posição (PSI-BLAST). Usado para encontrar proteínas distantemente relacionadas. Nesse caso, uma lista de proteínas proximamente relacionadas é criada. Essa lista serve de base para a criação de uma sequência média, que resume as características importantes do conjunto de sequências. A sequência média é usada para buscar sequências similares no banco de dados e um grupo maior de proteínas é encontrado. O grupo maior é usado na construção de uma nova sequência média e o processo é repetido. Ao incluir proteínas relacionadas na busca, o PSI-BLAST é muito mais sensível na percepção de relações evolutivas distantes que o BLAST proteína-proteína tradicional.

iv. *blastx*: tradução de nucleotídeos em 6 quadros-proteína. Compara os produtos de tradução conceitual nos 6 quadros de leitura de uma sequência de nucleotídeos contra o banco de dados de sequências proteicas.

v. *tblastx*: tradução de nucleotídeos em 6 quadros-tradução de nucleotídeos em 6 quadros. O mais lento dos programas BLAST, tem por objetivo encontrar relações distantes entre sequências de nucleotídeos. Ele traduz a sequência de nucleotídeo nos 6 possíveis quadros de leitura e compara os resultados contra a tradução nos 6 quadros de leitura das sequências de nucleotídeos depositadas no banco de dados.

vi. *tblastn*: proteína-tradução de nucleotídeos em 6 quadros. Compara uma sequência de proteína contra a tradução nos 6 quadros de leitura das sequências de nucleotídeos depositadas no banco

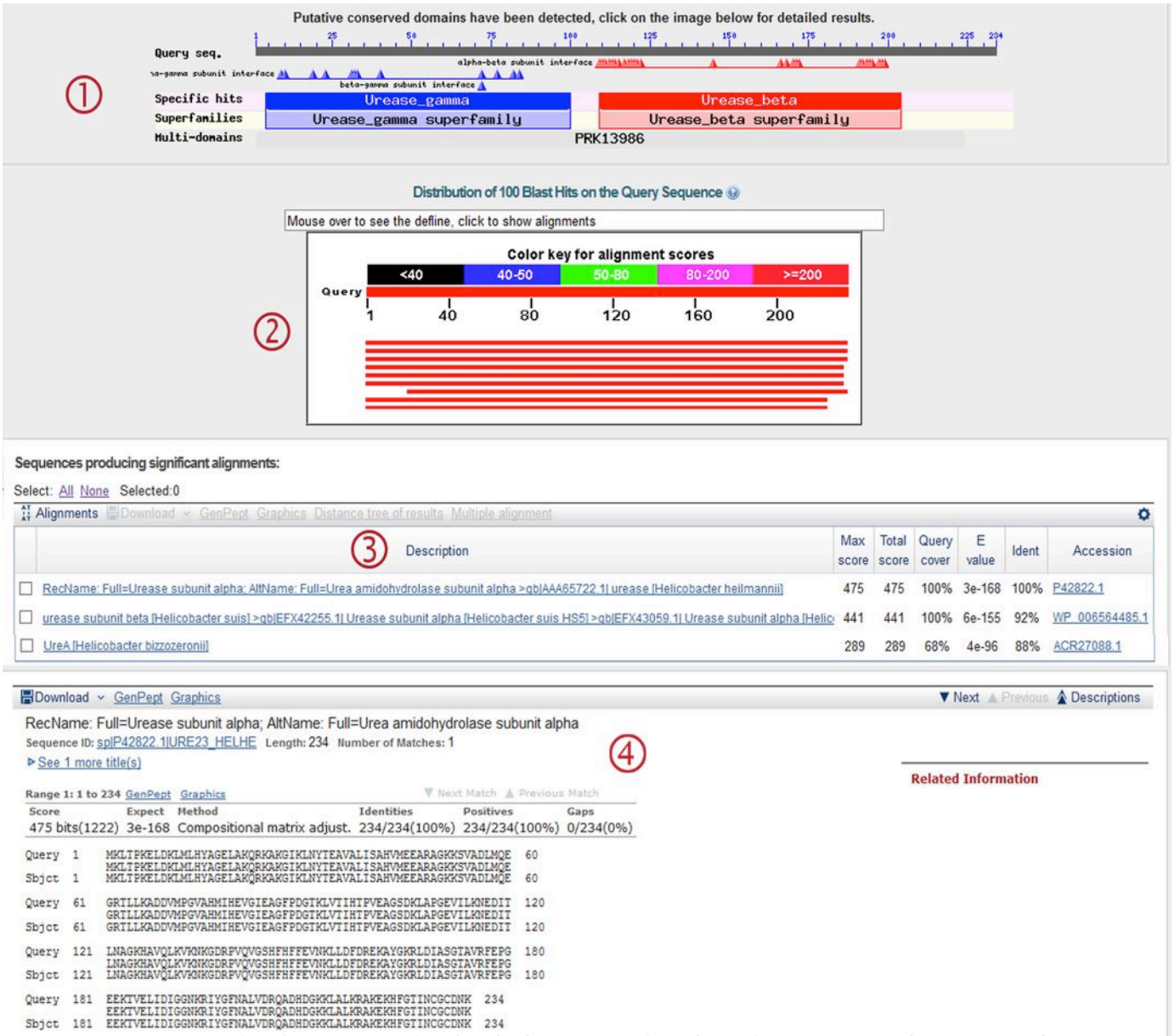


Figura 13-3: Exemplo de um resultado de busca realizada pelo BLAST. Diferentes informações são apresentadas: *1*) representação gráfica de domínios conservados identificados na sequência; *2*) representação gráfica de *matches*, indicando qualidade do alinhamento e cobertura das sequências identificadas; *3*) informações estatísticas dos resultados encontrados, incluindo identidade e valor *e*; *4*) alinhamento de cada sequência encontrada com a sequência de busca (*query*).

de dados.

vii. *megablast*: para empregar um grande número de sequências de busca. Quando se compara um grande número de sequências de busca (especialmente no BLAST por linha de comando), o megablast é muito mais rápido que o BLAST executado por várias vezes seguidas. Ele agrupa muitas sequências de busca, formando uma grande sequência, antes de realizar a busca no banco de dados. Os resultados são pós-analisados em busca de alinhamentos individuais.

3.8. Significância estatística

Em determinados casos, especialmente para buscar evidência de homologia entre sequências, o alinhamento é analisado sob o ponto de vista estatístico. Nessa óptica, podemos calcular quão bom pode ser um ali-

nhamento simplesmente levando em consideração as razões de chance de alinhamento entre nucleotídeos quaisquer. Para isso, sequências de nucleotídeos ou aminoácidos são geradas aleatoriamente, alinhadas em conjunto e avaliadas, segundo um determinado esquema de pontuação. Para alinhamentos globais, pouco se sabe a respeito destas distribuições randômicas. No entanto, felizmente, estas técnicas são bem entendidas para casos de alinhamentos locais e, atualmente, são amplamente utilizadas para a avaliação de similaridade, especialmente em bancos de dados que comportam grande quantidade de sequências.

Para analisar a probabilidade associada a determinado alinhamento é necessário, inicialmente, gerar um modelo aleatório das sequências em análise. Esses novos alinhamentos serão pontuados seguindo um determinado esquema de pontuação. Neste contexto, será calculada a probabilidade de se obter aleatoriamente uma pontuação pelo menos igual à pontuação do alinhamento original. O valor associado aos múltiplos testes realizados é chamado de valor e (*e-value*). Para banco de dados, este valor corresponde ao número de distintos alinhamentos, com uma pontuação igual ou melhor, que são esperados ocorrer na busca por sequências similares simplesmente por razões de chance (aleatórios). Estes cálculos estatísticos levam em consideração a pontuação do alinhamento e o tamanho do banco de dados. Quanto menor o valor e, menor o número de chances de uma determinada sequência ser alinhada aleatoriamente com outras e, portanto, mais significante é o resultado. Por exemplo, um valor e de 1e-3 ($1x10^{-3}$ ou 0,001) significa que há a chance de 0,001 de que a sequência alvo seja alinhada com uma sequência aleatória do banco de dados. Por exemplo, em um banco de dados que contém 10.000 sequências, neste caso, esperaríamos encontrar até 10 outras sequências que alinharão significativamente com a sequência alvo. É importante ressaltar que o fato de encontrarmos um valor e próximo de zero na comparação entre duas sequências não necessariamente denota a homologia destas sequências, dado que sequências não relacionadas podem conter similaridades devido à evolução convergente.

3.9. Alinhamento de 2 estruturas

O alinhamento de estruturas é um problema matematicamente complexo que só pode ser resolvido por algoritmos heurísticos. A Figura 14-3 apresenta um exemplo de alinhamento estrutural simples. Diferentes algoritmos oferecem resultados diferentes para o alinhamento, e algumas vezes essas diferenças são grandes. Por esse motivo é importante testar diferentes programas de alinhamento estrutural. Cada um deles tem pontos fortes e fracos, que podem ser explorados a partir da leitura dos artigos que os propuseram originalmente.

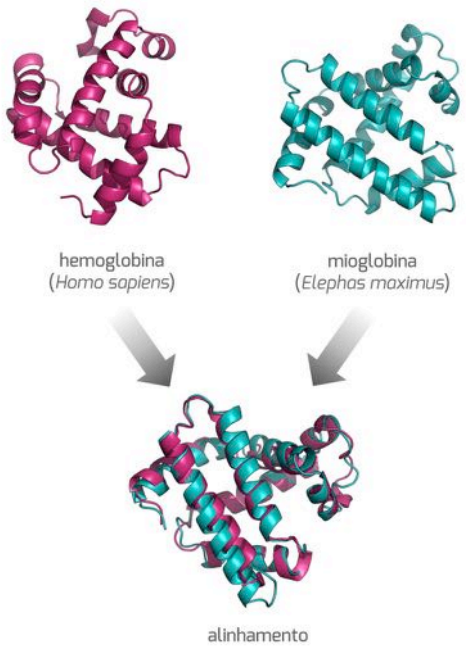


Figura 14-3: Exemplo de alinhamento de duas estruturas proteicas, oriundas de diferentes organismos: hemoglobina humana e mioglobina de elefante-asiático.

Existem três etapas essenciais para as diferentes estratégias de alinhamento estrutural: a representação, a otimização e a pontuação. A representação se refere às maneiras de representar as estruturas de uma forma que não seja dependente de coordenadas espaciais e que seja adequada ao ali-

nhamento. A otimização lida com a amostragem do espaço de possíveis soluções para o alinhamento entre as estruturas. A pontuação lida com a classificação dos resultados obtidos e com sua significância estatística. A seguir apresentamos as características específicas de alguns dos métodos mais utilizados para o alinhamento de duas estruturas.

DALI: emprega matrizes de distâncias para representar as estruturas, transformando as estruturas 3D em conjuntos 2D de distâncias entre Cα. Se imaginarmos a sobreposição das matrizes, as regiões de sobreposição na diagonal representam similaridades na estrutura 2ária (similaridades no esqueleto polipeptídico), e similaridades fora da diagonal representam similaridades na estrutura 3ária. As matrizes são então divididas em matrizes menores, de tamanho fixo, com base nas similaridades encontradas. Cada submatriz é unida a outras que sejam adjacentes para obter a matriz de sobreposição com maior abrangência. A significância estatística do alinhamento é calculada com base na distribuição encontrada em uma comparação de centenas de estruturas de baixa identidade. A pontuação é apresentada como número de desvios-padrão em relação a tal distribuição.

SSAP: cria vetores ligando resíduos a partir dos Cβ, representando a estrutura em duas dimensões, considerando posição e direção. Um algoritmo de programação dinâmica identifica similaridades entre as matrizes de vetores, gerando uma nova matriz que é posteriormente recalculada considerando as diferenças entre cada posição de similaridade encontrada na primeira etapa em relação às outras posições de similaridade, até que uma matriz ótima seja atingida. A pontuação do SSAP não é estatística, mas foi calibrada em relação ao banco de dados CATH. Assim, uma pontuação maior que 70 indica similaridade entre as estruturas comparadas.

VAST: cria vetores a partir de elementos de estrutura 2ária cujo tipo, direção e conexão estão relacionados com a topologia da proteína. Esses elementos (fragmentos) de estrutura 2ária são alinhados e comparados com alinhamentos gerados aleatoriamente. Alinhamentos com boa pontuação são agrupados e depois realinhados usando um procedimento de otimização por Monte Carlo. A significância estatística é dada pelo valor p (assim como ocorre no BLAST). O valor p é proporcional à probabilidade de se obter o alinhamento ao acaso.

SARF2: transforma as coordenadas em um conjunto de elementos de estrutura 2ária. Posteriormente, avalia pares desses elementos comparando o ângulo entre eles, a menor distância entre seus eixos e as distâncias mínimas e máximas entre cada elemento e a linha média. Um otimizador baseado em grafos é empregado para obter o maior número de conjuntos mutuamente compatíveis, e então o alinhamento final é calculado por adição de mais resíduos até que um valor mínimo de RMSD, definido pelo usuário, seja atingido. A pontuação final do alinhamento é calculada como função do RMSD e do número de Cα pareados entre as estruturas. A significância estatística é obtida por comparação à distribuição de pontuações obtidas pelo alinhamento da proteína leghemoglobina a centenas de estruturas não redundantes.

CE: representa as proteínas como conjuntos de distâncias entre Cα de oito resíduos consecutivos na estrutura. Primeiramente, são identificados todos os pares de octâmeros compatíveis entre as estruturas. Posteriormente, um algoritmo de extensão combinatória identifica e combina os pares mais similares entre as estruturas, adicionando mais pares a cada etapa do cálculo até a obtenção do melhor alinhamento. A significância estatística é dada por comparação às pontuações obtidas em um conjunto de alinhamentos entre estruturas com menos de 25% de identidade de sequência.

MAMMOTH: transforma as coordenadas da proteína em um conjunto de vetores unitários a partir dos Cα de heptâmeros consecutivos. A similaridade entre heptâmeros é calculada pela sobreposição de seus vetores, a matriz de similaridade ótima é identificada e então o melhor alinhamento local entre estruturas é identificado dentro de um valor de RMSD pré-definido. A significância estatística é dada pelo valor p, baseado na comparação com a pontuação de alinhamentos obtidos aleatoriamente.

SALIGN: representa as proteínas por um conjunto de propriedades ou características calculadas a partir da sequência e da estrutura ou definidas arbitrariamente pelo usuário. Tais propriedades incluem tipo de resíduo, distância entre resíduos, acessibilidade da cadeia lateral, estrutura 2ária, conformação local da estrutura e característica a ser definida pelo usuário. O programa calcula uma matriz de dissimilaridade entre propriedades equivalentes, e a pontuação da dissimilaridade é calculada pela soma das matrizes de cada característica. A melhor sobreposição de matrizes é

obtida por um algoritmo baseado em programação dinâmica. A significância estatística não é calculada pelo SALIGN e o usuário obtém apenas os valores da pontuação de dissimilaridade. O programa fornece, entretanto, um valor adicional de qualidade, apresentado como porcentagem de Cα cuja distância é menor que 3,5 Å entre os pares de estruturas alinhadas.

3.10. Alinhamento de >2 estruturas

A maior parte dos métodos disponíveis para o alinhamento múltiplo de estruturas inicia-se estabelecendo todos os alinhamentos entre pares de estruturas e, então, emprega-os para estabelecer um alinhamento consenso entre todas as estruturas. A Figura 15-3 apresenta um exemplo de alinhamento estrutural múltiplo. Os métodos para obter o alinhamento consenso variam entre os programas de alinhamento. A seguir apresentamos as características específicas de alguns dos métodos mais utilizados para o alinhamento de estruturas múltiplo.

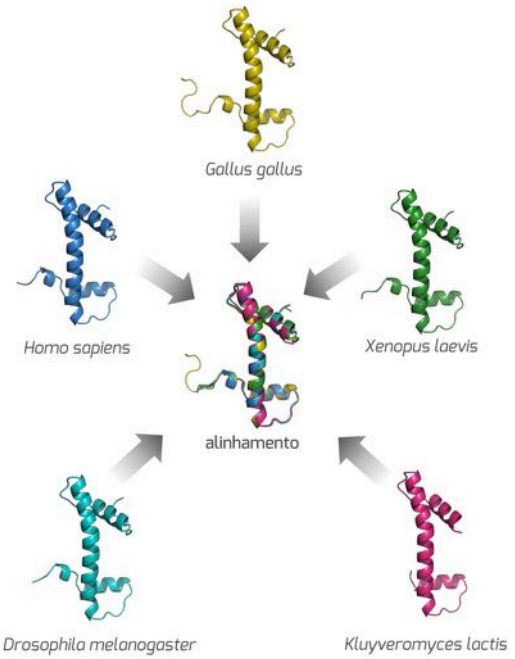


Figura 15-3: Exemplo de alinhamento de múltiplas estruturas proteicas, oriundas de diferentes organismos (histonas H3 de levedura, mosca-da-fruta, homem, frango, sapo-de-garras).

CE-MC: realiza o refinamento de um conjunto de alinhamentos de pares de estruturas empregando uma técnica de otimização de Monte Carlo. O algoritmo modifica o alinhamento múltiplo aleatoriamente, e as modificações são aceitas se houver melhoria na pontuação do alinhamento. O processo encerra quando o alinhamento múltiplo não puder mais ser melhorado por modificações aleatórias.

MAMMOTH-Mult: essa extensão do MAMMOTH gera inicialmente todos os alinhamentos de estruturas aos pares. Um procedimento de organização por médias é empregado para agrupar as estruturas com base em suas similaridades aos pares, gerando uma árvore. O alinhamento múltiplo é gerado por reorganização dessa árvore, onde ramos similares vão sendo agrupados aos pares, iterativamente.

SALIGN: pode realizar alinhamentos múltiplos de duas maneiras, baseado em uma árvore ou por alinhamento progressivo. O primeiro caso é muito similar ao MAMMOTH-Mult. No alinhamento progressivo, as estruturas são alinhadas na ordem em que são fornecidas para o programa. A vantagem desse método é o de seu custo computacional ser menor que o do método baseado em uma árvore.

3.11. Alinhamento flexível

O alinhamento de estruturas considerando sua flexibilidade está se tornando cada vez mais importante devido à melhor compreensão do enovelamento proteico. Cada vez mais, percebe-se que não existem enovelamentos estanques, mas sim um gradiente densamente populado por variantes conformacionais. Desta forma, torna-se mais difícil definir domínios proteicos, sendo mais adequado descrever as estruturas como conjuntos de estruturas supra-secundárias. Com base nessa proposta, a diferença entre proteínas relacionadas reside na orientação relativa desses subdomínios. A Figura 16-3 demonstra as diferenças que podem ser observadas ao alinhar um par de estruturas de maneira rígida ou flexível. A seguir apresentamos as características específicas de alguns dos métodos mais utilizados para este tipo de alinhamento de estruturas.

FATCAT: o algoritmo adiciona “torções” entre pares de fragmentos proteicos alinhados, que são tratados

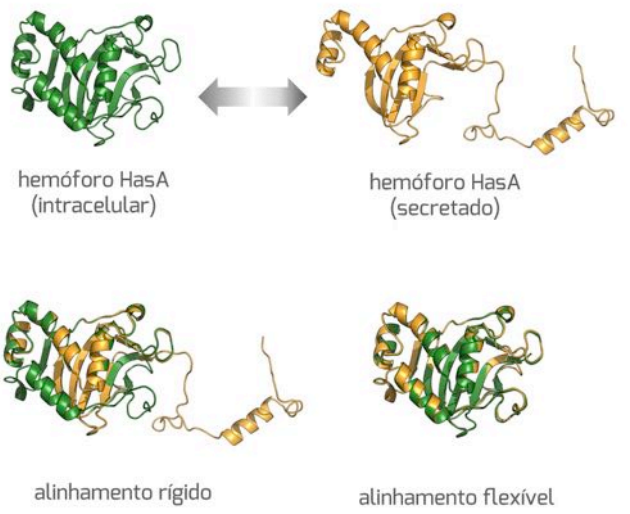


Figura 16-3: Comparação entre alinhamento estrutural rígido e flexível. A estrutura da proteína HasA (um captador bacteriano de grupamentos heme) foi obtida para suas formas intra- e extra-celular. Observe que o alinhamento rígido identifica similaridade parcial entre as estruturas, enquanto o alinhamento flexível detecta o rearranjo espacial de parte da proteína, evidenciando sua identidade.

como corpos rígidos. De maneira geral, o programa permite a inclusão dessas torções quando elas diminuem o valor final do RMSD, refletindo em um melhor alinhamento estrutural. O alinhamento final é obtido por programação dinâmica e se baseia na matriz de similaridade entre os fragmentos pareados, obtidos na primeira etapa do cálculo.

FLEXPROT: mantém uma das proteínas rígida, enquanto a outra pode sofrer alterações em busca de maior similaridade estrutural. As regiões potencialmente flexíveis da proteína são detectadas automaticamente e empregadas nas alterações conformacionais.

ALADYN: alinha pares de estruturas com base em sua dinâmica interna e similaridade entre seus movimentos de grande escala. O posicionamento ótimo entre as proteínas é encontrado ao maximizar as similaridades entre os padrões de flutuação estrutural, que são calculados pelo modelo de redes elásticas.

POSA: uma variante do FATCAT para o alinhamento múltiplo flexível de estruturas. Emprega uma metodologia combinada, introduzindo grafos de ordem parcial para visualizar e agrupar regiões similares entre as estruturas.

3.12. Conceitos-chave

Algoritmo: sequência lógica de instruções necessárias para executar uma tarefa.

Alinhamento: método de organização de sequências ou estruturas biológicas para evidenciar regiões similares e dissimilares. Estes métodos estão geralmente atrelados a inferências funcionais ou evolutivas.

Alinhamento Múltiplo: alinhamento que envolve mais de duas sequências ou estruturas

Alinhamento Simples: alinhamento que envolve apenas duas sequências ou estruturas.

BLAST: *Basic Local Alignment Search Tool* (Ferramenta de Busca por Alinhamento Local Básico), empregado para buscar sequências em bancos de dados com base em sua similaridade.

Homologia: é um termo essencialmente qualitativo que denota uma ancestralidade comum de determinada sequência.

HSP: pares de segmentos de alta pontuação (*high-scoring segment pairs*), zonas de similaridade entre sequências identificadas pelo BLAST.

Identidade: Porcentagem de caracteres similares entre duas sequências (excluindo-se as lacunas).

Indels: identifica inserções e deleções de caracteres ao longo do processo evolutivo.

Lacunas: regiões identificadas por hifens que representam a inserção/deleção de caracteres ao longo do processo evolutivo.

Matches: regiões que apresentam caracteres idênticos entre diferentes sequências.

Mismatches: regiões que apresentam caracteres não idênticos entre diferentes sequências.

Penalidades por lacuna (PL): conjunto de parâmetros necessários para atribuir a pontuação para uma lacuna em um sistema de alinhamento por pontuação.

RMSD: desvio médio quadrático.

Tradução: tradução (*in silico*) de uma sequência de mRNA em sua possível sequência proteica correspondente

3.13. Leitura recomendada

BOGUSKI, Mark S. A molecular biologist visits Jurassic Park. ***Biotechniques***, 12, 668-669, 1992.

CARUGO, Oliviero. Recent progress in measuring structural similarity between proteins. ***Curr. Protein. Pept. Sci.***, 8, 219-241, 2007.

MADDEN, Tom. The BLAST sequence analysis tool. In: McENTYRE, Jo; OSTELL, Jim (Org.). ***The NCBI Handbook***. Bethesda: National Center for Biotechnology Information, 2002.

MARTI-RENOM, Marc A.; et al. Structure comparison and alignment. In: GU, Jenny; BOURNE, Philip E. (Org.). ***Structural Bioinformatics***. 2.ed. Hoboken: John Wiley & Sons, 2009.

MAYR, Gabriele; DOMINGUES, Francisco S.; LACKNER, Peter. Comparative analysis of protein structure alignments. ***BMC Struct. Biol***., 7, 50, 2007.

MOUNT, David W. ***Bioinformatics: Sequence and Genome Analysis***. 2.ed. Cold Spring Harbor: Cold Spring Harbor Laboratory Press, 2004.

ROSSMANN, Michael G.; ARGOS, Patrick. The taxonomy of binding sites in proteins. ***Mol. Cell. Biochem.***, 21, 161-182, 1978.