Algoritmos para Análise de Sequências Biológicas

Blast

Sumário

- Algoritmos e ferramentas para procurar sequências similares em bases de dados de grande dimensão
- Descrever o algoritmo Blast
- Descrever, analisar e implementar uma simplificação do Blast

Blast AASB 2 / 10

Métodos heurísticos

- Não garantem solução óptima.
- São **mais rápidos** que os algoritmos de PD, na ordem das 50-100 vezes.
- Se as sequências a comparar forem pouco similares, o ideal será utilizar a PD pois tem maior sensibilidade.
- Ideais para procuras em BDs, onde se tem uma sequência e se procuram sequências similares em conjuntos de elevada cardinalidade.
- Métodos mais conhecidos: FASTA e BLAST.

Blast AASB 3/10

Critérios de avaliação

- Sensibilidade % das sequências homólogas na BD que são retornadas Precisão % das sequências similares retornadas que são homólogas Eficiência tempo de resposta a um pedido
 - VP verdadeiros positivos (sequências homólogas efetivamente detetadas)
 - FP falsos positivos (sequências detectadas que não são homólogas)
 - FN falsos negativos (sequências homólogas não detetadas)

$$Sensibilidade = \frac{VP}{FN + VP}$$

$$Precis$$
ã $o = VPVP + FP$

Blast AASB 4/10

BLAST – Basic Local Alignment Search Tool

- Algoritmo mais utilizado da atualidade na pesquisa em BDs de seguências.
- Procura bons alinhamentos locais entre uma sequência query e sequências de uma base de dados definida
- Usa pequenas "palavras" (e.g. 3 AAs ou 5-15 bases de DNA);
- Procura palavras do mesmo tamanho nas sequências da BD, que comparadas com as palavras da sequência query tenham alta similaridade.
- Os matches comuns formam a base de um alinhamento local que é posteriormente estendido nas duas direções
- A extensão ocorre até o alinhamento baixar de um score pré-definido

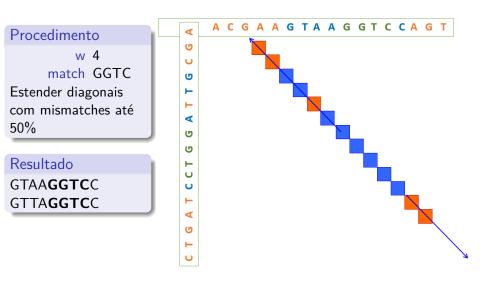
Blast AASB 5 / 10

Fases do Blast

- Remover zonas de pouca complexidade (e.g. repetições) que podem comprometer o alinhamento
- ② Obter todas as palavras de tamanho w da query
- **9** Para cada palavra, compilar lista de todas as palavras possíveis de tamanho w cujos scores, sejam maiores do que um dado limite T (e.g. para proteínas tipicamente T=13)
- Organizar as palavras identificadas de forma eficiente numa árvore de procura
- Procurar em cada sequência da BD todos os hits com as palavras recolhidas
- Estender os hits obtidos no passo 5 em ambas as direções enquanto o score for aumentando (ou mantendo-se acima de um dado limite)
- Escolher os alinhamentos de 6 com maior score normalizando para o seu tamanho (high scoring pairs - HSPs)
- Calcular a significância dos melhores HSPs (calculando E-value)

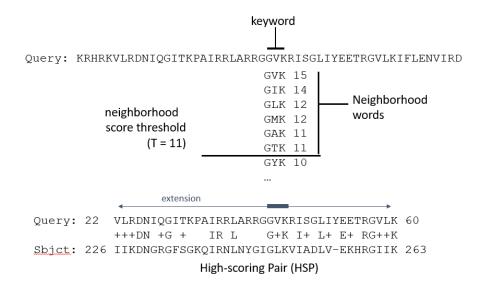
Blast AASB 6/10

Exemplo



Blast AASE 7/10

Exemplo com proteínas



Blast AASB 8/10

Refinamentos

- Dois hits próximos independentes acima de T, separados por uma distância não superior a um parâmetro dado
- Isto leva a menos extensões e logo a mais rapidez no algoritmo;
 resultados obtidos não pioram
- Esta estratégia permite incluir gaps na zona estendida entre os dois hits
- Diminuir os valores dos parâmetros w e T pode aumentar a sensibilidade mas também aumentam o tempo de processamento
- Outra melhoria: vários HSPs podem ser combinados para gerar alinhamentos maiores e de melhor qualidade; algoritmos de PD podem ser usados nesta tarefa

Blast AASB 9/10

Significância estatística do alinhamento

- Funções de mérito dão resultados relativos que não servem para avaliar a qualidade do alinhamento e a similaridade das sequências.
- Programas anteriores calculam medidas estatísticas para dar indicações sobre a qualidade dos alinhamentos.
- Medida mais importante: E indica o nº de HSPs com o mesmo score esperado usando uma quantidade de sequências aleatórias igual ao tamanho da BD.
- Valores de E muito **próximos de zero** indicam grande similaridade.
- Valores de E significativos tipicamente < 0,05.
- Para valores pequenos o valor de E é semelhante ao p-value
- ullet E = 1, significa que um match por acaso terá score similar ao obtido

Blast AASB 10 / 10