

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

Métodos heurísticos

- **Não** garantem solução ótima.
- 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.

Cr terios de avalia  o

- Sensibilidade** % das seq ncias hom logas na BD que s o retornadas
- Precis o** % das seq ncias similares retornadas que s o hom logas
- Efici ncia** tempo de resposta a um pedido
- VP** verdadeiros positivos (seq ncias hom logas efetivamente detectadas)
- FP** falsos positivos (seq ncias detectadas que n o s o hom logas)
- FN** falsos negativos (seq ncias hom logas n o detectadas)

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

$$Precis o = \frac{VP}{VP + FP}$$

BLAST – Basic Local Alignment Search Tool

- Algoritmo mais utilizado da atualidade na pesquisa em BDs de sequências.
- 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

Fases do Blast

- ➊ Remover zonas de pouca complexidade (e.g. repetições) que podem comprometer o alinhamento
- ➋ Obter todas as palavras de tamanho w da query
- ➌ 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 $T = 13$)
- ➍ Organizar as palavras identificadas de forma eficiente numa árvore de procura
- ➎ Procurar em cada sequência da BD todos os hits com as palavras recolhidas
- ➏ Estender os hits obtidos no passo 5 em ambas as direções enquanto o score for aumentando (ou mantendo-se acima de um dado limite)
- ➐ Escolher os alinhamentos de 6 com maior score normalizando para o seu tamanho (high scoring pairs - HSPs)
- ➑ Calcular a significância dos melhores HSPs (calculando E-value)

Exemplo com proteínas

keyword

Query: KRHRKVLRDNIQGITKPAIRRLARRGGVKRISGLIYEETRGVLKIFLENVIRD

neighborhood
score threshold
(T = 11)

GVK 15

GIK 14

GLK 12

GMK 12

GAK 11

GTK 11

GYK 10

...

Neighborhood
words

extension

Query: 22 VLRDNIQGITKPAIRRLARRGGVKRISGLIYEETRGVLK 60
+++DN +G + IR L G+K I+ L+ E+ RG++K
Sbjct: 226 IIKDNNGRGFSGKQIRNLNYGIGLKVIADLV-EKHRGIK 263

High-scoring Pair (HSP)

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

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
- $E = 1$, significa que um match por acaso terá score similar ao obtido