

Bioinformática – Visão Geral

Anotação de Genomas

- ⇒ Genes localizados → anotamos proteínas
- ⇒ Projetos transcriptoma → genes expressos (diferentes fases)
- ⇒ Vias bioquímicas (metabólicas) completas e incompletas \leftrightarrow ambiente
- ⇒ Mostrar mapa KEGG

Bioinformática – Visão Geral

Pós-genômica

- ⇒ Estudo das proteínas, suas estruturas e funções, vias metabólicas, desenvolvimento de drogas, mapas de interação
- ⇒ Análise de expressão: DNA → mRNA → proteína
- ⇒ Capturamos mRNA → fácil em eucariotos, pois usamos oligonucleotídeos (DNA com até 20 bp) poli-T que aderem a cauda poli-A + transcriptase reversa → cDNA
- ⇒ Técnicas EST e SAGE
- ⇒ SAGE: usa fragmentos de cDNA de 10 a 17 nucleots. chamados de SAGETags

Bioinformática – Visão Geral

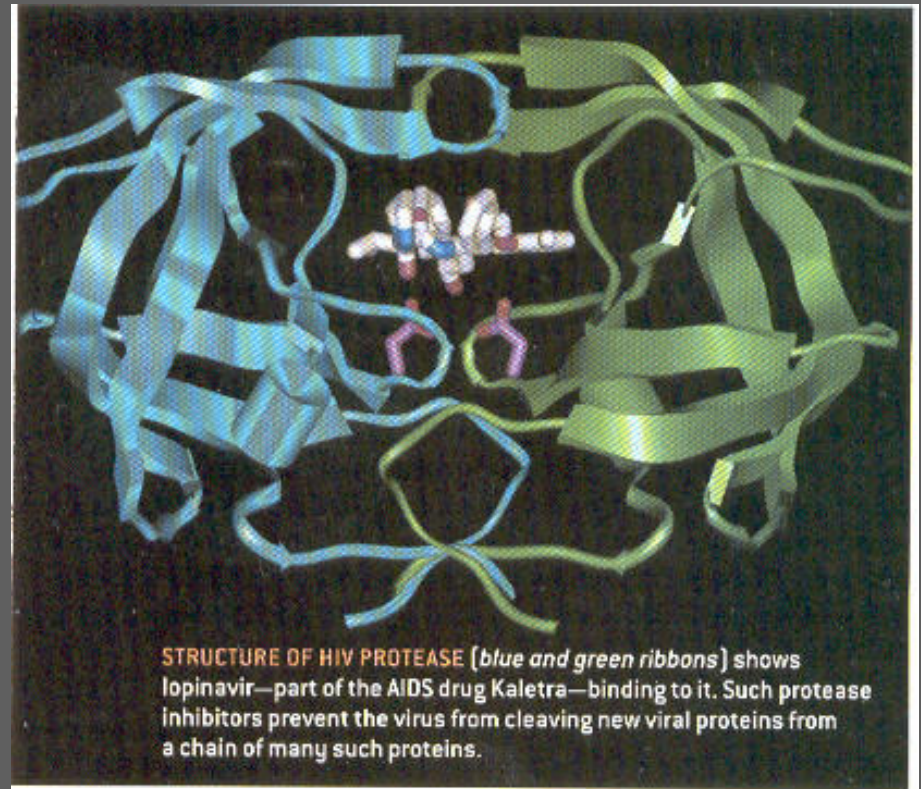
Pós-genômica

- ⇒ Chips de DNA: oligonucleotídeos
- ⇒ Microarrays: fragmentos de cDNA
- ⇒ A uma lâmina de vidro são presos os fragmentos de cDNA's conhecidos
- ⇒ mRNA's "hibridam-se" aos oligos no chip
- ⇒ Populações de mRNA's obtidos de células em diferentes condições (ex: normal vs. cancerosa) são jogadas nos chips
- ⇒ Os mRNA's são marcados com corantes fluorescentes, um verde e outro vermelho
- ⇒ A cor verde ou vermelha em cada spot do chip indica qual gene está sendo mais transcrito em cada estado celular
- ⇒ Cor amarela indica que o gene é igualmente transcrito
- ⇒ A intensidade da cor indica a intensidade do nível de expressão

Bioinformática – Visão Geral

Pós-genômica

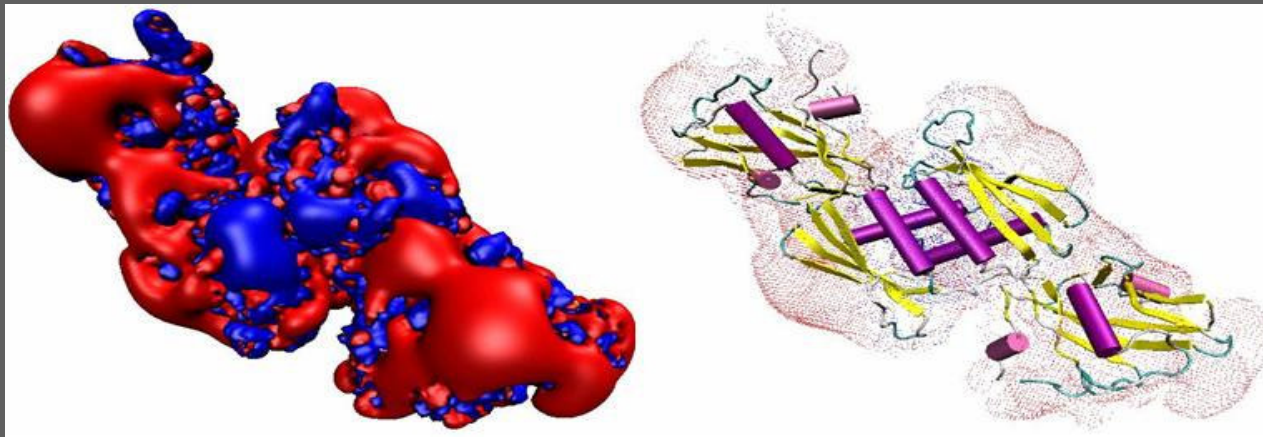
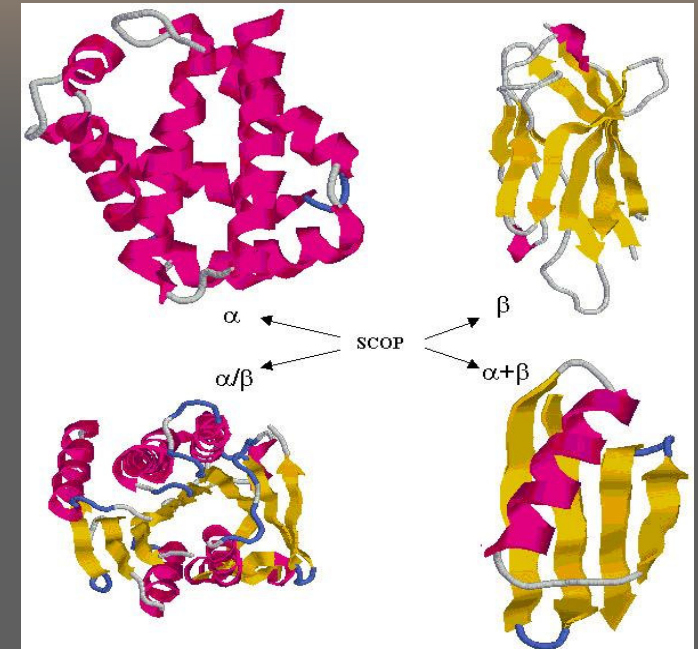
- ⇒ Nem sempre a quantidade de mRNA corresponde a de proteína
- ⇒ Analisa-se das proteínas expressas (projetos proteoma que são experimentais)
- ⇒ Cristalografia por raios-X
- ⇒ RMN



Bioinformática – Visão Geral

Pós-genômica

- Modelagem molecular
- Predição da classe SCOP de proteínas

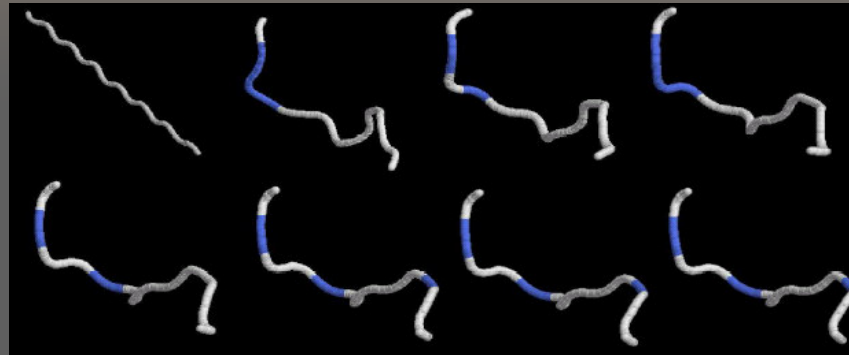


← Poisson-Boltzmann

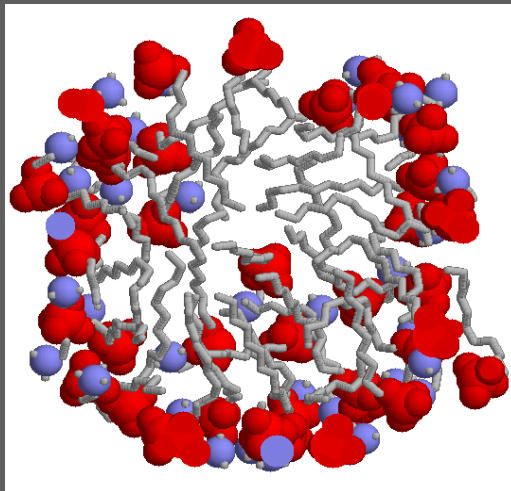
Bioinformática – Visão Geral

Pós-genômica

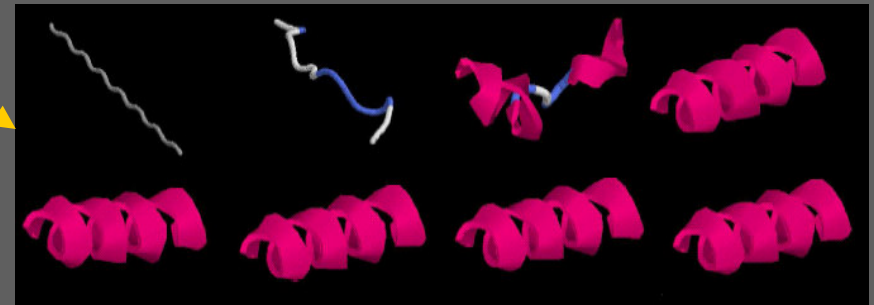
- Dinâmica molecular
+
▪ Rede neural



← Dinâmica
molecular



Sistema
híbrido

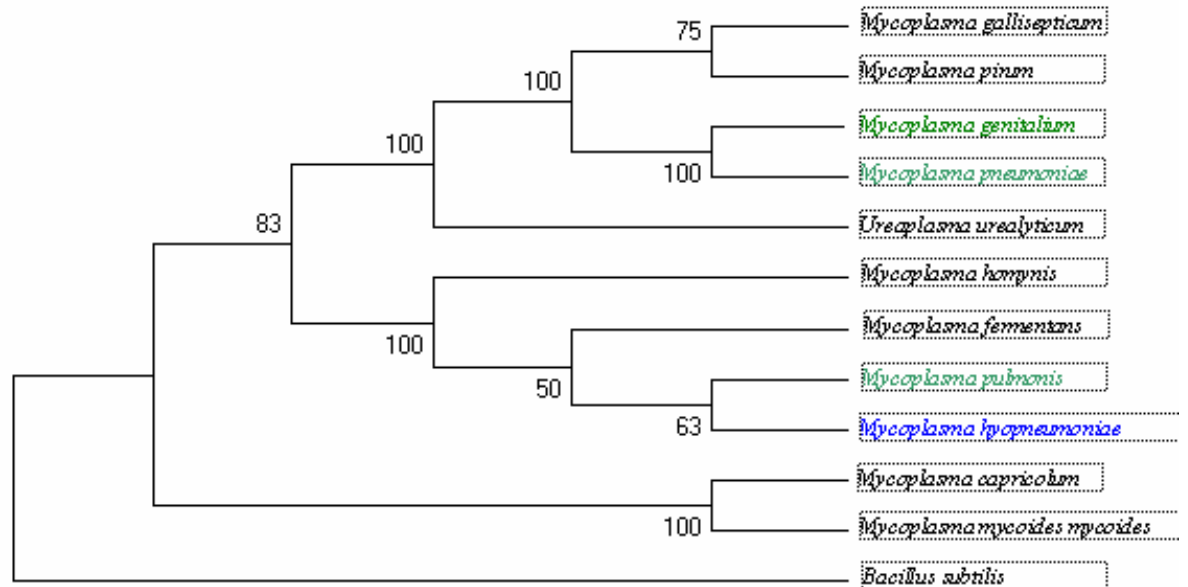


- Simulação com software GROMACS

Bioinformática – Visão Geral

Filogenética

- ⇒ Usar DNA para “parentesco” entre espécies
- ⇒ Parentesco entre espécies associando escala temporal → filogenia
- ⇒ Representação através de árvores



Bioinformática – Visão Geral

Filogenética

- ⇒ Usa-se seqs. homólogas (relacionadas evolutivamente)
- ⇒ Escolhidas as seqs., faz-se o alinhamento múltiplo associado a um método de distância: máxima parsimônia ou máxima verossimilhança
- ⇒ Exemplo de medida de distância: Distância p → número de sítios variáveis entre duas seqs em relação ao total de sítios comparados
- ⇒ Máxima parcimônia: a melhor hipótese explicativa é aquela que requer o menor número de passos, isto é a árvore que possuir o menor núm. de substituições é a mais próxima da real
- ⇒ Na MP não há cálculo de distância
- ⇒ O número de possibilidades de árvores faz com que seja lento

Bioinformática – Visão Geral

Filogenética

- ⇒ Máxima verossimilhança: busca-se uma árvore que maximize a probabilidade dos dados observados
- ⇒ Busca diferentes topologias e com variações nos tamanhos dos ramos
- ⇒ Muito mais lento
- ⇒ A cada rodada do programa uma árvore diferente pode ser gerada
- ⇒ Os ramos e comprimentos mais freqüentes são utilizados para a construção da árvore
- ⇒ Usa-se o método estatístico bootstrap que gera conjuntos modificados aleatoriamente
- ⇒ As árvores são comparadas e as probabilidades indicadas
- ⇒ Software:
- ⇒ Clustal: www.ebi.ac.uk/clustalw
- ⇒ Phylip: <http://evolution.genetics.washington.edu/phylip.html>
- ⇒ Mega: www.megasoftware.net

Modelo de dados do NCBI

- ⇒ Representação de seqs segmentadas:
- ⇒ NCBI usa o padrão Abstract Syntax Notation 1 (ASN.1)
- ⇒ O modelo é pensado p/ facilitar a descoberta de conhecimento pela conexão de informações que são armazenadas separadamente ou pelo processamento destas informações
- ⇒ NCBI usa 4 elementos para dados: citações bibliográficas, seqs de DNA, seqs de proteínas e estruturas tridimensionais

Modelo de dados do NCBI

Tipos de citações bibliográficas:

- ⇒ Artigos científicos → PubMed Central
- ⇒ Identificador (núm. Inteiro) → PMID (MUID, MEDLINE)
- ⇒ Patentes de seqs → contém menos informações
- ⇒ Submissão de seqs

Modelo de dados do NCBI

O que há no nome de uma seqüência?

Classe de objetos do modelo: Seq-id

- ⇒ Locus: caindo em desuso por conter inf. biológica (3 letras do nome do organismo e o nome do gene)
- ⇒ Núm. de acesso: criado por DDBJ/EMBL/GenBank para auxiliar na identificação de uma seqüência. Não contém informação biológica. 2 letras e 6 dígitos, letras estão associadas ao banco que forneceu a seq. Ex: U → GenBank
- ⇒ Núm. gi: (GenInfo) identificador único p/ seqs em muitas fontes, os outros acima não garantem isso, pois as seqs são atualizadas. Todo processamento interno do NCBI é baseado neste identificador

Modelo de dados do NCBI

⇒ Identificador ACESSO/VERSÃO combinados: é esperado ser o melhor identificador. A versão representa mudanças na seq.

Números de acesso em seqs de proteínas:

⇒ também recebem o mesmo tipo de identificador acesso/versão, mas com 3 letras, 5 dígitos e a versão

GenBank e EMBL

- ⇒ GenBank ou EMBL flatfile é uma unidade elementar de inf. nestes bancos. É o formato para troca de informação entre os bancos principais
- ⇒ Exercício: obter para os genes abaixo, a partir do EMBL e GenBank, a seq. de nucleotídeos e seu tamanho, a seq. de aminoácidos, o id de acesso/versão, o id da proteína, o tipo de molécula de onde se obteve a seq., o artigo científico (PDF) e a sua estrutura espacial:
- ⇒ Myosin heavy chain IIx/d (Homo sapiens), Dopachrome tautomerase precursor (Homo sapiens), cytochrome b (E. coli), telomerase reverse transcriptase 1 (S. pombe), shikimate dehydrogenase (H. influenza)