

Ferramentas de Bioinformática para Sequenciamento e Anotação

Ana Lúcia C. Bazzan

Universidade Fed. do Rio Grande do Sul

Instituto de Informática

{bazzan@inf.ufrgs.br}

Roteiro

- Bioinformática e Biologia Computacional
 - características e principais objetivos
- Biologia Molecular: uma introdução
- Projetos Genoma
- Ferramentas de Bioinformática
 - sequenciamento
 - anotação

O que é Bioinformática?

- Bioinformática
 - uso de computadores em aplicações da área de biologia (tecnologia e infra-estrutura computacional)
 - análise de dados provenientes do genoma (elucidar processos biológicos complexos)
 - diagnósticos de doenças, desenvolvimento de novos fármacos
- Biologia computacional
 - estudo de sistemas moleculares naturais e artificiais
 - novos paradigmas de computação baseada em DNA

Principais Tarefas

- Algumas tarefas associadas à bioinformática em projetos Genoma:
 - receber e armazenar sequências
 - montar o genoma
 - disponibilizar ferramentas para anotação do genoma
 - comunicação com repositórios de dados (incluindo envio)

O Negócio de Bioinformática

- mercado para o setor de bioinformática nos próximos 4 a 5 anos é de dois bilhões de dólares
- mais de 50 empresas; privadas e com ações na bolsa
- coleta e armazenamento de dados, pesquisas em banco de dados e sua interpretação
- acesso a bases de dados é cobrado; clientes grandes companhias farmacêuticas e de biotecnologia

Motivação

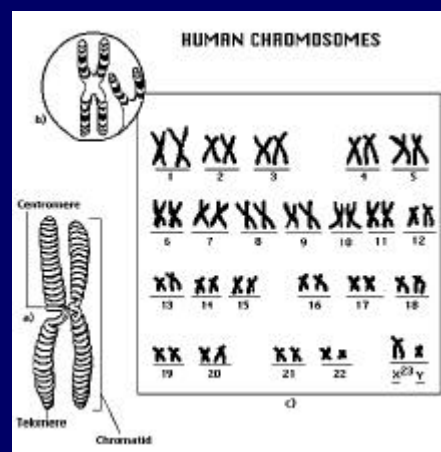
- encontrar mais rapidamente um fármaco para uma patologia específica; alguns aspectos do processo de descoberta científica e de desenvolvimento de fármacos migrarão para biologia *in silico*
- bioinformática tem o potencial de mudar a biologia e a medicina
- genoma sem bioinformática não conseguirá sobreviver face à concorrência, especialmente no mundo dos negócios associados à biotecnologia

Motivação

- volume de dados é de altíssima ordem
 - Incyte Genomics: 20 milhões de pares de bases de DNA / dia
 - Celera Genomics: afirma possuir 50 terabytes de dados armazenados (≈ 80 mil CDs)
 - apenas os bancos de dados que armazenam sequências!
 - adicionalmente: expressões gênicas (quando e onde um gene se expressa), diferenças genéticas entre indivíduos (*single-nucleotide polymorphisms* ou SNP), estruturas de proteínas, interação de proteínas, referências a publicações científicas (artigos, revistas, livros, etc.).

Uma Analogia

- genoma ≡ livro
- cromossomos ≡ capítulos
- genes ≡ histórias
- exons ≡ parágrafos (interrompidos por introns)
- codons ≡ palavras
- bases ≡ letras (A,C,G,T)



O Genoma Humano

- ~3 bilhões pares de bases (bp)
- 30K - 60K genes
- ~1 gene/~30kbp
- ~6 exons/gene
- ~150 bp/exon
- genes: onde estão e como se parecem?

Seleção e Evolução

- busca em espaço de características possíveis para os organismos
- espaço precisamente definido
 - 4 letras em uma dupla fita (2 sequências lineares)
 - tradução: complexa
- genótipo x fenótipo: codificação genética x características físicas
- busca é sobre espaço do genótipo
- seleção é sobre espaço do fenótipo

Biologia Molecular

- célula
 - procariotos: sem núcleo ou outras organelas; DNA circular => bactérias
 - eucariotos: núcleo e organelas (e.g. mitocôndria)
=> animais, plantas, leveduras, fungos
- célula somática: divisão em 4 fases; é especializada (e.g. 14 tipos em tecidos)
 - questão em aberto: COMO, QUANDO???
- célula reprodutiva: divide-se via meiose

Biologia Molecular

- composição da célula: núcleo, citoplasma, material genético e mecanismos para sua tradução em proteínas, membrana
 - membrana: lipídio (hidrofóbico) + grupo fosfatado (hidrofílico)
 - material genético: DNA ou RNA
 - proteína:
 - molécula responsável por maioria das funções da célula
 - estrutura primária: sequência de amino-ácidos (20 tipos)
 - cadeia: até 4500 a.a. => espaço de busca = 20^{4500}

DNA e RNA

- ácidos nucleicos
- polímeros formados a partir de nucleotídeos: adenina, citosina, guanina, timina
- dupla fita, hélice dupla
 - A com T (U no RNA)
 - C com G
- cada fita: cabeça (5') e cauda (3')
- *motifs*: padrões de amino-ácidos na sequência

Proteínas

- funções: estrutura da célula, enzimas, ativação/desativação de genes, sensores, atuadores, detectores (sistema imunológico)
- cada célula tem mesmo DNA; cada tipo de célula gera proteínas diferentes
- constituição: amino-ácidos
 - átomo C + grupo amino (NH_3) + grupo carboxil (COOH) + cadeia variável

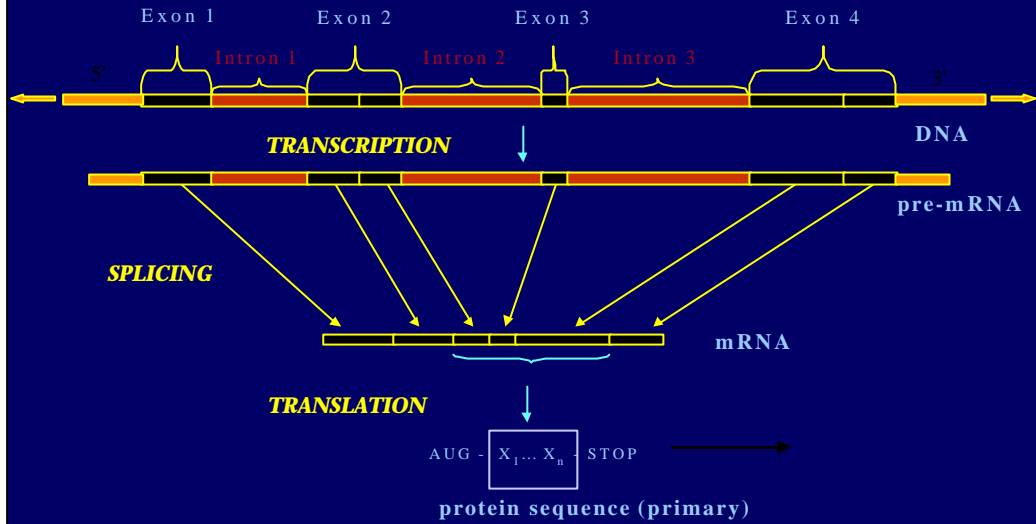
Proteínas

- estrutura primária: cadeia de amino-ácidos
 - 3 nucleotídeos formam um codon que especifica um amino-ácido (são 20 ao todo)
 - $3^4 = 64 \Rightarrow$ redundância
- estrutura secundária: arranjos locais (alfa hélices, fita beta, coils etc.)
- estrutura terciária: depende da posição dos átomos após o *folding*
- estrutura quaternária: partes ativas / inativas

O Dogma Central

- DNA - mRNA - tRNA - proteína
 - etapas:
 - transcrição de uma porção de DNA (via promotores) em RNA mensageiro (ligação de uma RNA polimerase com uma parte da molécula de DNA)
 - splicing: junta os exons; transporte para fora do núcleo
 - tradução: mRNA é usado como *blueprint* para produção de amino ácidos (no ribossomo) que constituem as proteínas
 - folding da proteína, transformações pós translacionais (elementos distintos podem mudar a forma e a atividade)
 - transporte da proteína para onde ela exerce sua função

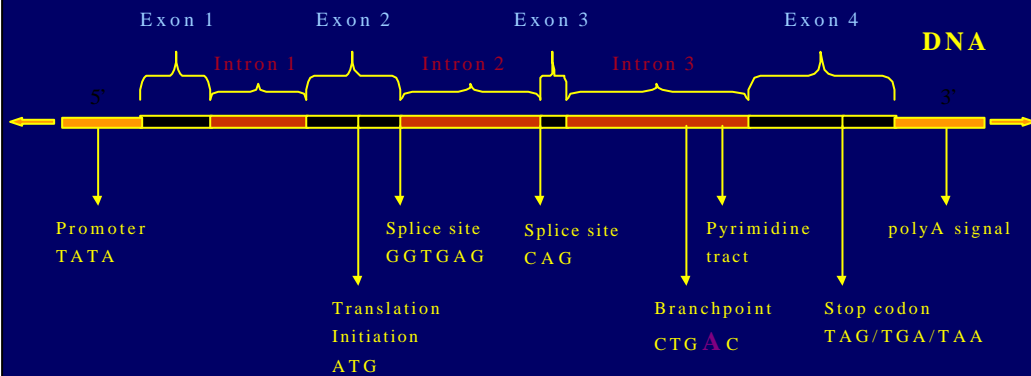
O Dogma Central



Ferramentas de Bioinformática para Sequenciamento e Anotação

Ana Lúcia C. Bazzan

Sinais em Genes

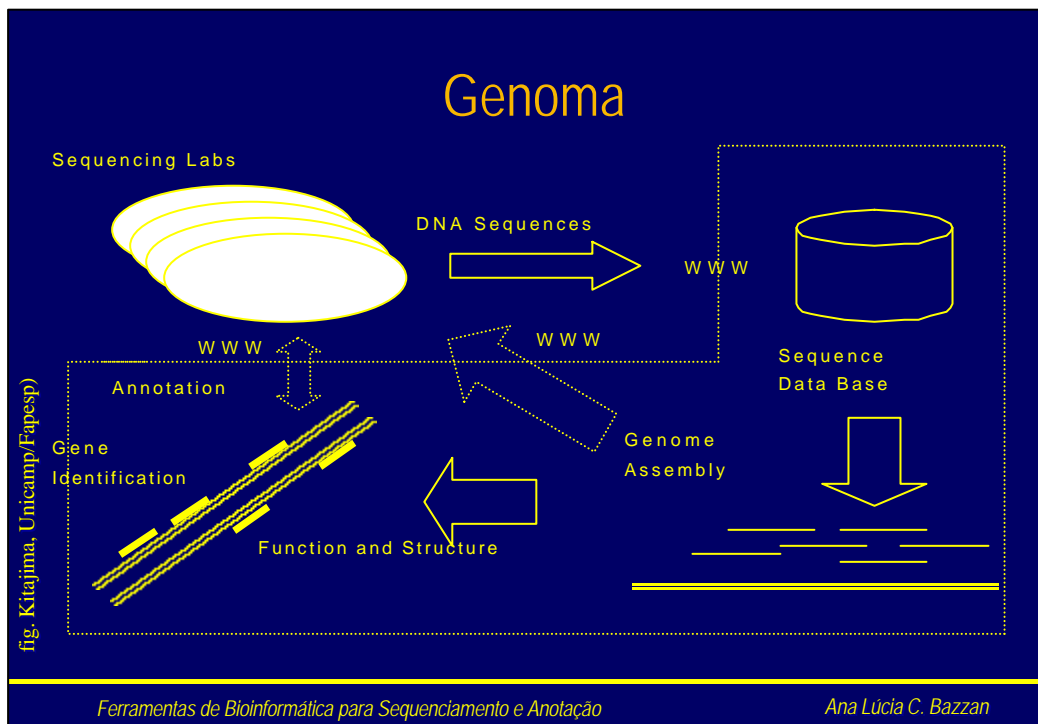


Ferramentas de Bioinformática para Sequenciamento e Anotação

Ana Lúcia C. Bazzan

Genoma, Transcriptoma, Proteoma

- Genoma
 - Sequenciar e montar o material genético (DNA) de um organismo
- Transcriptoma
 - Encontrar os genes expressos (EST)
- Proteoma
 - Encontrar as proteínas sintetizadas



Anotação e Submissão

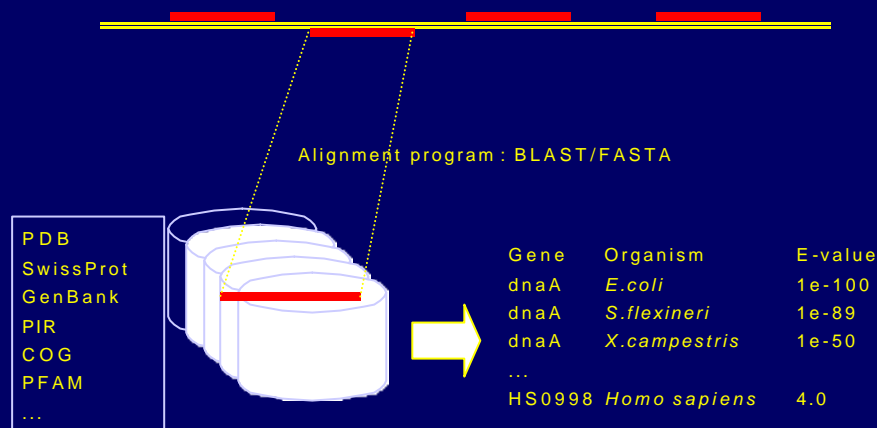


fig. Kitajima, Unicamp/Fapesp)

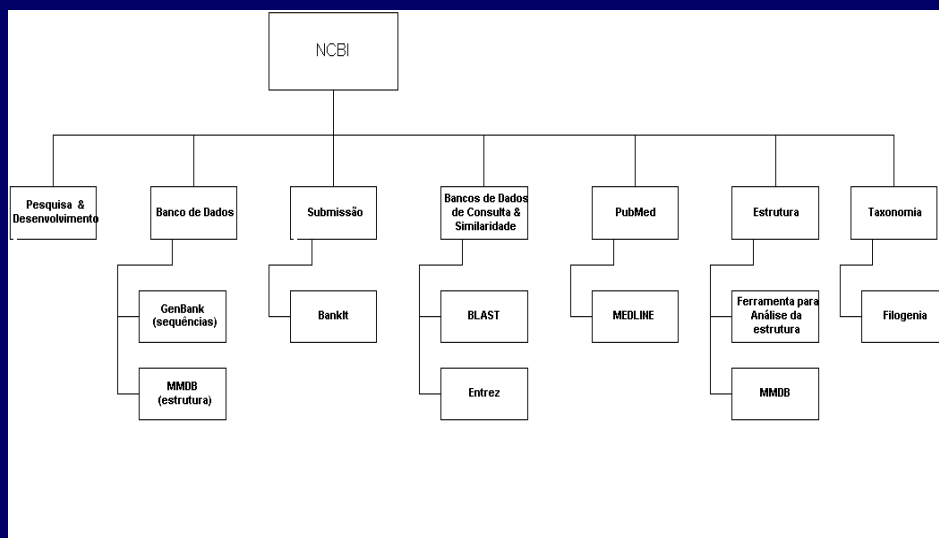
BankIt

- formulário em HTML desenvolvido para submissão de seqüências de forma conveniente e rápida via Web
- equipe de funcionários responsáveis pela anotação no GenBank revê a informação textual submetida, incorporando-a nos campos estruturados apropriados

Repositórios de Dados

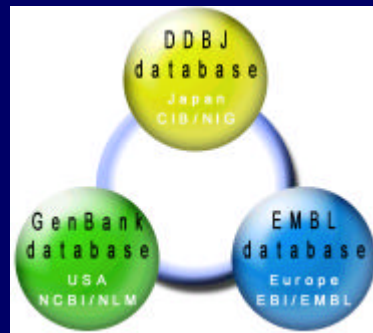
- NCBI (*national center for biotechnology information*): criado em 1988
- responsável por:
 - criar Bancos de Dados públicos
 - conduzir pesquisas na área de biologia computacional
 - desenvolver ferramentas de software próprias para análise de dados de genoma
 - disseminar informações biomédicas

Estrutura do NCBI



Genbank

- Banco de dados de seqüências de nucleotídeos gerados a partir do sequenciamento de DNA
- Desde 1982 o número de bases dobra a cada 14 meses



Modelo de Dados

- Bancos de dados de seqüências e as ferramentas de acesso do NCBI foram construídos a partir de um Modelo de Dados particular
- Modelo de dados simples e poderoso o suficiente para agregar dados heterogêneos
 - seqüências de nucleotídeos e amino-ácidos
 - estruturas tridimensionais
 - publicações (Medline)

Ferramentas para se Encontrar a Estrutura de um Gene

- Porque algoritmos?
 - mais rápido e limpo que bancada
 - poda possibilidades
 - fornece pistas para bancada
- Encontrar similaridades: localização rápida e montagem das regiões conservadas e com similaridade
- Métodos não baseados em similaridade

BLAST

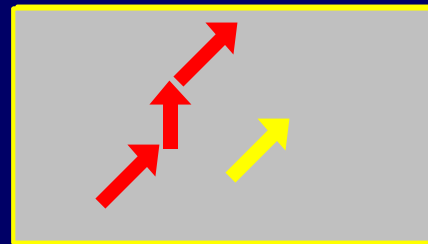
- *basic local alignment search tool*
- BLAST é um algoritmo usado por uma família de cinco programas que procuram por similaridade
- Métodos estatísticos são utilizados para julgar se a similaridade é significativa
- O resultados final da busca é apresentado por ordem de significância

Alinhamento por Similaridade

- Alinhar para encontrar pontos chave entre sequências
- Por comparação: quanto melhor similaridade, maior o escore

Exact matches of individual sequence letters are vital to an alignment. The alignment score is a measure of the number of matches. The score is the sum of the scores of the matches. The score is the sum of the scores of the matches. The score is the sum of the scores of the matches.

Human: GCC TTG GCC



Complexidade

- predição envolve rearranjos no alinhamento (substituição, inserção, deleção, duplicação, inversão de bases)
- alinhamento de sequências curtas: complexidade para atingir alinhamento ótimo é desprezível
- alinhamento de segmentos de cromossomos: deseja-se um resultado ótimo em período de 1 semana

Métodos não Baseados em Similaridade

- Codon bias
- Sinais em sequências (podem levar a sobre estimação do número de genes se sinal for fraco)

GENSCAN

- exemplo de utilização de *hidden markov models* (HMM)
- aspectos positivos:
 - flexível, processos probabilísticos
 - não necessita de similaridades com genes conhecidos
 - melhores resultados em termos de sensibilidade e especificidade
- aspectos negativos:
 - podem ser complicados, necessita treinamento
 - viés em relação ao conjunto de treinamento

Comparação GENSCAN/ Outros

1738 *Human Molecular Genetics, 1997, Vol. 6, No. 10 Review*

Table 2. Estimated performances of the various programs

Program	Original ref.	Test ref.	Prediction type	Sensitivity (%total)	Specificity (%total)	Sensitivity (%exact exon)	Specificity (%exact exon)	Missed exons	Wrong exon
FGENEH	59	52	Gene structure	83	93	73	78	15	11
GeneID	57	5	Gene structure	69	77	42	46	28	24
GeneParser	63	5	Gene structure	66	79	35	40	29	17
Genie	71	71	Gene structure	87	88	69	70	10	15
GenLang	58	5	Gene structure	72	79	51	52	21	21
GENSCAN	72	72	Gene structure	93	93	78	81	9	5
GRAIL II	46	52	Internal exons	79	85	51	57	23	28
GRAIL II/GAP	66	43	Gene structure	83	87	-	32	25	10
HEXON	50	52	Internal exons	88	80	71	65	10	27
MORGAN	-	-	Gene structure	83	79	58	51	14	-
MZEF	52	52	Internal exons	87	95	78	86	14	7
SurFind	24	5	Internal exons	71	85	42	47	24	14
VEIL	70	70	Gene structure	83	72	53	49	19	-
Xpound	51	5	Internal exons	61	87	15	18	32	13

Ferramentas de Bioinformática para Sequenciamento e Anotação

Ana Lúcia C. Bazzan

Algumas Referências

1736 *Human Molecular Genetics, 1997, Vol. 6, No. 10 Review*

Table 1. Contact addresses and availability of the programs cited in this article

Program (ref)	Electronic address	Type of access*
GeneID (57)	geneid@darwin.bu.edu www.imim.es/GenelDetermination/GenelD/geneid_input.html	ES HP
GeneParser (63)	beagle.colorado.edu/~essayden/GeneParser.html	HP, EX
Genie (71)	www.hgc.fbi.gov/inf/genie.html	HP, WS, ES
GenLang (58)	www.ncbi.upenn.edu/~sdong/genlang_home.html	HP, WS, SC
GENSCAN (72)	gnomic.stanford.edu/GENSCANW.html	HP, WS, ES
GENVIEW (65)	www.itba.mt.cnr.it/webgene	HP, WS
GRAIL (66)	avalon.epm.cml.gov	HP, ES, CL
HEXON/FGENEH (59)	dot.imgen.bcm.tmc.edu:9331/gene-finder/gf.html	HP, WS, ES
MORGAN (-)	www.cs.jhu.edu/labs/complio/morgan.html	HP, WS, EX
MZEF (52)	clio.cshl.org/genefinder	HP, WS, EX
ORFgene (75)	www.itba.mt.cnr.it/webgene	HP, WS
PROCRUSTES (74)	www-hio.usc.edu/software/procrustes/index.html	HP, WS, ES, EX
SurFind (24)	www.rabbitbunch.com	HP, EX
VEIL (70)	www.cs.jhu.edu/labs/complio/veil.html	HP, WS, ES, EX
Xpound (51)	ftp://igs-server.cnr-mrs.fr/pub/Banbury/xpound	SC
Banbury Cross	igs-server.cnr-mrs.fr	HP

Ferramentas de Bioinformática para Sequenciamento e Anotação

Ana Lúcia C. Bazzan

Projetos Genoma no Brasil

- Rede da FAPESP: 4-5 centros de bioinformática
- Genoma National: *C. violaceum* (LNCC/RJ)
- ProGeNE: EST Leishmania (UFPE)
- Bahia: patógeno Vassoura de Bruxa (UNICAMP)
- Minas Gerais: *Schistosoma mansoni* (Cenapad)
- Rio: bactéria da fixação do nitrogênio (LNCC)
- Paraná: SIMEPAR
- Genoma sul (em fase de proposta)
- Iniciativas autônomas

Maiores Informações

- Sites "da moda" estão constantemente mudando
- Populares
 - www.ncbi.nlm.nih.gov
 - www.tigr.org
 - www.fruitfly.org/~nomi/annotation-links.html
- Links atualizados, mailist, disciplinas no II/UFRGS:
<http://www.inf.ufrgs.br/~bazzan/bioinfo> ou
www.inf.ufrgs.br/bazzan