

Introdução à Bioinformática

Ronaldo Rodrigues Ferreira

Centro de Biologia Genômica e Molecular (PUCRS)
Instituto de Informática (UFRGS)

Uberaba, junho de 2005



Objetivos

- O que é Bioinformática
- Bioinformática *versus* Biocomputação
- Diálogo entre biólogos moleculares e cientistas da computação
- Bioinformática: proteômica e genômica
- A Bioinformática no Brasil – caso Genesul
- A Bioinformática e a iniciativa privada
- Sequenciamento de genomas: técnicas e equipamentos
- Bases de Dados biológicos
- Anotação de Genomas
- Softwares
- Principais locais de pesquisa no Brasil em Bioinformática
- Congressos

O que é Bioinformática

Desenvolvimento de ferramentas e métodos computacionais para análise, manipulação, construção, edição e gerenciamento de dados biológicos.

Análise em laboratório de dados biológicos é difícil e custosa. Portanto, técnicas computacionais são essenciais [1].

Bioinformática

versus

Biocomputação

*“... Or to come right to the point, how do we train
bacteria to make transistors?”*

Seymour Cray

pai do supercomputador [2]

Diálogo entre biólogos moleculares e cientistas da computação

- Falta de uma formação multidisciplinar
- Objetos de estudo diferentes

Proteômica ou Bioinformática Estrutural:

- Predição de estrutura de proteínas
- Visualização 3D de estruturas protéicas

Genômica:

- Análise, edição, manipulação de Genomas
- Banco de Dados biológicos
- Genômica funcional
- Anotação de genomas

Começou como uma iniciativa da Coopercitrus e Fapesp com o objetivo de sequenciar o genoma completo da *Xylella fastidiosa*. <http://www.xylella.Incc.br/>

Diversos projetos de sequenciamento estão em curso no País:

- | | |
|-----------------------|---|
| brGene | - http://www.brgene.Incc.br/ |
| OMM | - http://www.omm.Incc.br/ |
| PIGS | - http://www.genesul.Incc.br/ |
| <i>Leifsonia xyli</i> | - http://www.leifsonia.Incc.br/ |
| Genoma Café | - http://www.cenargen.embrapa.br/biotec/genomacafe/index.html |
| Genoma Banana | - http://genoma.embrapa.br/musa/index.html/ |
| RioGene | - http://www.riogene.Incc.br/ |
| entre outros | |

Genesul

Rede de laboratórios de Bioinformática, de Sequenciamento e de Diagnóstico dos Estados de Santa Catarina, do Paraná e do Rio Grande do Sul. O projeto é financiado pelo MCT e pela FAPERGS.

O projeto tem como objetivo sequenciar a bactéria *Mycoplasma hyopneumoniae*. Essa bactéria causa grandes danos ao porco, o que prejudica a produção no Sul do País. Após o sequenciamento, o objetivo é o desenvolvimento de vacinas.

ESTADO ATUAL: *Mycoplasma hyopneumoniae*, *Mycoplasma hyopneumoniae* 7448 e *Mycoplasma hyopneumoniae* 7442 já sequenciados

A Bioinformática e a iniciativa privada

O sequenciamento de genomas atrai grande interesse comercial.

Coopercitrus e Embrapa são hoje as principais empresas que usam da bioinformática para auxiliar o estudo dos genomas.

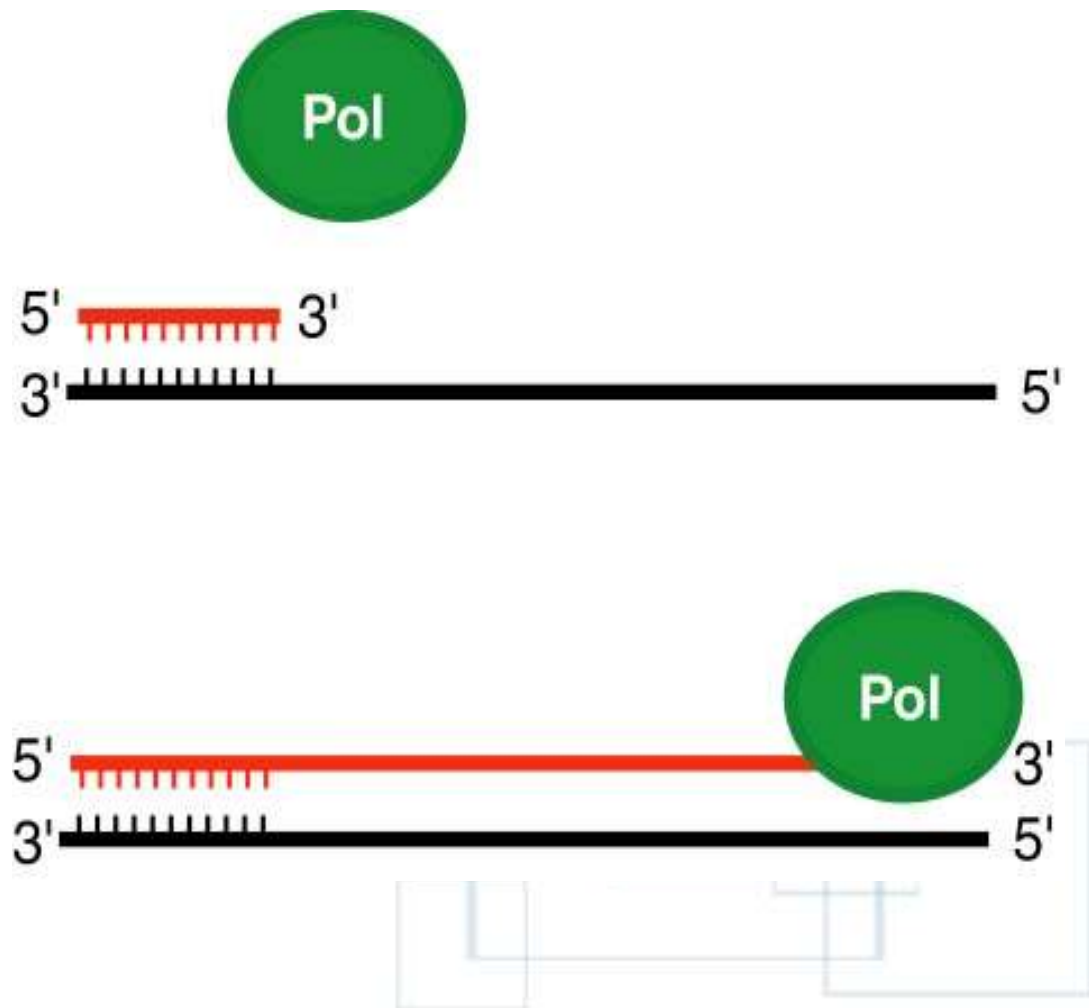
A Bioinformática acelera os estudos do genoma. Velocidade é imprescindível no mercado.



Sequenciamento de genomas: técnicas e equipamentos

Sequenciamento: técnicas

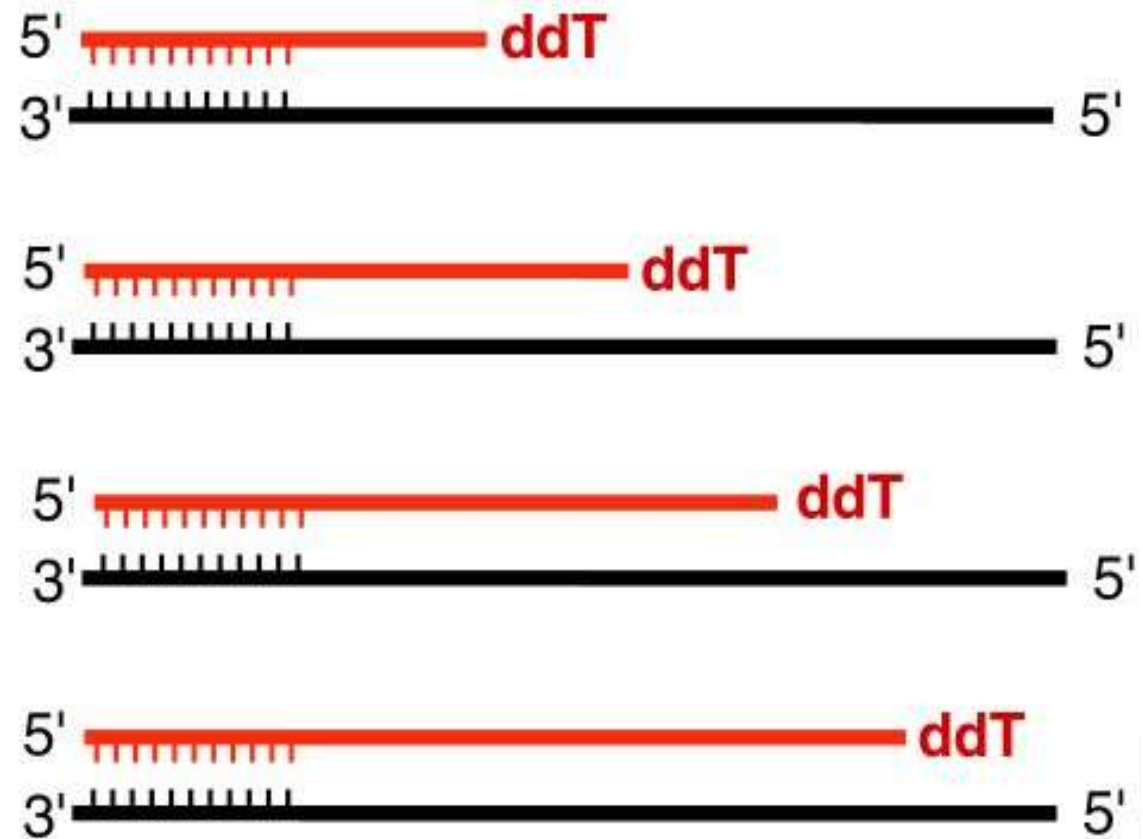
- Extensão direta da fita a partir do *primer* de uma fita de única face.
- Utiliza-se uma DNA polimerase.
- Os *primers* têm comprimento de 18 a 25 bases.



Sequenciamento de genomas: técnicas e equipamentos

Sequenciamento: técnicas

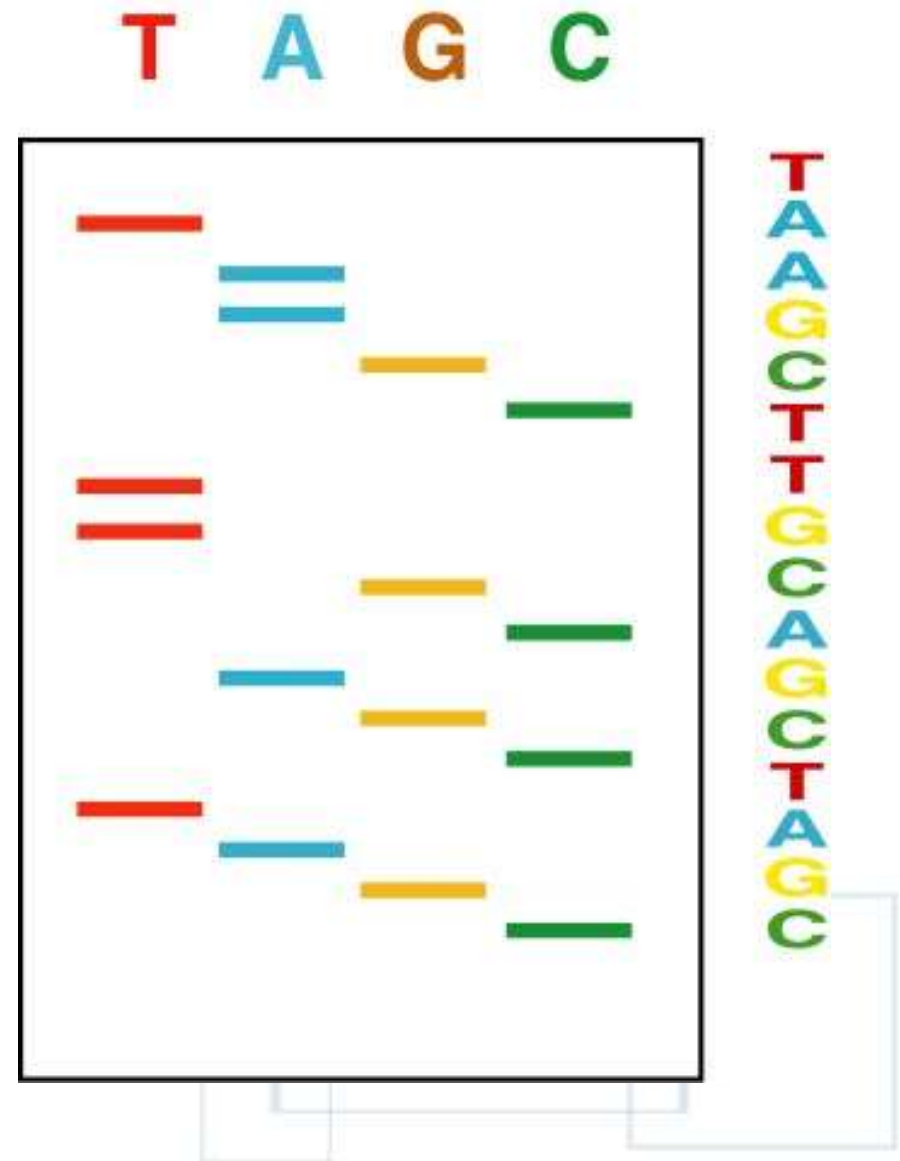
- Terminações em 4 dideoxidos (uma para cada base).
- A polimerase não se estende após essas terminações.
- São formadas diversas sequências de vários tamanhos.
- Produzem sequências terminadas sempre em A, C, G e T.



Sequenciamento de genomas: técnicas e equipamentos

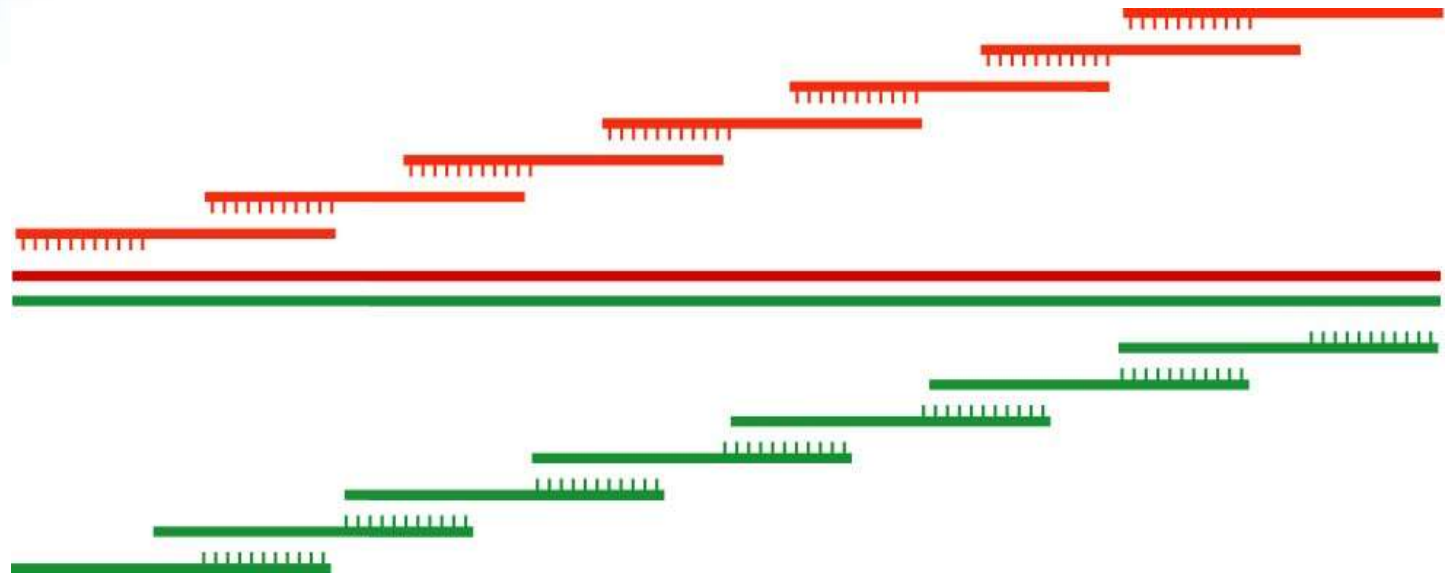
Sequenciamento: técnicas

- Utiliza-se marcadores com índices de refração diferentes (um para cada base).
- Utilizado para sequenciamento automatizado.
- Os produtos são separados por tamanho.



Sequenciamento de genomas: técnicas e equipamentos

Shotgun

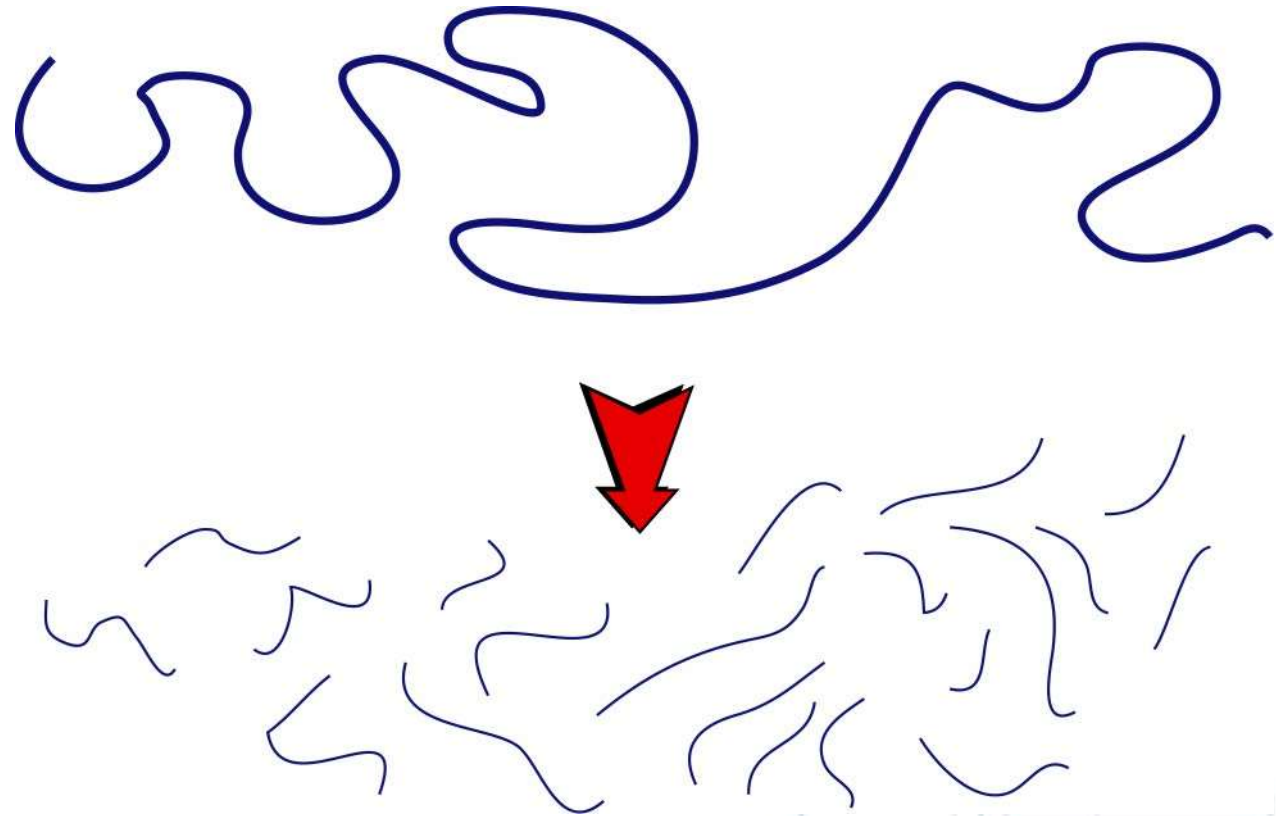


Particionar pequenas porções de DNA em vetores e depois montar a sequência

Sequenciamento de genomas: técnicas e equipamentos

Shotgun

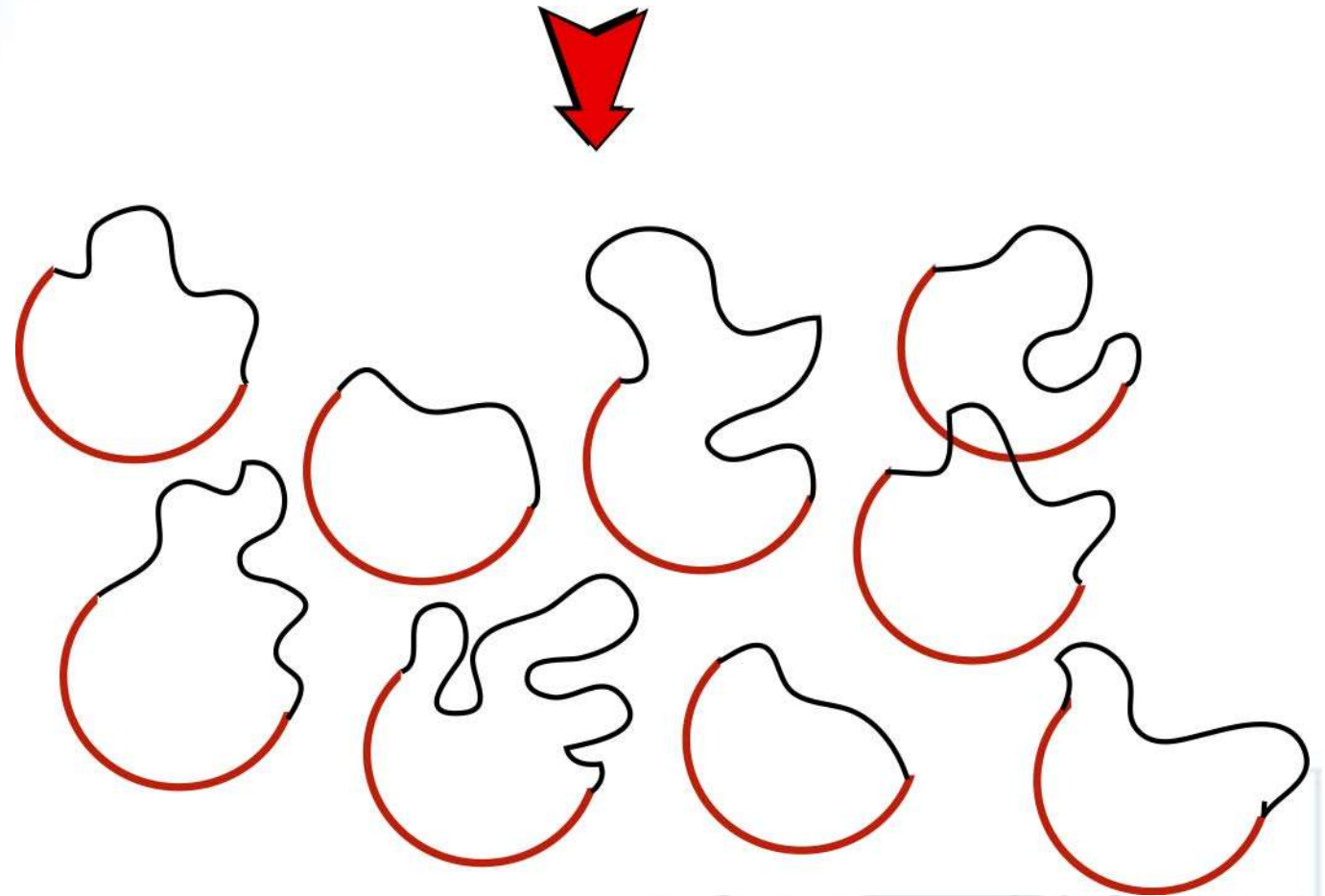
O DNA Genômico é partido em pedaços de tamanho estipulado



Sequenciamento de genomas: técnicas e equipamentos

Shotgun

O DNA é inserido em vetores universais

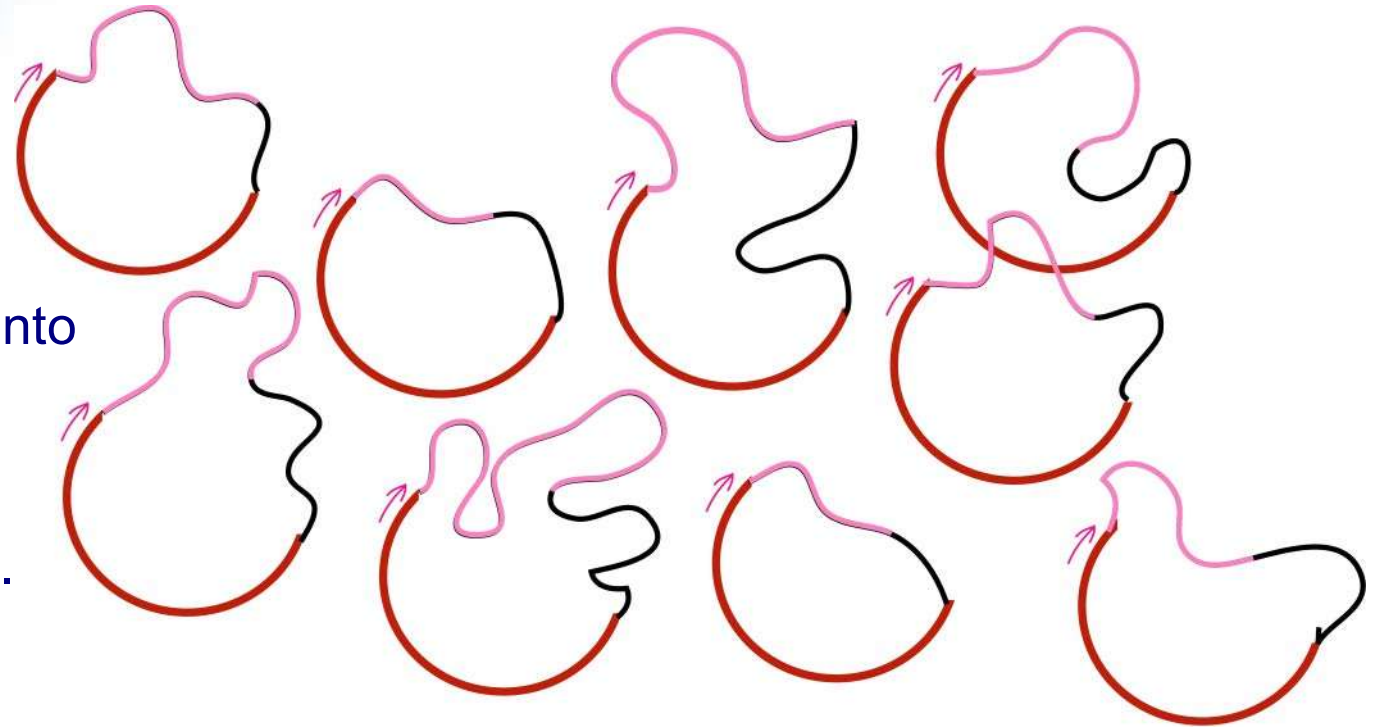


Sequenciamento de genomas: técnicas e equipamentos

Shotgun

Reações de sequenciamento são realizadas com um *primer* universal.

As reações são aleatórias.

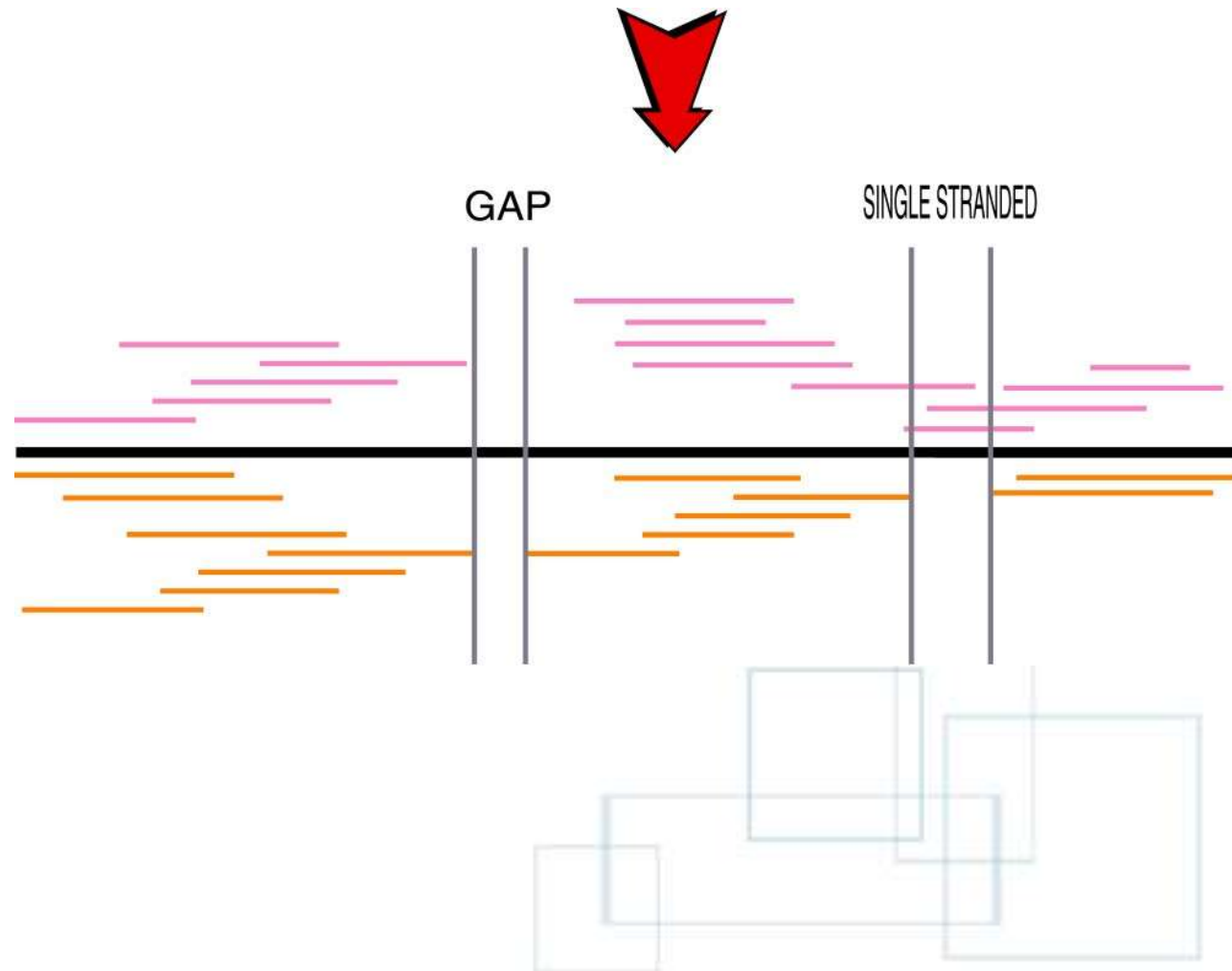


Sequenciamento de genomas: técnicas e equipamentos

Shotgun

Os *reads* são montados em *contigs* e regiões *single-stranded* (há sequência para somente uma fita).

Essas regiões são identificadas para posterior sequenciamento, produzindo assim a sequência completa.



Shotgun

A técnica *Shotgun* é aplicável a tamanhos diversos de DNA.

A única consideração é sobre o tamanho da porção do vetor de clonagem. Essa deve ser a menor possível.

Por exemplo, a porcentagem de DNA do Cosmídio é de aproximadamente 20%. Com isso, ao resequenciar um DNA que usa o Cosmídio como vetor de clonagem, 20% dos *reads* desse DNA são perdidos.

Nucleotídeos / Aminoácidos:

- GenBank

Proteínas:

- Swiss-Prot
- TrEMBL

Motivos:

- Prosite
- Interpro

Estruturas 3D:

- PDB

NCBI - GenBank

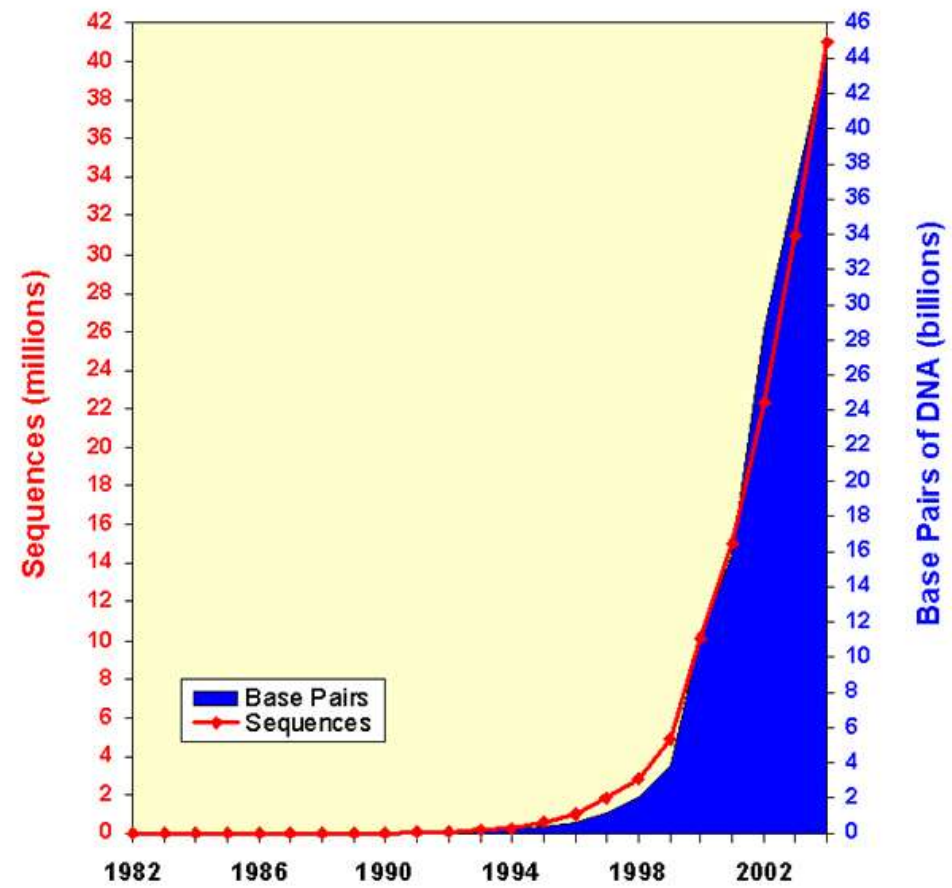
O *National Center of Biotechnology Information* começou suas atividades em 1988. Seus principais objetivos são:

- Estabelecimento de Banco de Dados públicos
- Pesquisa em Biologia Computacional e processos ligados à doenças
- Desenvolvimento de Software
- Análise de dados de genomas e informática médica

OBS: Ao contrário do Swiss-Prot, o Genbank não é curado.
Portanto, os seus dados podem possuir inconsistências.

NCBI - GenBank

Growth of GenBank
(1982 - 2004)



Swiss-Prot - TrEMBL

O Swiss-Prot é uma base de dados curada de proteínas que tem como objetivo:

- prover um alto nível de anotação (descrição da função de proteínas, seus domínios estruturais, modificações pós-translacionais, variantes, etc.)
- um nível mínimo de redundância
- alta integração com outras bases de dados.

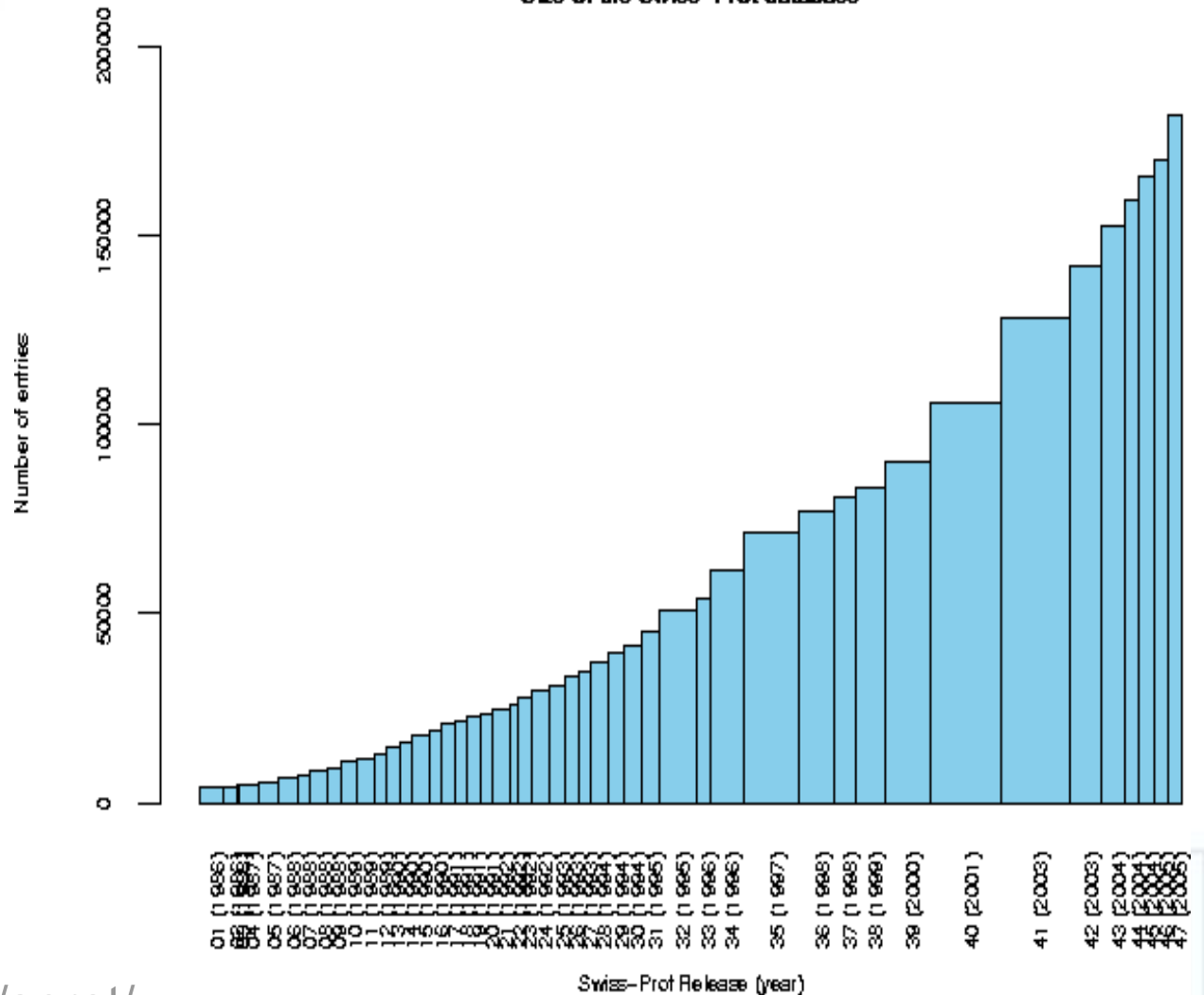
O TrEMBL é um suplemento de anotação por computador do Swiss-Prot que contém todas as sequências de nucleotídeos do EMBL ainda não integradas no Swiss-Prot.

Swiss-Prot - TrEMBL

Início: 1986

<http://br.expasy.org/sprot/>

Size of the Swiss-Prot database

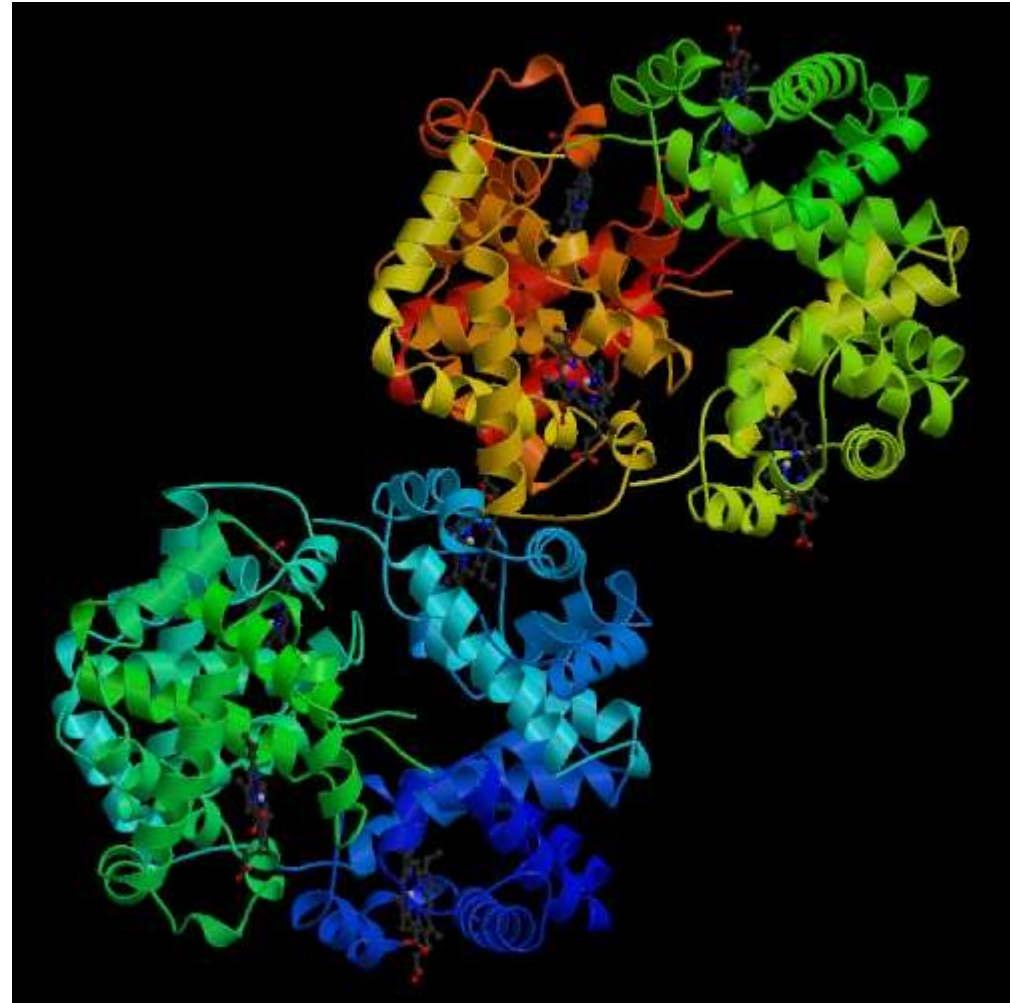


PDB

O *Protein Data Bank* armazena estruturas 3D de proteínas e macromoléculas.

<http://www.rcsb.org/pdb/index.html>

PDB



Descobrir, analisar, mapear, pesquisar as funções e características dos genes no genoma dos organismos

Montagem e Consenso

- Phred
- Phrap
- Consed

Alinhamento

- Sequencher
- ClustalW
- Staden

Edição

- BioEdit
- Sequencher

Propósitos Gerais

- EMBOSS

Phred

Interpreta os arquivos de sequenciamento de DNA, verifica os nucleotídeos e designa valores de qualidade para cada base.

Phrap

Montagem de sequências de DNA obtidas através do processo de *Shotgun*.

Consed

Visualização, edição e acabamento da montagem de sequências criadas com o *Phrap*.

Staden Package

Pacote de *softwares* de Bioinformática para:

- Montagem (*gap, pregap, vectorClip, screenSeq, findRenz, trev*)
- Detecção de Mutações (*traceDiff, hetscan, gap4*)
- Análise de Sequências (*spin, makeWeights*)
- Manipulação e Leitura de arquivos de sequenciamento (*convertTrace, getComment*)

ClustalW

Software de propósitos gerais para alinhamento múltiplo de DNA ou proteínas. Ele produz alinhamentos múltiplos de sequências divergentes com significado biológico. Ele calcula o melhor alinhamento para as sequências, alinha-as umas com as outras. Assim, as identidades, similaridades e diferenças entre as sequências podem ser vistas. Relação evolutiva pode ser visualizada através de Cladogramas ou Filogramas.



EMBOSS

European Molecular Biology Open Software Suite

Nasceu da necessidade dos biólogos moleculares por softwares específicos para o seu campo de estudos e da necessidade de uma plataforma de propósitos gerais para o desenvolvimento acadêmico de software de análise de sequenciamento [4].

O EMBOSS é composto por mais de cem aplicativos [4].

Há a possibilidade da utilização de qualquer formato de sequências. Novos formatos são facilmente adicionados. Além do suporte às bases públicas, é possível utilizar o EMBOSS com bases privadas [4].

<http://emboss.sourceforge.net/>

[4] EMBOSS: The European Molecular Biology Open Software Suite / Rice,P. Longden,I. and Bleasby,A. Trends in Genetics 16, (6) pp276277

Locais de Pesquisa em Bioinformática no Brasil

Nordeste:

- **Laboratório de Bioinformática da UFPE** <http://biolab.cin.ufpe.br/>

Centro Oeste:

- **EMBRAPA** <http://asparagin.cenargen.embrapa.br/pt/>

Sudeste:

- **UFRJ** <http://www.bioinfo.ufrj.br/>
- **LNCC** <http://www.lncc.br/~labinfo>
- **USP São Paulo** <http://www.ime.usp.br/posbioinfo/>
- **UFMG** <http://www.ufmg.br/bioinformatica/>
- **FIOCRUZ**
- **UNICAMP** <http://www.lbi.ic.unicamp.br>

Sul:

- **PUCRS**
- **UFRGS** <http://www.inf.ufrgs.br/~crym/LabBioInf/>

Locais de Pesquisa em Bioinformática no Brasil

Congresso Nacionais

- Congresso Brasileiro de Genética
Sociedade Brasileira de Genética

<http://www.sbg.org.br>

- Simpósio Brasileiro em Bioinformática
Sociedade Brasileira de Computação

http://www.unisinos.br/simposio/bsb/index_port.php

Congressos Internacionais

- ISMB <http://www.iscb.org/ismb2005/sigs.html>
- RECOMB <http://www.broad.mit.edu/recomb2005/>

International Society for Computational Biology

<http://www.iscb.org/>

- [1] **Machine Learning Approaches to Gene Recognition**
Mark W. Craven and Jude W. Shavlik IEEE – AI in Molecular Biology (1994)
- [2] **Imitation of Life: How Biology is Inspiring Computing**
FORBES, Nancy / MIT Press (2004)
- [3] **Slides da disciplina “Tópicos Especiais em Computação VI: Introdução à Bioinformática” do semestre 2005/1 do Instituto de Informática da UFRGS**
Prof. Dra. Ana Lúcia C. Bazzan (2005)
- [4] **EMBOSS: The European Molecular Biology Open Software Suite**
Rice,P. Longden,I. and Bleasby,A. -Trends in Genetics 16, (6) pp276277
- [5] **Bioinformatic tools for DNA/protein sequence analysis, functional assignment of genes and protein classification**
Rehm, B.H.A . -Applied Microbiology Biotechnology 57, pp579-592 (2001)

www.inf.ufrgs.br/~rrferreira/bioinf

rrferreira@inf.ufrgs.br

