



Universidade de São Paulo
Instituto de Química



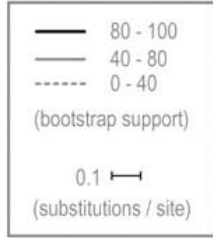
Comparação de sequências

João Carlos Setubal

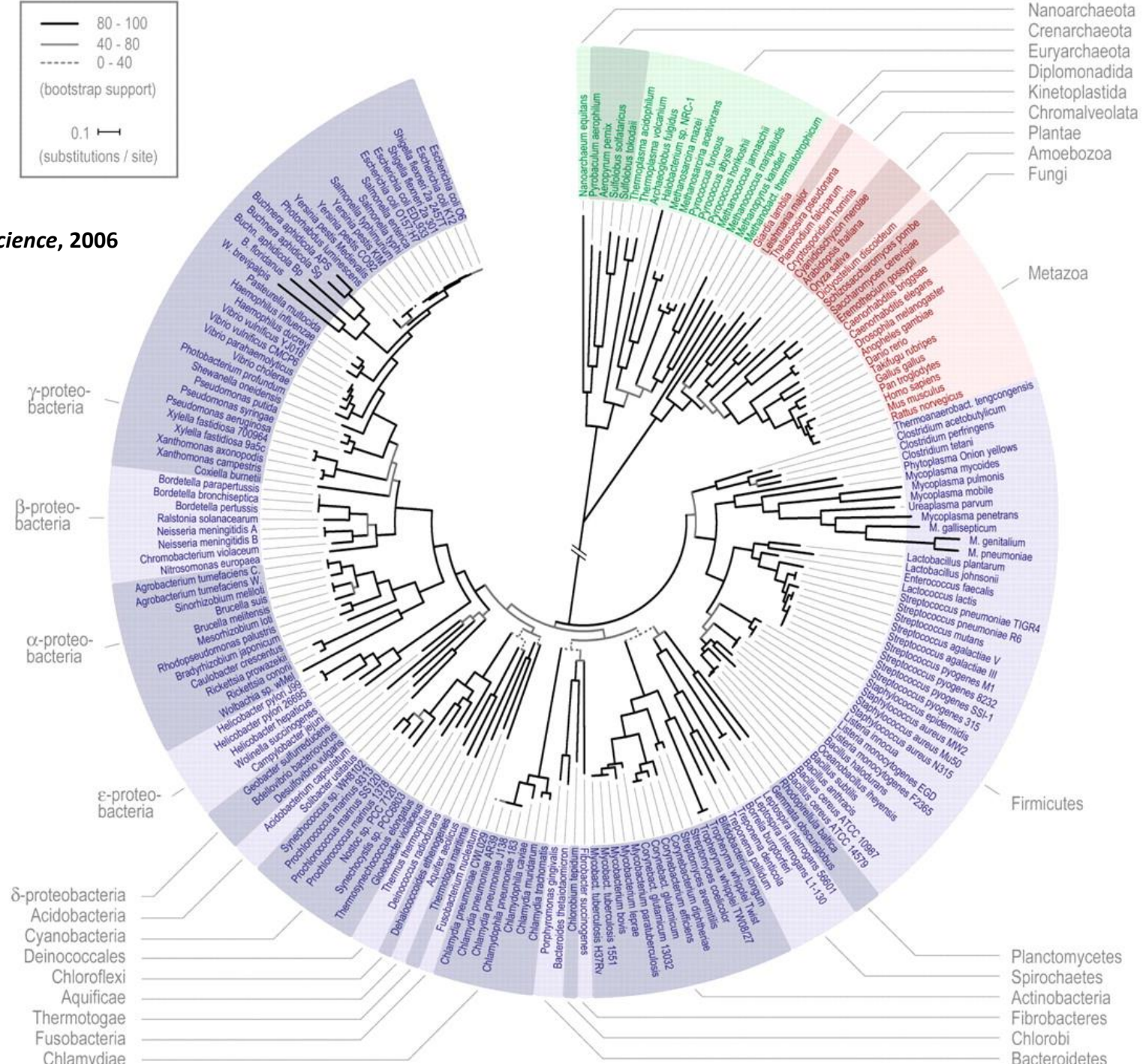
IQ-USP

2017

Motivação



Ciccarelli et al, *Science*, 2006



Como foi possível criar tal árvore?

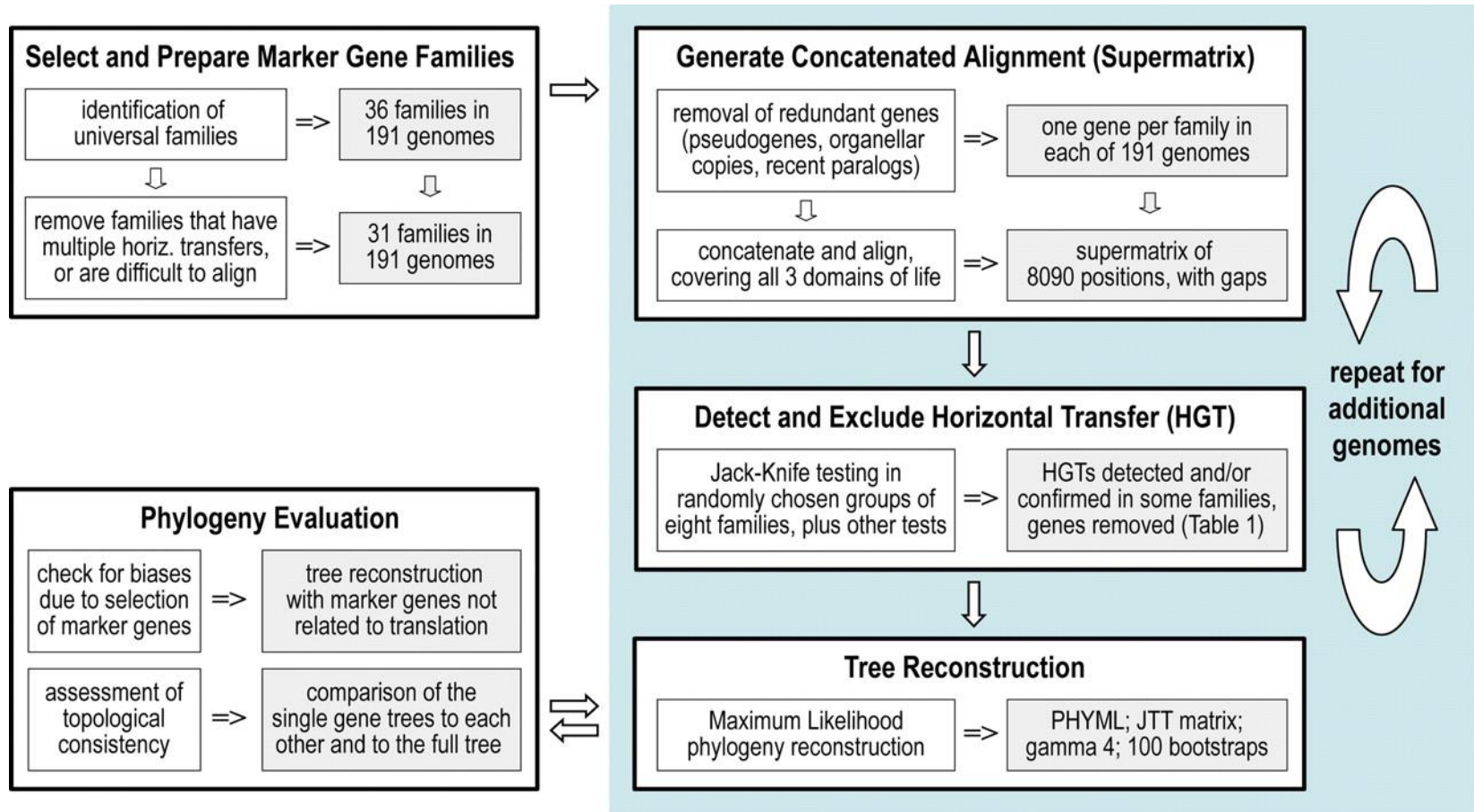
- Determinar genes que são **compartilhados** por todos os seres vivos
- O que é “compartilhar”?
 - Gene x_1 em organismo A tem função α
 - Gene x_2 em organismo B é **homólogo** e tem a **mesma função α**
 - Então x_1 e x_2 são “o mesmo gene x ” e portanto ele é compartilhado por A e B
- Quais genes são esses?

Orthologous Group	Av. Length	Annotation	Genes in Prok.	Genes in Euk.	Total Genes
COG0012	380	Predicted GTPase, probable translation factor	171	30	201
COG0016	423	Phenylalanine-tRNA synthetase alpha subunit	168	42	210
COG0018†	548	Arginyl-tRNA synthetase	175	45	220
COG0048	137	Ribosomal protein S12	168	48	216
COG0049	182	Ribosomal protein S7	169	41	210
COG0052	240	Ribosomal protein S2	168	79	247
COG0060*	956	Isoleucyl-tRNA synthetase	172	42	214
COG0080	154	Ribosomal protein L11	170	61	231
COG0081	230	Ribosomal protein L1	168	61	229
COG0085†	1138	DNA-directed RNA polymerase, beta subunit	178	60	238
COG0087	288	Ribosomal protein L3	168	54	222
COG0091	157	Ribosomal protein L22	168	75	243
COG0092	240	Ribosomal protein S3	168	30	198
COG0093	130	Ribosomal protein L14	168	41	209
COG0094	182	Ribosomal protein L5	169	36	205
COG0096	131	Ribosomal protein S8	168	55	223
COG0097	177	Ribosomal protein L6P/L9E	168	65	233
COG0098	220	Ribosomal protein S5	168	110	278
COG0099‡	133	Ribosomal protein S13	168	49	217
COG0100	145	Ribosomal protein S11	169	51	220
COG0102	167	Ribosomal protein L13	168	54	222
COG0103	172	Ribosomal protein S9	168	52	220
COG0124*	472	Histidyl-tRNA synthetase	178	31	209
COG0143*†	646	Methionyl-tRNA synthetase	180	35	215
COG0172	442	Seryl-tRNA synthetase	177	37	214
COG0184	154	Ribosomal protein S15P/S13E	168	41	209
COG0186	122	Ribosomal protein S17	170	46	216
COG0197	175	Ribosomal protein L16/L10E	168	54	222
COG0200	166	Ribosomal protein L15	168	70	238
COG0201	445	Preprotein translocase subunit SecY	178	37	215
COG0202	323	DNA-directed RNA polymerase, alpha subunit	171	45	216
COG0256	178	Ribosomal protein L18	168	50	218
COG0495	854	Leucyl-tRNA synthetase	172	43	215
COG0522	199	Ribosomal protein S4 and related proteins	174	46	220
COG0525*‡	880	Valyl-tRNA synthetase	169	37	206
COG0533	375	Metal-dependent proteases with chaperone activity	168	35	203

Questões

- Onde podemos achar as sequências dos genes?
- Como determinar compartilhamento?
- Como preparar esses dados para construir uma árvore?
- Como construir uma árvore?
- Como saber se ela está correta?
- Algo novo desde 2006?

Fig. 1. Overview of the procedure.



Respostas

- Onde podemos achar as sequências dos genes?
 - Bancos de dados públicos (NCBI)
 - BLAST
- Como determinar compartilhamento?
 - Através de comparação de sequências
- Como preparar esses dados para construir uma árvore?
 - Alinhamento múltiplo concatenado
- Como construir uma árvore?
 - Métodos de reconstrução filogenética
- Como saber se ela está correta?
 - Inferência é um termo melhor do que construção
 - Argumentos probabilísticos
 - Transferência Horizontal (Lateral) de Genes
- Algo novo desde 2006?
 - Next Generation Sequencing (NGS)
 - Comparação de genomas completos

Por que comparar sequências?

- Achar similaridades
 - Dadas 2 sequências, **quão parecidas** elas são?
 - DNA e proteína
- **Buscas** em banco de dados
 - Achar quais sequências do banco são parecidas com minha sequência-consulta
 - Consulta (*query*) é tipicamente uma sequência **nova**
 - “google”

Motivação (cont.)

- Construir **famílias de proteínas**
 - Saber quais organismos tem membros da família
 - Determinar uma “assinatura” para a família
- Construir **filogenias**
 - Entender a **história evolutiva** de genes e organismos

Premissas

- Em geral buscamos sequências “aparentadas”
- Sequências “aparentadas” são similares
- “aparentadas” = **homólogas**
 - Descendem de um mesmo ancestral
- Descendentes sofreram mutações ao longo do tempo

Alinhamento de DNA

GTGGTGGCCTACGAAGGT

GTAGTGCCTTCGAAGGGT

Como avaliar um alinhamento?

- Sistema de pontuação
 - Match: +1
 - Mismatch: -1

Pontuação do alinhamento

GTGGTGGCCTACGAAGGT

GTAGTGCCTTCGAAGGGT

+1+1-1+1+1+1-1+1-1+1-1-1-1+1-1+1+1+1 = 4

É possível melhorar o alinhamento?

- Sim
- Pela introdução de espaços

Alinhamento com espaços

GTGGTGGCCTACGAA-GGT
GTAGTG-CCTTCGAAGGGT

Sistema de pontuação com espaços

- Match: +1
- Mismatch: -1
- Espaço: -2
- (Buraco: sequência de espaços)
 - Em inglês: *gaps*

Pontuação do alinhamento

GTGGTGGCCTACGAA-GGT
GTAGTG-CCTTCGAAAGGT

$$+1+1-1+1+1+1-2+1+1+1-1+1+1+1+1-2+1+1+1 = 9$$

Justificativa para o sistema de pontuação

- Matches tem que ser recompensados (> 0)
- Mismatches e espaços tem que ser penalizados (< 0)
- Mismatches representam **substituições**
 - **Mutações** (ocorrem com frequência)
 - Podem não trazer letalidade
- Espaços representam **inserções** ou **remoções**
 - Mais prováveis de causarem letalidade
 - Alteram quadro de leitura
 - Ocorrem com muito menos frequência

Alinhamentos ótimos

- São os alinhamentos de pontuação **máxima**
- **similaridade** entre duas sequências
 - É o valor da pontuação do alinhamento ótimo
- No exemplo anterior
 - Similaridade = 9

Comparação de sequências de aminoácidos

Pontuação de alinhamento de proteínas

```
H:  I I W G E D T L M E Y L E N P K K Y I P G T K M I F V G I K K K E E R A D L I A Y L K K A T N E
C:  V V W T K E T L F E Y L L N P K K Y I P G T K M V F A G L K K A D E R A D L I K Y I E V E S A K S L
      *      **   ***  ***** *  *  **   ***** *
```

% de identidade é uma medida simples mas válida de similaridade de sequências de proteínas

Aminoácidos se dividem em famílias

- Hidrofóbicos
 - Ala, Val, Phe, Pro, Met, Ile, Leu
- Com carga
 - Asp, Glu, Lys, Arg
- Polares
 - Ser, Thr, Tyr, His, Cys, Asn, Gln
 - Trp
- Gly

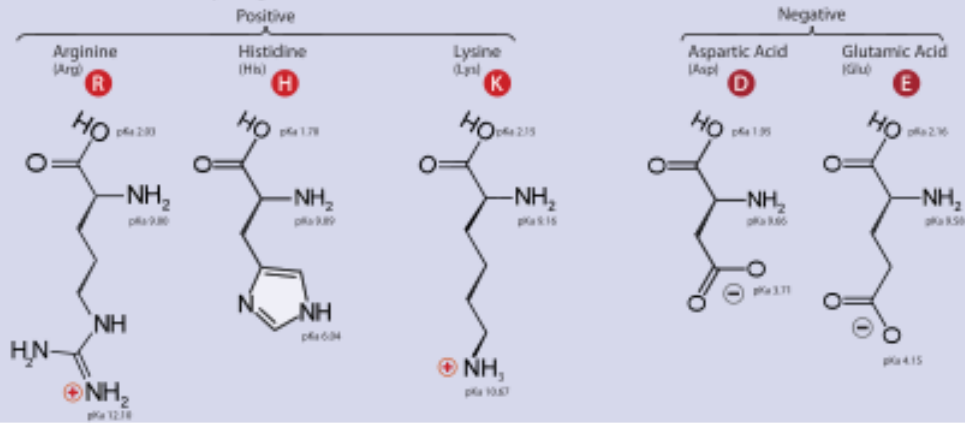
Twenty-One Amino Acids

⊕ Positive

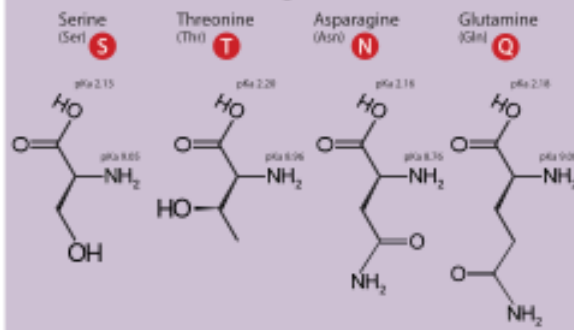
⊖ Negative

• Side chain charge at physiological pH 7.4

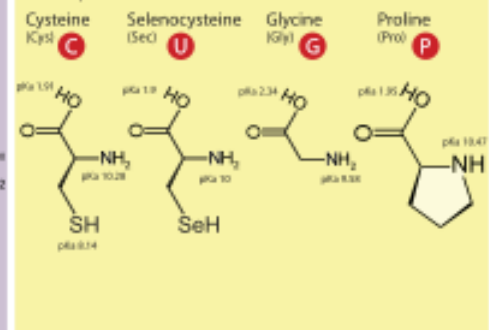
A. Amino Acids with Electrically Charged Side Chains



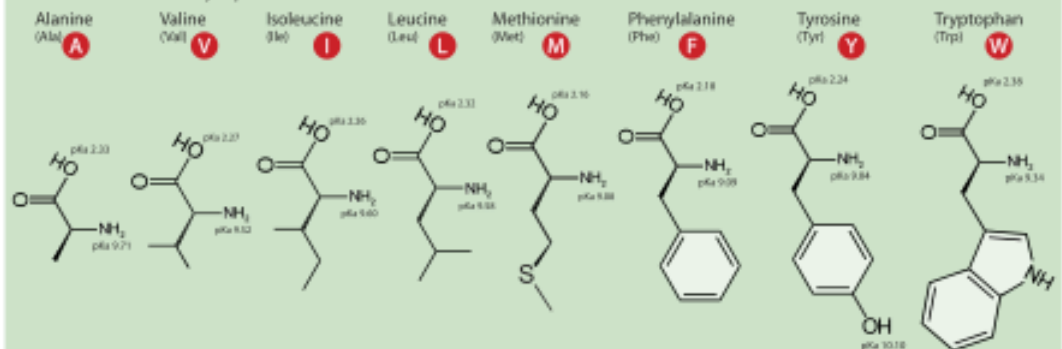
B. Amino Acids with Polar Uncharged Side Chains



C. Special Cases



D. Amino Acids with Hydrophobic Side Chain



Mutações e proteínas

- Substituições que não alteram a estrutura da proteína tendem a ser preservadas durante a evolução
- A troca de um aminoácido de uma família por outro da **mesma** família em geral cai nessa categoria
- (Indels podem ter consequências mais drásticas)
- Então: como avaliar mismatches?

Matriz de substituição de amino ácidos BLOSUM62

```
# Matrix made by matblas from blosum62.iiij
# * column uses minimum score
# BLOSUM Clustered Scoring Matrix in 1/2 Bit Units
# Blocks Database = /data/blocks_5.0/blocks.dat
# Cluster Percentage: >= 62
# Entropy = 0.6979, Expected = -0.5209
  A  R  N  D  C  Q  E  G  H  I  L  K  M  F  P  S  T  W  Y  V  B  Z  X  *
A  4 -1 -2 -2  0 -1 -1  0 -2 -1 -1 -1 -1 -2 -1  1  0 -3 -2  0 -2 -1  0 -4
R -1  5  0 -2 -3  1  0 -2  0 -3 -2  2 -1 -3 -2 -1 -1 -3 -2 -3 -1  0 -1 -4
N -2  0  6  1 -3  0  0  0  1 -3 -3  0 -2 -3 -2  1  0 -4 -2 -3  3  0 -1 -4
D -2 -2  1  6 -3  0  2 -1 -1 -3 -4 -1 -3 -3 -1  0 -1 -4 -3 -3  4  1 -1 -4
C  0 -3 -3 -3  9 -3 -4 -3 -3 -1 -1 -3 -1 -2 -3 -1 -1 -2 -2 -1 -3 -3 -2 -4
Q -1  1  0  0 -3  5  2 -2  0 -3 -2  1  0 -3 -1  0 -1 -2 -1 -2  0  3 -1 -4
E -1  0  0  2 -4  2  5 -2  0 -3 -3  1 -2 -3 -1  0 -1 -3 -2 -2  1  4 -1 -4
G  0 -2  0 -1 -3 -2 -2  6 -2 -4 -4 -2 -3 -3 -2  0 -2 -2 -3 -3 -1 -2 -1 -4
H -2  0  1 -1 -3  0  0 -2  8 -3 -3 -1 -2 -1 -2 -1 -2 -2  2 -3  0  0 -1 -4
I -1 -3 -3 -3 -1 -3 -3 -4 -3  4  2 -3  1  0 -3 -2 -1 -3 -1  3 -3 -3 -1 -4
L -1 -2 -3 -4 -1 -2 -3 -4 -3  2  4 -2  2  0 -3 -2 -1 -2 -1  1 -4 -3 -1 -4
K -1  2  0 -1 -3  1  1 -2 -1 -3 -2  5 -1 -3 -1  0 -1 -3 -2 -2  0  1 -1 -4
M -1 -1 -2 -3 -1  0 -2 -3 -2  1  2 -1  5  0 -2 -1 -1 -1 -1  1 -3 -1 -1 -4
F -2 -3 -3 -3 -2 -3 -3 -3 -1  0  0 -3  0  6 -4 -2 -2  1  3 -1 -3 -3 -1 -4
P -1 -2 -2 -1 -3 -1 -1 -2 -2 -3 -3 -1 -2 -4  7 -1 -1 -4 -3 -2 -2 -1 -2 -4
S  1 -1  1  0 -1  0  0  0 -1 -2 -2  0 -1 -2 -1  4  1 -3 -2 -2  0  0  0 -4
T  0 -1  0 -1 -1 -1 -1 -2 -2 -1 -1 -1 -1 -2 -1  1  5 -2 -2  0 -1 -1  0 -4
W -3 -3 -4 -4 -2 -2 -3 -2 -2 -3 -2 -3 -1  1 -4 -3 -2 11  2 -3 -4 -3 -2 -4
Y -2 -2 -2 -3 -2 -1 -2 -3  2 -1 -1 -2 -1  3 -3 -2 -2  2  7 -1 -3 -2 -1 -4
V  0 -3 -3 -3 -1 -2 -2 -3 -3  3  1 -2  1 -1 -2 -2  0 -3 -1  4 -3 -2 -1 -4
B -2 -1  3  4 -3  0  1 -1  0 -3 -4  0 -3 -3 -2  0 -1 -4 -3 -3  4  1 -1 -4
Z -1  0  0  1 -3  3  4 -2  0 -3 -3  1 -1 -3 -1  0 -1 -3 -2 -2  1  4 -1 -4
X  0 -1 -1 -1 -2 -1 -1 -1 -1 -1 -1 -1 -1 -1 -2  0  0 -2 -1 -1 -1 -1 -1 -4
* -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4  1
```

Fonte: NCBI

Pontuação leva em conta a matriz

- Match: $\text{blosum62}(i,i)$ sempre positivo
- Mismatch: $\text{blosum62}(i,j)$ positivo, nulo, negativo
- Espaço: -2

```
H: GDVEKGKKIFIMKCSQCHTVEKGGKHKTGPNLHGLEFGRKTGQAPGYSYTAANKNKGIIWG
GD EKGKK++ +C QCH V+      KTGP LHG+ GR +G   G+ Y+AANKNKG++W
C: GDYEKGKKVYKQRCLQCHVVDSTAT-KTGPTLHGVIGRTSGTVSGFDYSAANKNKGVVWT
```

Como obter alinhamentos ótimos?

- Precisamos de um **algoritmo**
- Algoritmos são diferentes de **programas**
 - Algoritmo é um método abstrato
 - Programa é uma **encarnação física** (numa linguagem de programação) de um algoritmo
 - Um programa pode ser **executado** num computador
 - Dizemos que um certo algoritmo **foi implementado** em tal ou qual linguagem
- Quem escrever **algoritmo** na prova perde 3 pontos!

Algoritmo para achar alinhamentos ótimos de DNA

- Desenvolvido com a técnica de **Programação dinâmica**
- Técnica desenvolvida na década de 1950 por Richard Bellman
- **Não** é programação e **não** é dinâmica!

Programação Dinâmica

- PD se usa para problemas que tem uma estrutura de **subproblemas**
- Num alinhamento com sequências s e t um **subproblema** é qualquer alinhamento entre s' e t' tal que
 s' = um **prefixo** de s e t' = um **prefixo** de t

```
H:  I IWGEDTLMEYLENPKKYI PGTKMIFVGIKKKEERADLIAYLKKATNE
C:  VVWTKETLFEYLLNPKKYI PGTKMVFAGLKKADERADLIKYIEVESAKSL
    *  **  ***  ****  *****  **  *  *  **  *****  *
```

```
H:  I IWGEDTLMEYLENPKKYI PGT
C:  VVWTKETLFEYLLNPKKYI PGT
    *  **  ***  ****  *****
```



Um prefixo

Ideia básica da PD

- Achar soluções de subproblemas e armazená-las numa **tabela** (matriz)
- Para achar a solução **ótima**:
 - Ir achando as soluções na direção dos subproblemas **menores** para os **maiores**
 - Último elemento da tabela a ser preenchido contém a solução do problema “completo”

	j	0	1	2	3	4
i		t	G	A	T	C
0	s					
1	G					
2	T					
3	C					

Preenchimento da tabela

- Este processo precisa começar com o “menor subproblema possível”. Qual seria?
 - Quando pelo menos uma das sequências é vazia
- Inicialização: Alinhar s com cadeia vazia e alinhar t com cadeia vazia

	j	0	1	2	3	4
i		t	G	A	T	C
0	s	0	-2	-4	-6	-8
1	G	-2				
2	T	-4				
3	C	-6				

Continuação

- Alinhar caracter X com caracter Y
- 3 possibilidades
 - X com Y
 - Aplicar pontuação respectiva, dependendo se for DNA ou proteína
 - X com espaço
 - Cobrar -2
 - Y com espaço
 - Cobrar -2

Preenchimento de (1,1)

- Significa determinar qual é o melhor alinhamento dos prefixos de s e t com apenas um caracter cada um
- Alternativas

G-	-G	G
-G	G-	G

Todos eles usam valores determinados na inicialização

	j	0	1	2	3	4
i		t	G	A	T	C
0	s	0	-2	-4	-6	-8
1	G	-2	1			
2	T	-4				
3	C	-6				

	j	0	1	2	3	4
i		t	G	A	T	C
0	s	0	-2	-4	-6	-8
1	G	-2	1	-1	-3	-4
2	T	-4	-1	0	0	-1
3	C	-6	-3	-2	-1	1

Exercícios

- Inventar duas sequências de DNA curtas e “rodar” (na mão) o algoritmo de PD
- Para se auto-corrigir:
 - <http://www.codeproject.com/Articles/304772/DNA-Sequence-Alignment-using-Dynamic-Programming-A>
 - (busque **dna sequence alignment code project**)
 - Demo parece segura
 - Exige registro
 - Tem código fonte

Complexidade computacional de PD

- Queremos saber quanto tempo gasta o algoritmo
- Mas algoritmos são abstratos, não estão associados a computadores físicos; o que significa “tempo”?
- Então contamos **número de operações elementares**
 - Operações aritméticas
 - Uso de posições da memória
 - Etc
- Porém queremos apenas algo **aproximado**
- Uma **ordem de grandeza**

Complexidade computacional

- Queremos expressar o número de operações como uma **função matemática** do tamanho das entradas
- Por exemplo
 - n (linear)
 - n^2 (quadrático)
 - $n \log n$
- Vamos desprezar constantes (n e $30n$ **dão na mesma**)
- Só nos interessa **o termo de maior grau** (desprezar os demais)
 - $7n^2 + 1400n + 3000 \log n$ vira apenas n^2
- Vamos usar a notação **$O(f(n))$** para denotar tudo isso

Complexidade computacional de PD

- A matriz tem tamanho $n+1$ por $m+1$
- Todos os elementos da matriz precisam ser preenchidos
- Supondo tempo constante para o preenchimento
 - $n+1 \times m+1 = nm + n + m + 1$
 - $O(nm)$
 - Se $n \approx m$, $O(n^2)$
 - Quadrático
- Memória: quadrático também

Algorithm *Similarity*

input: sequences s and t

output: similarity between s and t

$m \leftarrow |s|$

$n \leftarrow |t|$

for $i \leftarrow 0$ **to** m **do**

$a[i, 0] \leftarrow i \times g$

for $j \leftarrow 0$ **to** n **do**

$a[0, j] \leftarrow j \times g$

for $i \leftarrow 1$ **to** m **do**

for $j \leftarrow 1$ **to** n **do**

$a[i, j] \leftarrow \max(a[i - 1, j] + g,$
 $a[i - 1, j - 1] + p(i, j),$
 $a[i, j - 1] + g)$

return $a[m, n]$

FIGURE 3.2

Basic dynamic programming algorithm for comparison of two sequences.

Penalização de espaços pode ser mais sofisticada

GTGGTGGCCTACGAAGGT

GTGGTCGC---CGAAGGT

GTGGTGGCC-ACGAAGGT

GT-GTCGCCTACGA-GGT

- No sistema de pontuação apresentado, k espaços consecutivos (um buraco ou *gap*) custam **o mesmo** que k espaços separados
- Seria melhor distinguir os dois casos

Penalização de espaços feita por uma **função matemática**

- k = número de espaços
- $p(k) = a + bk$
- $p(k)$ é **subtraído** do score
- a = custo para **abrir** um buraco
- b = custo para **continuar** um buraco
- Por exemplo: $p(k) = 2 + k$
- 5 espaços consecutivos custam **7**
- 5 espaços separados custam **15**
- Compare com a função implícita do sistema simples: $p(k) = 2k$

Algoritmo de PD com penalização de buracos por função **afim**

- O algoritmo é mais complexo
 - São necessárias 3 tabelas ao invés de 1
- Mas a complexidade permanece a mesma (quadrática)
- Algoritmo de **Smith-Waterman**

Smith-Waterman

$a[i, j]$ = maximum score of an alignment between $s[1..i]$ and $t[1..j]$ that ends in $s[i]$ matched with $t[j]$.

$b[i, j]$ = maximum score of an alignment between $s[1..i]$ and $t[1..j]$ that ends in a space matched with $t[j]$.

$c[i, j]$ = maximum score of an alignment between $s[1..i]$ and $t[1..j]$ that ends in $s[i]$ matched with a space.

The entries (i, j) of these arrays depend on previous entries according to the following formulas, valid for $1 \leq i \leq m$ and $1 \leq j \leq n$:

$$a[i, j] = p(i, j) + \max \begin{cases} a[i - 1, j - 1] \\ b[i - 1, j - 1] \\ c[i - 1, j - 1] \end{cases}$$

$$b[i, j] = \max \begin{cases} -(h + g) + a[i, j - 1] \\ -g + b[i, j - 1] \\ -(h + g) + c[i, j - 1] \end{cases}$$

$$c[i, j] = \max \begin{cases} -(h + g) + a[i - 1, j] \\ -(h + g) + b[i - 1, j] \\ -g + c[i - 1, j]. \end{cases}$$

As before, $p(i, j)$ indicates the score of a matching between $s[i]$ and $t[j]$.

- Se a função for genérica [ex. $p(k) = a + b \log k$], então a complexidade passa para $O(n^3)$
 - Algoritmo de **Needleman-Wunsch**

Queremos descobrir sequências aparentadas

- Aparentadas = ancestral comum = homólogas
- Alinhamentos **biologicamente relevantes**
- Nota máxima por si só não nos informa sobre parentesco
 - Alinhamentos de nota máxima não necessariamente correspondem a alinhamentos biologicamente relevantes
- Como fazer?

Bancos de sequências

- Situação típica
 - Tenho uma sequência consulta
 - Quero saber se existem sequências já publicadas que são “parentes” dela
- Tenho que fazer uma busca em bancos de sequências

Bancos de sequências

- Resultado do sequenciamento em geral é publicado
- “bancos de dados” de sequências
- Na verdade **catálogos**
- Mais importante: **GenBank**
 - Mantido pelo *National Center for Biotechnological Information*
 - **NCBI**
 - <http://www.ncbi.nlm.nih.gov>



National Center for
Biotechnology Information

Search All Databases

Search

Clear

NCBI Home

Site Map (A-Z)

All Resources

Chemicals & Bioassays

Data & Software

DNA & RNA

Domains & Structures

Genes & Expression

Genetics & Medicine

Genomes & Maps

Homology

Literature

Proteins

Sequence Analysis

Taxonomy

Training & Tutorials

Variation

Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

[About the NCBI](#) | [Mission](#) | [Organization](#) | [Research](#) | [RSS Feeds](#)

Get Started

- [Tools](#): Analyze data using NCBI software
- [Downloads](#): Get NCBI data or software
- [How-To's](#): Learn how to accomplish specific tasks at NCBI
- [Submissions](#): Submit data to GenBank or other NCBI databases

Human Microbiome Project

NIH Roadmap Initiative designed to characterize the community of microorganisms living on and in the human body.



1 2 3 4 5

Popular Resources

- [BLAST](#)
- [Bookshelf](#)
- [Gene](#)
- [Genome](#)
- [Nucleotide](#)
- [OMIM](#)
- [Protein](#)
- [PubChem](#)
- [PubMed](#)
- [PubMed Central](#)
- [SNP](#)

NCBI News

[New NCBI News Issue](#)

08 Jul 2011

Information on the redesigned PopSet resource, as well as new

[Preliminary genomic assemblies from two isolates from the European E. coli outbreak now available](#)

07 Jun 2011

Preliminary genomic assemblies of two isolates are in the

[More...](#)

UniProt <http://www.uniprot.org/>

UniProt

Search Blast Align Retrieve ID Mapping

Search in Query

Search Advanced Search > Clear

WELCOME

The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.

What we provide

UniProtKB	Protein knowledgebase, consists of two sections: <ul style="list-style-type: none">★ Swiss-Prot, which is manually annotated and reviewed.★ TrEMBL, which is automatically annotated and is not reviewed. Includes complete and reference proteome sets .
UniRef	Sequence clusters, used to speed up sequence similarity searches.
UniParc	Sequence archive, used to keep track of sequences and their identifiers.
Supporting data	Literature citations , taxonomy , keywords , subcellular locations , cross-referenced databases and more.

Getting started

- [Text search](#)
- [Sequence similarity searches \(BLAST\)](#)
- [Sequence alignments](#)
- [Batch retrieval](#)
- [Database identifier mapping \(ID Mapping\)](#)



NEWS

UniProt release 2013_01 - Jan 9, 2013

Hereditary sensory and autonomic neuropathy type IA: New dietary hope? | UniRef news

- › [Statistics for UniProtKB: Swiss-Prot · TrEMBL](#)
- › [Forthcoming changes](#)
- › [News archives](#)

Follow @uniprot < 501 followers

SITE TOUR



Learn how to make best use of the tools and data on this site.

PROTEIN SPOTLIGHT

unusual liaisons December 2012

Sex for procreation. It doesn't sound in the least bit eccentric. But how about sex between a flower and an insect? We all know that flowers depend very much on insects to perpetuate their species. It is their answer to a lack of legs or wings...

Estatística de alinhamentos

- Com um banco, temos uma “população” de sequências
- Com essa população, posso criar uma teoria estatística que vai me permitir separar os alinhamentos **estatisticamente significativos** daqueles obtidos por mero acaso
- Diremos que os alinhamentos estatisticamente significativos são biologicamente relevantes
- A significância estatística precisa ser quantificada: **e-value**



E-value

- Teoria de **Karlin e Altschul**
- Calcula o **e-value** (expect value) de um alinhamento
- $E = Kmne^{-\lambda S}$
- m e n são os tamanhos das sequências
- S é a pontuação
- K e λ são parâmetros
- Um banco de sequências pode ser tratado como uma longa sequência de tamanho n
- A fórmula dá o **número de alinhamentos** que se esperaria obter com pontuação pelo menos S **ao acaso**

e-value

- Não é uma probabilidade
- Pode resultar maior do que 1
- Mas em geral os alinhamentos biologicamente relevantes tem e-value $< 10^{-5}$
- Para valores assim ou menores, o e-value se comporta como uma probabilidade
- p-values e e-values $P = 1 - e^{-E}$

- E-value depende do tamanho do banco
- Não se pode comparar diretamente e-values obtidos de consultas a **bancos diferentes**
- Mas existe uma fórmula de conversão
 - Dado o e-value contra banco X, é possível saber qual seria o e-value contra banco Y
 - Essa mesma fórmula pode ser usada para dar o e-value para comparação de apenas duas sequências entre si (supondo que Y seja genBank)

The statistics of sequence similarity scores	The statistics of PSI-BLAST scores	Iterated profile searches with PSI-BLAST	BLAST Home
--	------------------------------------	--	------------

- The statistics of global sequence comparison
- The statistics of local sequence comparison
- Bit scores
- P-values
- Database searches
- The statistics of gapped alignments
- Edge effects
- The choice of substitution scores
- The PAM and BLOSUM amino acid substitution matrices
- DNA substitution matrices
- Gap scores
- Low complexity sequence regions
- References

▶ Introduction

To assess whether a given alignment constitutes evidence for homology, it helps to know how strong an alignment can be expected from chance alone. In this context, "chance" can mean the comparison of (i) real but non-homologous sequences; (ii) real sequences that are shuffled to preserve compositional properties [1-3]; or (iii) sequences that are generated randomly based upon a DNA or protein sequence model. Analytic statistical results invariably use the last of these definitions of chance, while empirical results based on simulation and curve-fitting may use any of the definitions.

▶ The statistics of global sequence comparison

Unfortunately, under even the simplest random models and scoring systems, very little is known about the random distribution of optimal global alignment scores [4]. Monte Carlo experiments can provide rough distributional results for some specific scoring systems and sequence compositions [5], but these can not be generalized easily. Therefore, one of the few methods available for assessing the statistical significance of a particular global alignment is to generate many random sequence pairs of the appropriate length and composition, and calculate the optimal alignment score for each [1,3]. While it is then possible to express the score of interest in terms of standard deviations from the mean, it is a mistake to assume that the relevant distribution is normal and convert this Z-value into a P-value; the tail behavior of global alignment scores is unknown. The most one can say reliably is that if 100 random alignments have score inferior to the alignment of interest, the P-value in question is likely less than 0.01. One further pitfall to avoid is exaggerating the significance of a result found among multiple tests. When many alignments have been generated, e.g. in a database search, the significance of the best must be discounted accordingly. An alignment with P-value 0.0001 in the context of a single trial may be assigned a P-value of only 0.1 if it was selected as the best among 1000 independent trials.

Programação dinâmica é **cara**

- Especialmente quando
 - Comparação contra muitas sequências
 - Buscas em banco de dados
 - Comparação de muitas sequências entre si
 - Todas contra todas
- Alternativa: **BLAST**
- Basic Local Alignment Search Tool

BLAST

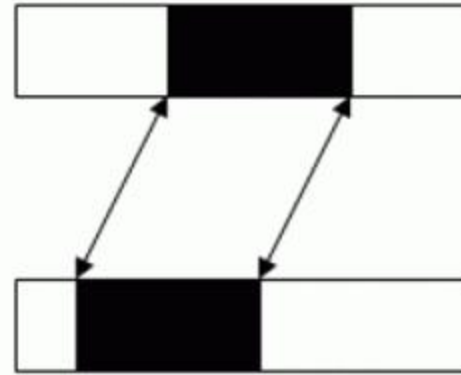
- Altschul et al., 1990, 1997
- Cerca de 75 mil citações (10/2014)
- Programa mais citado em ciência
 - (mas não são os papers mais citados; o mais citado tem 305 mil citações)
- Heurística
 - Não tem **garantia** de que **sempre** consegue achar os alinhamentos de pontuação máxima
 - Sacrifica garantia de otimalidade por velocidade
 - Mas na vasta maioria das vezes tais alinhamentos são de fato encontrados
 - Reporta e-values
 - (É possível fazer cálculo de e-values com PD)

BLAST

- Acha **alinhamentos locais**



global



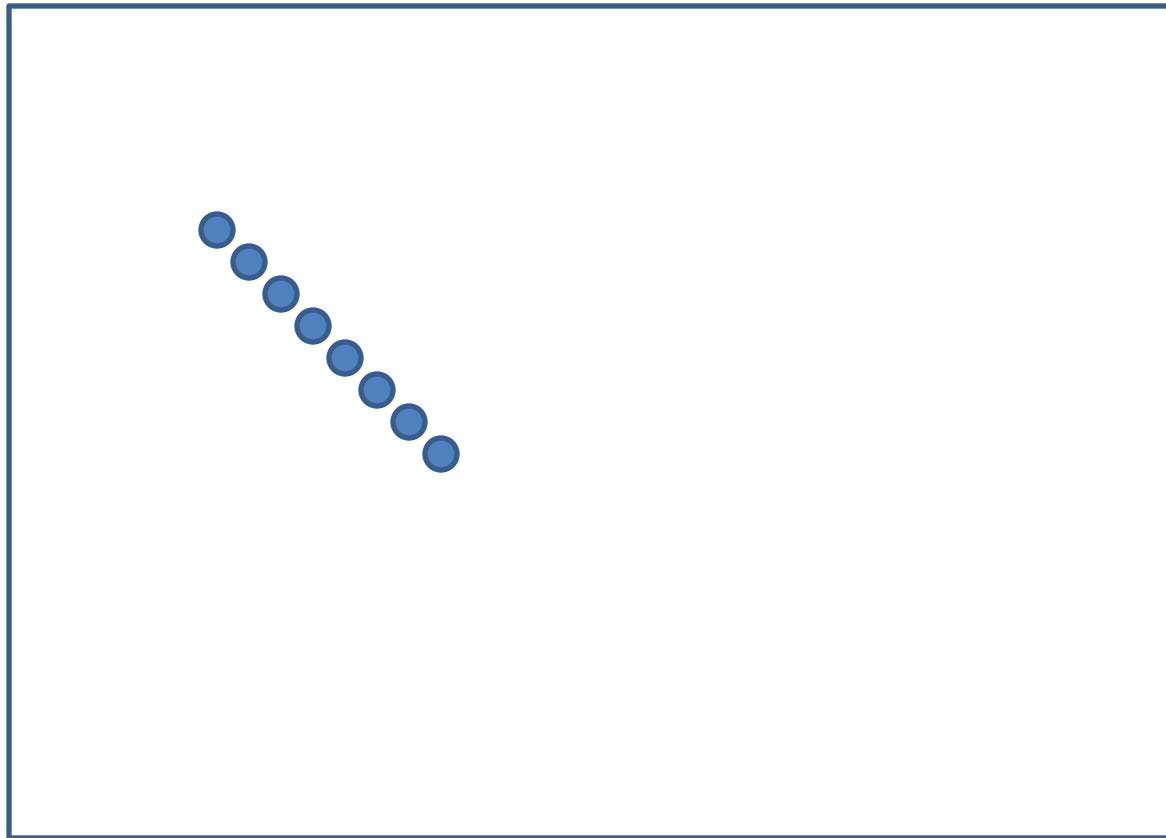
Local alignment

É útil pensar na matriz de PD como se fosse um dotplot

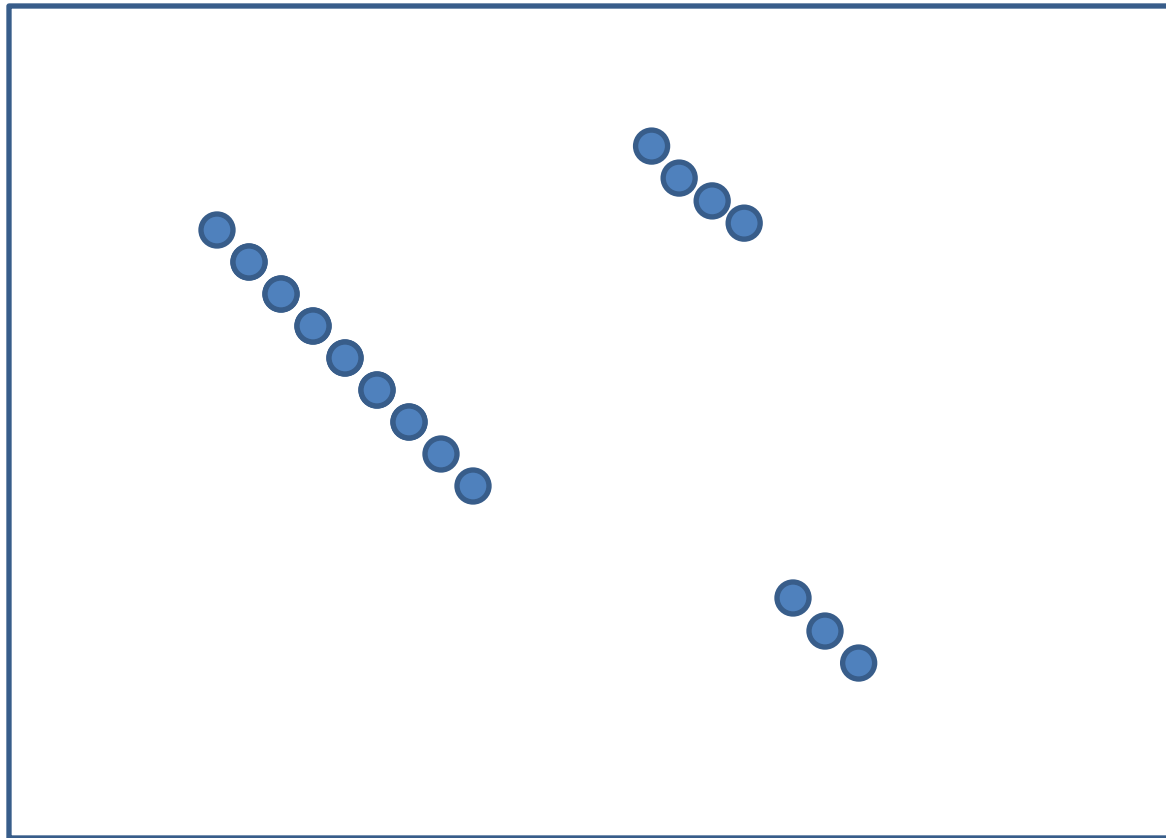
	j	0	1	2	3	4
i		t	G	A	T	C
0	s	0	-2	-4	-6	-8
1	G	-2	1	-1	-2	-4
2	T	-4	-1	0	0	-1
3	C	-6	-3	-2	-1	1

	1	2	3	4
1				
2				
3				

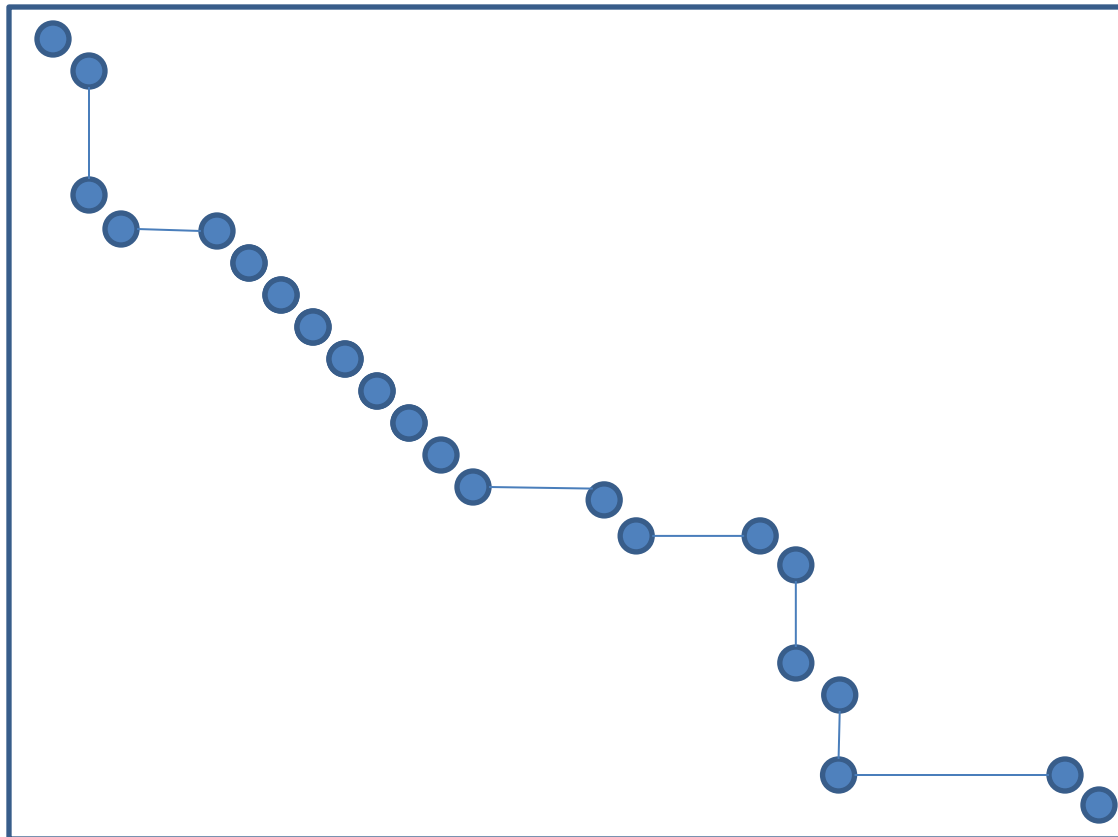
Um alinhamento local é um trecho de células consecutivas num dotplot



Pode haver vários alinhamentos locais



Alinhamento global



Exercício

- Como modificar o algoritmo de PD para que ele encontre o melhor alinhamento local?
- Dica: comece definindo as características do melhor alinhamento local

Alinhamento local com PD

- Um alinhamento global pode ter nota negativa
- um alinhamento local nunca pode ter nota negativa
- Pois o alinhamento entre sequências vazias tem nota zero
- bons alinhamentos locais são trechos com **notas positivas** na matriz de PD

Implementação

1. Inicialização da coluna zero e da linha zero com zeros
2. Ao preencher um elemento da matriz, fazer a operação de máximo, mas nunca deixar que o valor escolhido seja negativo

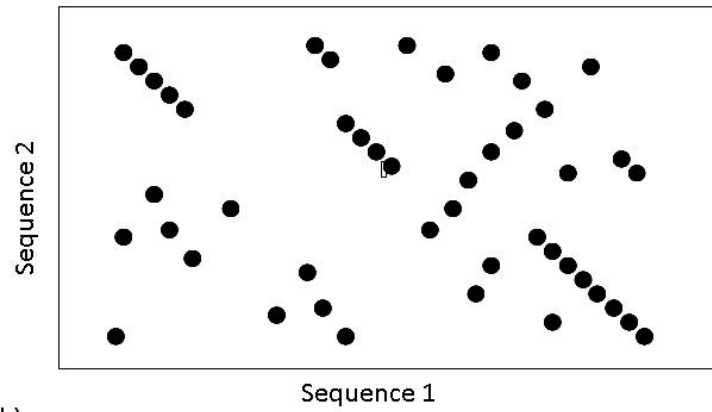
$$\text{Valor} \leftarrow \max (M[i-1,j], M[i, j-1], M[i-1,j-1], 0)$$

3. Ao final: procurar o elemento da matriz de valor **máximo**
4. Para recuperar o alinhamento: percorrer de trás para diante até chegar em zero

De onde vem a eficiência de BLAST?

- BLAST busca trechos parecidos (*palavras* ou *words*) entre as sequências = **alinhamentos-
semente**
- Para nt, esses alinhamentos tem que ser **exatos**
- Para aa, esses alinhamentos tem que ter **nota
positiva**
- Estende esses alinhamentos-mente

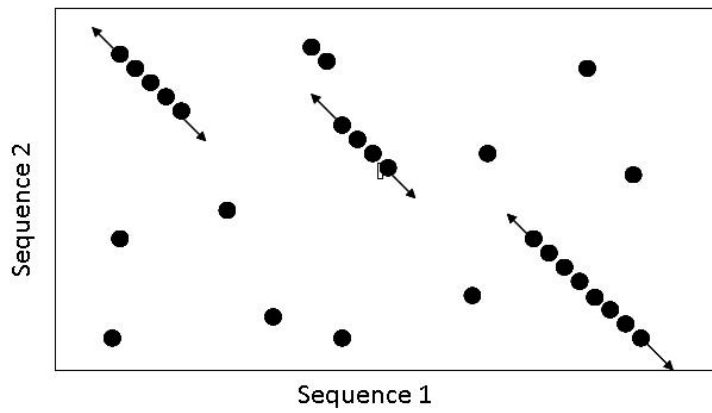
a)



b)



c)



Exemplo de alinhamento ótimo que BLAST não encontraria

- Suponha n e tamanho mínimo de palavras = 4

GTG-TGGCCTA-GAAGCT

GTGGTCG-CTACGAA-GT

Características de BLAST

- Tamanho default das palavras
 - DNA: 11 nt
 - Proteínas: 3 aa
- Reporta bit score, raw score, e-value, identidades, positivos, buracos

>lcl|35099 t
Length=499

Score = 604 bits (1558), Expect = 0.0, Method: Compositional matrix adjust.
Identities = 301/499 (60%), Positives = 365/499 (73%), Gaps = 25/499 (5%)

```
Query 21 DAACSEAAGDKSAMMHDALFERFSARLKAQVGPEVYASWFARLKLHTVSKSVVRFTVPTT 80
          DA C E ++ LF+ S++L+ QVG +VYASWF RLK +VS ++V +VPT
Sbjct 23 DARCLETTCEE-----LFGNVSSKLEDQVGSVDVYASWFQRLKFRSVSHNIVYLSVPTN 75

Query 81 FLKSWINNRYMDLITSLVQSEDPDVLKVEILVRSASRPVRPAQTEERAQPVQEVGAAPRN 140
          FLK+WI NRY+D IT L Q + VEI+VRS+ + P++T +
Sbjct 76 FLKAWIKNRYIDTITKLFQESISSIQGVEIIVRSAA--LMPSETS-----S 119

Query 141 KSFIPSQSATAPAAQPMAAQATLRQGGSGPLFGSPLDTRFTFDTFVEGSSNRVALAAAKT 200
          S I +A P + P+FGSPLD++F F F+EG SNRVALAAA T
Sbjct 120 SSAIAHTTAKPPIINTGKISTIQQKQSINPVFGSPLDSKFVFSNFIEGSPSNRVALAAAHT 179

Query 201 IAEAGAGA--VRFNPLFIHAGVGLGKTHLLQAIANAANAIDS PRNPRVVYLTAEYFMWRFAT 258
          IAE + + VRFNPLFIHA VGLGKTHLLQAIANAANI N RVVYLTAEYFMWRFAT
Sbjct 180 IAEENSSSCTVRFNPLFIHASVGLGKTHLLQAIANAANAIAKKQNNLRVVYLTAEYFMWRFAT 239

Query 259 AIRDNDALTLKDTLRNIDLLVIDDMQFLOGKMIQHEFCHLLNMLLDSAKQVVVAADRAPW 318
          AIRDN AL KD LRNIDLL+IDDMQFLOGK+IQHEFCHLLN LLDSAKQ+V AADR P
Sbjct 240 AIRDNYALNFKDCLRNIDLLLIDDMQFLOGKLIQHEFCHLLNSLLDSAKQIVAAADRPPS 299

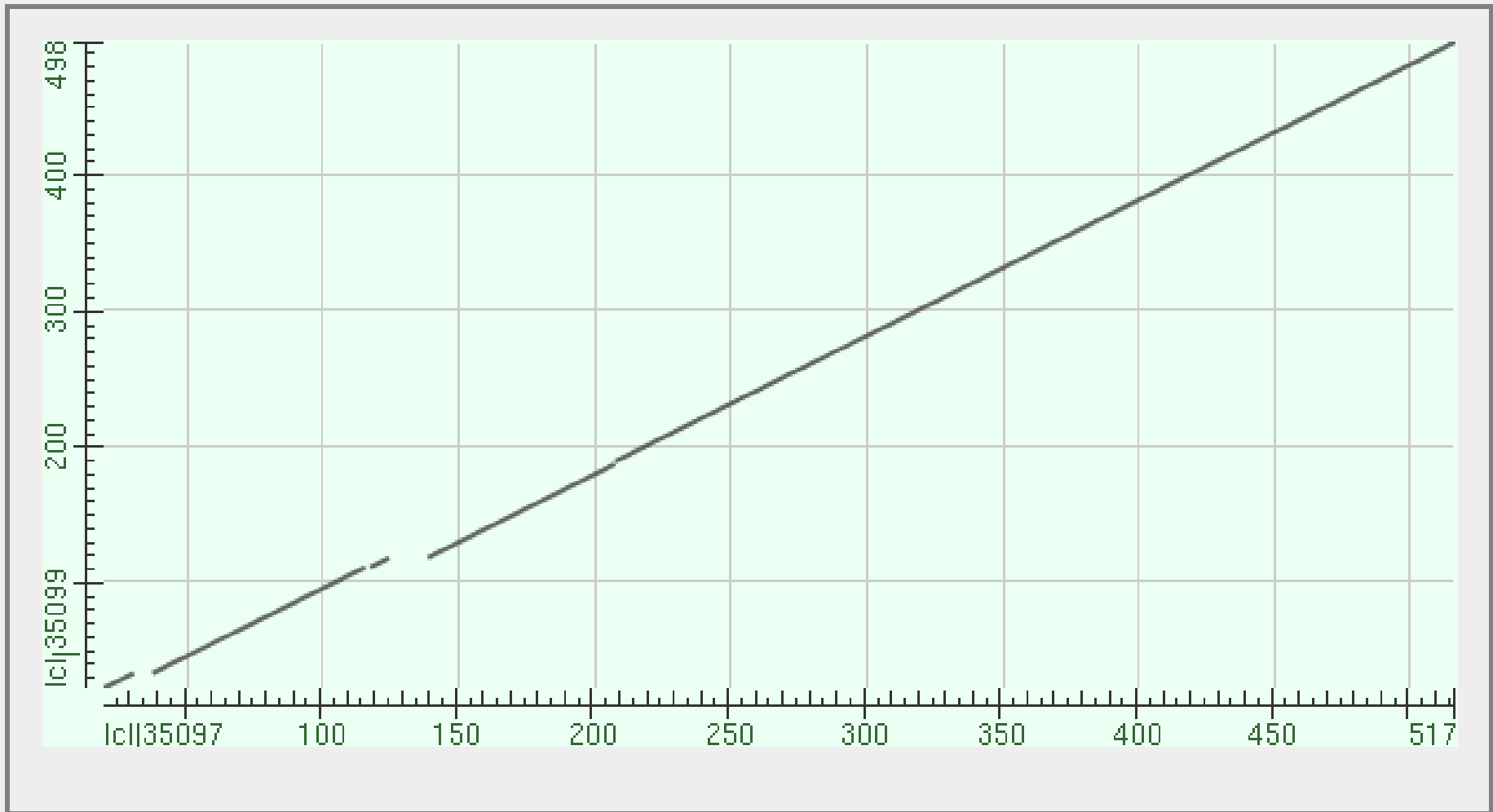
Query 319 ELESLDPRVRSRLQGGMAIEIEGPDYDMRYEMLNRRMGSARQDDPSFEISDEILTHVAKS 378
          ELESLD R+RSRLQGG+A+ + D +MR +L R+ A++D+P IS+EIL VA++
Sbjct 300 ELESLSRIRSRLQGGVAVPLGAHDIEMRLTILKNRLKMAKKDNPPLYISEEILQ RVAQT 359

Query 379 VTASGRELEGAFNQLMFRRSFEPNLSVDRVDELLSHLVSGEAKRVRIEDIQRIVARHYN 438
          VT SGREL+GAFNQL+FR SFEP L++ VDELLSHLV +GE K++RIEDIQR+V++HYN
Sbjct 360 VTTSGRELDGAFNQLVFRNSFEPVLTIKMVDELLSHLVSAGETKKIRIEDIQRMVSKHYN 419

Query 439 VSRQELVSNRRTRVIVKPRQIAMYLAKMLTPRSFPEIGRRFGGRDHTTVLHAVRKIEDLI 498
          +SR +L+SNRR R IV+PRQIAMYL+K++TPRSFPEIGRRFG RDHTTVLHAVRKIE +
Sbjct 420 ISRTDLLSNRRVRTIVRPRQIAMYLSKIMTPRSFPEIGRRFGDRDHTTVLHAVRKIEKSM 479

Query 499 SGDTKLGHEVELLKRLINE 517
          DT + EVELLKRLI+E
Sbjct 480 EKDTVIKKEVELLKR LI SE 498
```

Plot of |c|₃₅₀₉₇ vs 35099



Buscando no GenBank

t (499 letters)

Query ID |d|78035
Description t
Molecule type amino acid
Query Length 499

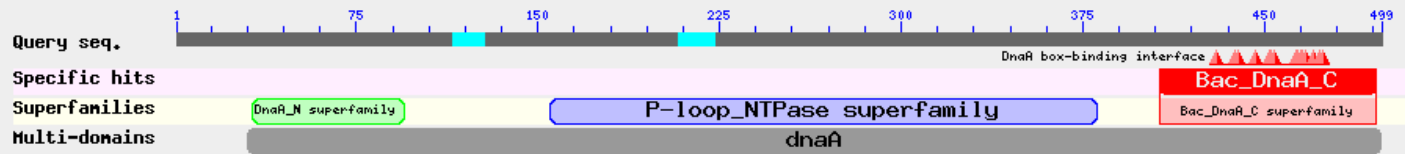
Database Name nr
Description All non-redundant GenBank CDS translations+PDB+SwissProt+environmental samples from WGS projects
Program BLASTP 2.2.26+ [Citation](#)

Other reports: [Search Summary](#) [Taxonomy reports](#) [Distance tree of results](#) [Multiple alignment](#)

Graphic Summary

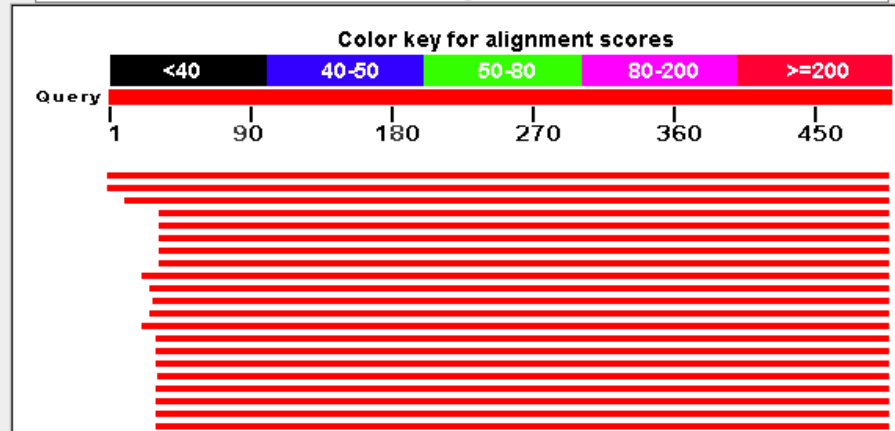
Show Conserved Domains

Putative conserved domains have been detected, click on the image below for detailed results.



Distribution of 100 Blast Hits on the Query Sequence

Mouse over to see the define, click to show alignments



Lista de hits

Descriptions

Legend for links to other resources: [U](#) UniGene [E](#) GEO [G](#) Gene [S](#) Structure [M](#) Map Viewer [P](#) PubChem BioAssay

Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
YP_004062317.1	chromosome replication initiator DnaA [Candidatus Liberibacter solanacearum]	785	785	99%	0.0	76%	G
YP_003065040.1	dnaA gene product [Candidatus Liberibacter asiaticus str. psy62] >gb ACT57	755	755	99%	0.0	76%	G
YP_002543179.1	chromosomal replication initiation protein [Agrobacterium radiobacter K84] >g	615	615	97%	0.0	61%	G
YP_765982.1	dnaA gene product [Rhizobium leguminosarum bv. viciae 3841] >sp Q1MMD6.	613	613	93%	0.0	63%	G
YP_001976569.1	chromosomal replication initiation protein [Rhizobium etli CIAT 652] >gb ACEE	612	612	93%	0.0	63%	G
YP_002973852.1	dnaA gene product [Rhizobium leguminosarum bv. trifolii WSM1325] >gb ACS	612	612	93%	0.0	63%	G
YP_467907.1	chromosomal replication initiation protein [Rhizobium etli CFN 42] >gb ABC89:	611	611	93%	0.0	63%	G
YP_002279530.1	dnaA gene product [Rhizobium leguminosarum bv. trifolii WSM2304] >gb AC15	608	608	93%	0.0	64%	G
EGP58677.1	chromosomal replication initiation protein [Agrobacterium tumefaciens F2]	607	607	95%	0.0	61%	
EHS51424.1	Chromosomal replication initiator protein dnaA [Rhizobium sp. PDO1-076]	605	605	94%	0.0	62%	
EHH08270.1	chromosomal replication initiation protein [Agrobacterium tumefaciens CCNW0	600	600	93%	0.0	62%	
YP_002548273.1	chromosomal replication initiation protein [Agrobacterium vitis S4] >gb ACM3	601	601	94%	0.0	61%	G
NP_353356.2	chromosomal replication initiation protein [Agrobacterium tumefaciens str. C5	598	598	95%	0.0	60%	G
ZP_08526429.1	chromosomal replication initiation protein [Agrobacterium sp. ATCC 31749] >s	595	595	93%	0.0	61%	
YP_004277622.1	chromosome replication initiator DnaA [Agrobacterium sp. H13-3] >gb ADY63	593	593	93%	0.0	61%	G
YP_001325697.1	dnaA gene product [Sinorhizobium medicae WSM419] >gb ABR58862.1 chroi	590	590	93%	0.0	63%	G
ZP_02164856.1	chromosomal replication initiation protein [Hoeflea phototrophica DFL-43] >gb	578	578	93%	0.0	61%	
ZP_05929413.1	chromosomal replication initiator protein dnaA [Brucella abortus bv. 3 str. Tul	578	578	93%	0.0	60%	
P35890.3	RecName: Full=Chromosomal replication initiator protein DnaA	573	573	93%	0.0	62%	
NP_384474.1	chromosomal replication initiation protein [Sinorhizobium meliloti 1021] >ref Y	574	574	93%	0.0	62%	G
AAA26258.1	dnaA [Sinorhizobium meliloti] >gb AAA91097.1 dnaA [Sinorhizobium meliloti]	574	574	93%	0.0	62%	
YP_001608612.1	dnaA gene product [Bartonella tribocorum CIP 105476] >emb CAK00617.1 c	573	573	92%	0.0	59%	G
AFL48605.1	chromosomal replication initiator protein DnaA [Sinorhizobium fredii USDA 257	571	571	93%	0.0	62%	
YP_004547390.1	unnamed protein product [Sinorhizobium meliloti AK83] >gb AEG51776.1 Chr	573	573	93%	0.0	62%	G
YP_002824558.1	chromosomal replication initiation protein [Sinorhizobium fredii NGR234] >gb A	571	571	93%	0.0	62%	G
CBI78638.1	chromosomal replication initiator protein DnaA [Bartonella sp. AR 15-3]	572	572	92%	0.0	60%	
YP_002971177.1	chromosomal replication initiator protein DnaA [Bartonella grahamii as4aup] >	571	571	92%	0.0	60%	G
ZP_10237186.1	chromosomal replication initiation protein. partial [Nitratireductor aauibiodomi	569	569	97%	0.0	56%	

Sabores de BLAST

Subject	nucleotídeos	aminoácidos
Query		
nucleotídeos	BLASTN TBLASTX	BLASTX
aminoácidos	TBLASTN	BLASTP

Também: megablast, psi-blast, phi-blast, delta-blast

Parâmetros de BLASTn

BLAST

Search **database Nucleotide collection (nr/nt)** using **Blastn (Optimize for somewhat similar sequences)**

Show results in a new window

Algorithm parameters

General Parameters

Max target sequences

100

Select the maximum number of aligned sequences to display

Short queries

Automatically adjust parameters for short input sequences

Expect threshold

10

Word size

11

Max matches in a query range

0

Scoring Parameters

Match/Mismatch Scores

2,-3

Gap Costs

Existence: 5 Extension: 2

Filters and Masking

Filter

Low complexity regions

Species-specific repeats for: Homo sapiens (Human)

Mask

Mask for lookup table only

Mask lower case letters

Regiões de baixa complexidade

- Sequências com elementos repetitivos e que aparecem com frequência
- Exemplo em DNA
 - AAAAAAA
- Exemplo em proteína
 - AGNLLGRNVVVGAG
- Uso do filtro é default
- Pode excluir alinhamentos relevantes

Parâmetros de BLASTp

BLAST Search **database Non-redundant protein sequences (nr)** using **Blastp (protein-protein BLAST)**

Show results in a new window

Algorithm parameters

General Parameters

Max target sequences Select the maximum number of aligned sequences to display

Short queries Automatically adjust parameters for short input sequences

Expect threshold

Word size

Max matches in a query range

Scoring Parameters

Matrix

Gap Costs

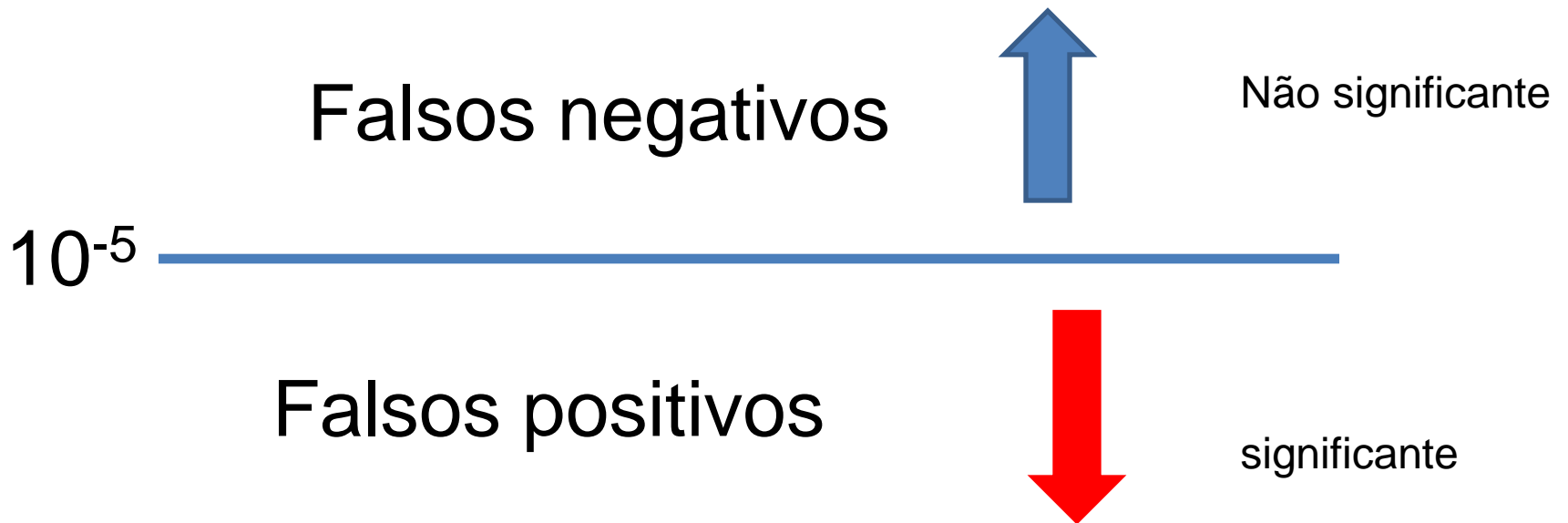
Compositional adjustments

Filters and Masking

Filter Low complexity regions

Mask Mask for lookup table only
 Mask lower case letters

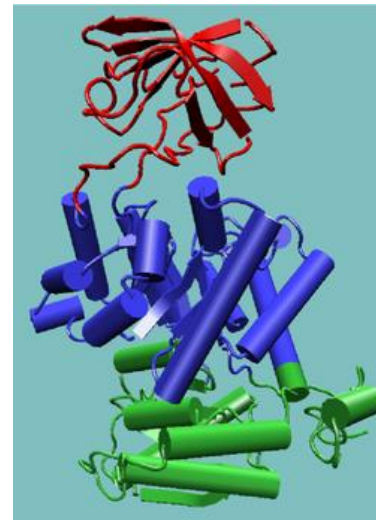
Posso acreditar nos resultados do BLAST?



- 1) Nem todos os alinhamentos estatisticamente significativos são biologicamente relevantes
- 2) Nem todos os alinhamentos que **não são** estatisticamente significantes **não são** relevantes

Exemplo do caso (1)

- Duas proteínas podem compartilhar um domínio e não serem relacionadas
 - falsos positivos de BLAST
- Acontece quando as proteínas tem múltiplos domínios



Pyruvate kinase

Referências



[Ian Korf, Mark Yandell, Joseph Bedell](#)

Artigo por Ingrid Lobo (Write Science Right) © 2008 Nature Education

<http://www.nature.com/scitable/topicpage/basic-local-alignment-search-tool-blast-29096>



Library

Updates

- ▶ New post in [Eyes on Environment: Unique and Alone On the EDGE of Existence](#)
- ▶ New topic in [Women in Science: The First Woman to Win a Fields Medal: Maryam Mirzakhani](#)
- ▶ New post in [Viruses101: Ebola Outbreak Declared an International Public Health Emergency](#)
- ▶ New post in [Accumulating Glitches: Does Biology Have Laws?](#)
- ▶ New post in [Accumulating Glitches: The Language of DNA](#)

Topic Rooms

Within this Subject (21)

- ▶ Comparative Genomics (5)
- ▶ Functional Genomics (4)
- ▶ Genome Sequencing and Annotation (6)
- ▶ Translational Genomics (6)

Other Topic Rooms

- ▶ **Genetics**
 - ▶ Gene Inheritance and Transmission
 - ▶ Gene Expression and Regulation
 - ▶ Nucleic Acid Structure and Function
 - ▶ Chromosomes and Cytogenetics
 - ▶ Evolutionary Genetics
 - ▶ Population and Quantitative

ADVANCED

▶ GENOMICS | Lead Editor: [Michael Goldman](#), [Christopher D. Smith](#)



Basic Local Alignment Search Tool (BLAST)

By: [Ingrid Lobo](#), Ph.D. (*Write Science Right*) © 2008 Nature Education

Citation: [Lobo, I. \(2008\) Basic Local Alignment Search Tool \(BLAST\). *Nature Education* 1\(1\):215](#)



Awash in a sea of data, how do scientists identify the function of a newly cloned gene? Online resources like the Basic Local Alignment Search Tool (BLAST) provide a helping hand.

Aa Aa Aa

Since the discovery of the [genetic code](#), biological research has undergone a sea of change in the way it is performed. Until the early twentieth century, biology focused on the processes of living organisms and almost always involved experiments in laboratories and in the field. The growth of molecular biology during the twentieth century moved research into the test tube, where biological systems could be painstakingly dissected and reassembled. Then, beginning in the 1970s, scientists started accumulating [DNA](#) and [protein](#) sequence data at an exponential rate; in fact, researchers currently have approximately 97 billion bases sequenced and over 93 million records. Amazingly, this sequence data doubles every 18 months!

But how do investigators make sense of this massive amount of data? How can they identify the functions of newly cloned genes? And is it possible to estimate the evolutionary relationships between genes or proteins just by examining their [nucleotide](#) or [amino acid](#) sequences? To address these important issues, researchers must first tease out the relationships between different [species](#) that are descended from a common ancestor. Any sequence similarity can then be used to infer function and evolutionary relationships. In fact, one common method for examining and comparing [genes](#) is to search for similarities between newly sequenced DNA and databases of gene sequences that have already been described. By identifying related genes or gene families with known functions, scientists can infer the functions and evolutionary relationships of newly cloned genes or even whole genomes.

As [gene](#) and protein sequence databases grew at the end of the twentieth century, scientists turned to computers to help analyze this abundant and ever-growing amount of data. Today, one of the most common tools used to examine DNA and protein sequences is the Basic Local Alignment Search Tool, also known as [BLAST](#) ([Altschul et al., 1990](#)). BLAST is a computer algorithm that is available for use online at the [National Center for Biotechnology Information \(NCBI\) website](#), as well as many other sites. BLAST can rapidly align and compare a query DNA sequence with a database of sequences, which makes it a critical tool in ongoing genomic research. In fact, the initial paper describing the program, published in the *Journal of Molecular Biology* and entitled "[Basic Local Alignment Search Tool](#)," was the most highly cited publication of the 1990s ([Taubs, 2000](#)). In recent years, the parallel [development](#) of large-scale sequencing projects and bioinformatic tools like BLAST has enabled scientists to study the genetic blueprint of life across many species, and it has also helped connect biology and computer science in the maturing field of [bioinformatics](#).

Alignment Theory

Although the computer science principles behind BLAST have been around for some time, prior to BLAST, they had not been applied to biology. Before BLAST, alignment programs used [dynamic programming algorithms](#), such as the Needleman-Wunsch and Smith-Waterman algorithms, that required long processing times and the use of supercomputers or parallel computer processors ([Collins & Coulson, 1984](#); [Gotoh & Tagashira, 1986](#); [Smith & Waterman, 1981](#)).

Novos programas

- **Usearch** [Edgar 2010]
 - Até 400x mais rápido do que BLAST
 - Com algum sacrifício de precisão
- **Pauda** [Huson e Xie, 2014]
 - Blastx “dos pobres”
 - 10.000x mais rápido do que blastx!
 - Com mais sacrifício de precisão
- **Diamond** [Buchfink, Xie, Huson 2014]
 - 20.000x mais rápido do que blastx!
 - Sem sacrifício de precisão!

Alinhamento múltiplo

Alinhamento múltiplo

```
C: ----SDIPAGDYEKGGKVKYKQRCLQCHVVDSTAT-KTGPTLHGVIIGRTSGTVSGFDYSAA
Y: ----TEFKAGSAKKGATLFFKTRCLQCHTVEKGGPHKVGPNLHGIFGRHSGQÆEGYSYTDA
A: MASFDEAPPGNPKAGEKIFRTKCAQCHTVEKAGGKHKQGPNLNGLFGRQSGTTPGYSYSAA
D: -----GVPAGDVEKGGKLFVQRCQCHTVEAGGKHKVGPNLHGLIGRKTGQAAAGFA YTDA
H: -----GDVEKGGKIFIMKCSQCHTVEKGGKHKKTGPNLHGLFGRKTGQAPGYSYTAA
M: -----GDVEKGGKIFVQKCAQCHTVEKGGKHKKTGPNLHGLFGRKTGQAAAGFSYTDA
      * . : * .:: . * ** . * : . * ** . * : * : * * : * . * : * : *
```

```
C: NKNKGVVWTKETLFEYLLNPKKYIPGTKMVFA GLKKADERADLIKYIEVESAKSL
Y: NIKKNVLWDENNMSEYLTNPKKYIPGTKMAFGGLKKEKDRNDLITYLKKACE---
A: NKSMAVNWEEKTLFYLLNPKKYIPGTKMVFPGLKKPQDRADLIAYLKEGTA---
D: NKAAGITWNETLFEYLENPKKYIPGTKMIFAGLKKPNERGDLIAYLKSATK---
H: NKNKGIIWGEDTLMEYLENPKKYIPGTKMIFVGIKKKEERADLIAYLKKATNE--
M: NKNKGIIWGEDTLMEYLENPKKYIPGTKMIFAGIKKKGERADLIAYLKKATNE--
      * : * : . . : : ** ***** * * : ** : * *** * : :
```

Para construir filogenias é necessário criar AMs

Avaliação de AMs por soma de pares

- Numa coluna, podemos separar todos os pares de aminoácidos
 - 1 com 2, 1 com 3, 1 com 4, etc
 - 2 com 3, 2 com 4, etc
 - A cada par corresponde uma nota na matriz BLOSUM62
 - A soma de todas as notas dos pares dá a nota da coluna
 - A soma das notas das colunas dá a nota do alinhamento

Coluna de um alinhamento

L
I
V
V
I

BLOSUM62

L/L: 4
L/I: 2
L/V: 1
I/I: 4
I/V: 3
V/V: 4

	L	I	V	V	I	
L		2	1	1	2	6
I			3	3	4	10
V				4	3	7
V					3	3
I						26

Nota da coluna

```

C:      ----SDIPAGDYEGKKVYKQRCLQCHVVDSTAT-KTGPTLHGVI GRTSGTVSGFDYSAA
Y:      ----TEFKAGSAKKGATL FKT RCLQCHTVEKGGPHKVGPNLHGIFGRHSGQAEGYSYTDA
A:      MASFDEAPPGNPKAGEKIFRTKCAQCHTVEKGAGHKQGP NLNGLFGRQSGTTPGYSYSAA
D:      ----GVPAGDVEKGGKLFVQRC AQCHTVEAGGKHKVGP NLHGLIGRKTGQAAGFAYTDA
H:      -----GDVEKGGKIFIMKCSQCHTVEKGGKHKTGP NLHGLFGRKTGQAAGFSYTAA
M:      -----GDVEKGGKIFVQKCAQCHTVEKGGKHKTGP NLHGLFGRKTGQAAGFSYTDA
          * . : * .:: . * *** . * : . * ** . * : * : * * : * . * : * : *

```

What do the consensus symbols mean in the alignment?

An * (asterisk) indicates positions which have a single, fully conserved residue

A : (colon) indicates conservation between groups of strongly similar properties - scoring > 0.5 in the Gonnet PAM 250 matrix

A . (period) indicates conservation between groups of weakly similar properties - scoring ≤ 0.5 in the Gonnet PAM 250 matrix

Não existe padrão universalmente aceito para avaliar AMs

- Ou seja, não existe o equivalente de e-values em BLAST
- Diferentes programas produzem diferentes notas

Multiple Sequence Alignment

Cthe_1566_Clostridium_thermoce	LRDIIENTYK	VLDT-DLOVV	LTGCTAGIVG	DDVDSLVSSEF	AO-----
Fisuc_1086_Fibrobacter_succino	LDOLIKSTLK	VFDG-DLYVV	LTGCVGGLIG	DDVPSLVNEY	RD-----
Metvu_1085_Methanocaldococcus_	LVEGLRNLVA	RYDP-ELISV	VTTCSSETIG	DDIEAFIRAA	RKKIAS ⁺ EFGE
MFS40622_0035	LVEGLRNLVA	RYDP-DLISV	VTTCSSETIG	DDIEAFIRAA	RKKIAAEFGE
Metin_0037_Methanocaldococcus_	LVEGIRNLVA	RYDP-DLISV	VTTCSSETIG	DDIEAFIRAA	RKKIAKEFGE
Csac_2462_Caldicellulosiruptor	LIEGIRNLVL	RYSP-TVIGV	ITTCSETIG	DDIEAFI ⁺ KEA	YKKLSEELSS
Daud_0147_Candidatus_Desulforu	FTEGIRNLVV	RYRP-DLITV	VTTCSSEIIG	DDMVSFIKVA	RKRLVSELGP
Slip_2126_Syntrophothermus_lip	VIEGIRNLVV	RYWP-GLIGV	VTTCSSEIMG	DDMVSFLKEA	RARLSREIGR
CT1536_nifD_Chlorobium_tepidum	LKVAIQEAYD	LFHP-KAIAI	FSTCPVGLIG	DDVHAVAREM	KEKLG ⁺ D----
Cphamnl_1754_Chlorobium_phaeob	LKEAIQEAYD	IFRP-KAIGI	FSTCPVGLIG	DDVHAVAREM	KEKLG ⁺ D----
MM0722_NifD_Methanosarcina_maz	LKKAIDEVVK	IFNP-EAVTI	CATCPVGLIG	DDIEAVSREA	EKEHG ⁺ ----
Avin_01390_nifD_Azotobacter_vi	LAKLIDEVET	LFPLNKGISV	OSECPIGLIG	DDIESVSKVK	GAELS ⁺ ----
Moth_0551_Moorella_thermoaceti	LEOACLEAIR	LFPEAKGLII	FTTCTTGLIG	DDVOAVARSV	EKKTG ⁺ ----
Slip_2127_Syntrophothermus_lip	LKASCLEAFR	LFPEARGMII	FTTCTTGLIG	DDVOGVAROV	EKEVG ⁺ ----
Daud_0146_Candidatus_Desulforu	LLKSALEAVR	LFPEATGIIM	YTTCTTGLIG	DDIGSVAKOI	ERETG ⁺ ----
Metvu_1084_Methanocaldococcus_	LEKACLEAAA	EFPEAKGIII	YATCTTGLIG	DNLGAVAKKV	EEKIG ⁺ ----
MFS40622_0034_Methanocaldococc	LEKACLEAAA	EFPOAKGIII	YATCTTGLIG	DNLEAVARKV	EEKIG ⁺ ----
Metin_0038_Methanocaldococcus_	LEKACIEAAE	EFPEAKGIFI	YATCPTALIG	DNLEAVARKV	EEKIK ⁺ ----
Csac_2463_Caldicellulosiruptor	LYNAIIEANO	EFPEAKAVFI	YATCPTALIG	DDLEAVAKKA	SKAIG ⁺ ----
RoseRS_1199_Roseiflexus	LLOSIIEANA	EFPNAKAVFV	YNTCSTALIG	DDGRDVAKOA	EAIIG ⁺ ----
Rcas_4041_Roseiflexus	LLOSIIEASA	EFPDAKAVFV	YNTCSTALIG	DDGRDVAKOA	EAIIG ⁺ ----
CT1538_nifE_Chlorobium_tepidum	LYKSLIELID	OYOP-NAAFI	YSTCIIGLIG	DDIDAVCKKV	AKEK ⁺ G----
MM0724_nifE_Methanosarcina_maz	LSNAIDELAG	IYRP-PVIFV	YSTCIVGIIG	DDLEAVCKTA	SKKHN ⁺ ----
Avin_01450_nifE_Azotobacter_vi	LFHAIROAVE	SYSP-PAVFV	YNTCVPALIG	DDVDVAVCKAA	AERFG ⁺ ----
Cthe_1565_Clostridium_thermoce	LANTIREVYE	RTHA-NAIFV	LTTCAAGIIG	DDVESVCNEA	EEELG ⁺ ----
Fisuc_1087_Fibrobacter	LROTIRDAKE	RFNP-KAIFI	GMACATAIIG	EDIDSIAEEM	EPEVG ⁺ ----
Ccel_1615_Clostridium_cellulol	LVDSLNEVNS	RYNP-KIIAV	LTNCCADIIG	DDVEGCI EGL	PDEIR ⁺ ----
Mlab_1039_Methanocorpusculum_1	LLNKILOECA	SHHP-KFVAI	LGTPVPALIG	CDISGIATEV	FDTTK ⁺ ----
Mlab_1040_Methanocorpusculum_1	LCNAIDELLP	QIQRPKVFLV	YICCVLYLAG	FDEQSTIDEL	KKRNP ⁺ D----

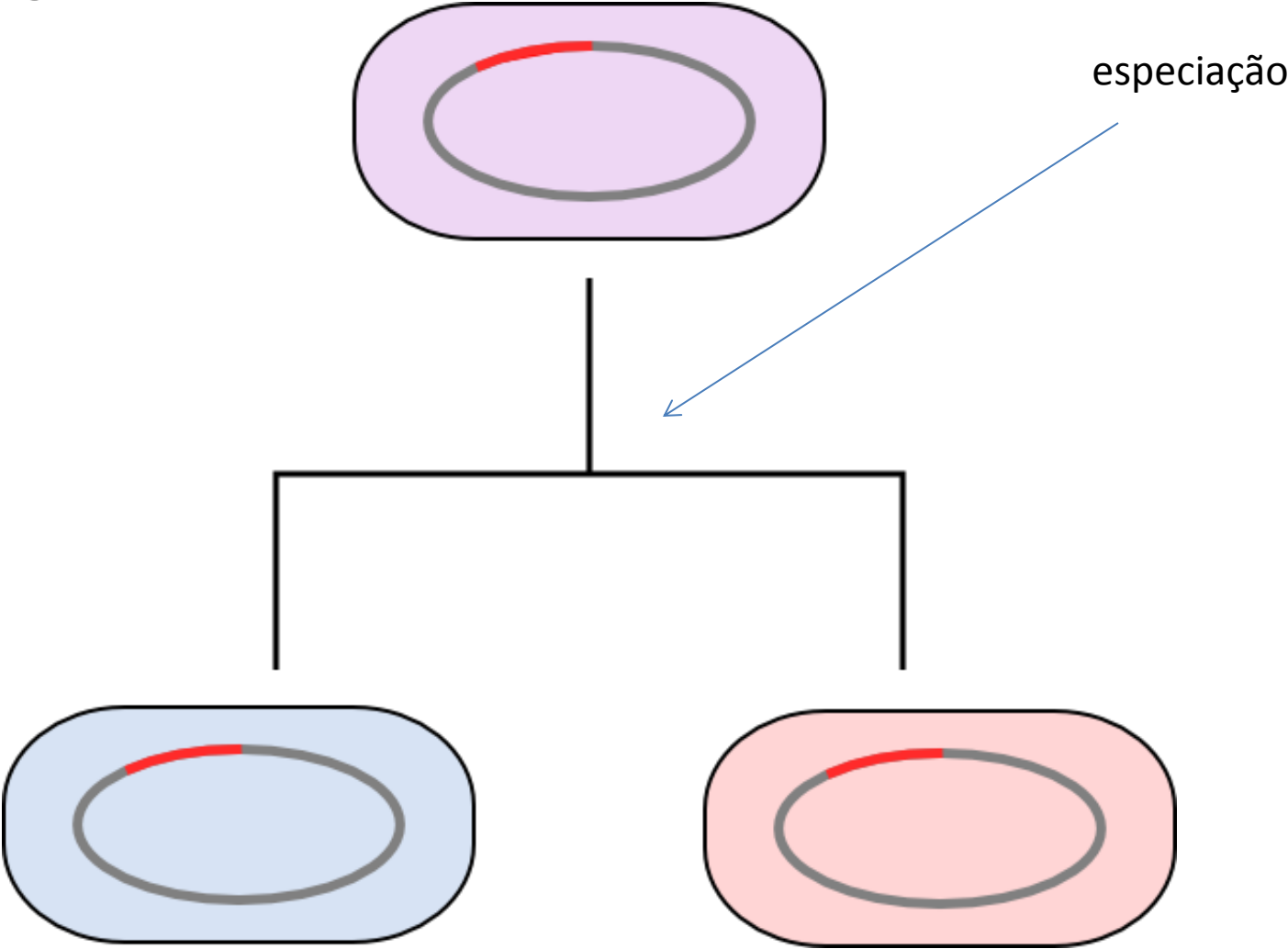
Sequências de entrada

- Dois conceitos importantes
 - Homologia
 - Família

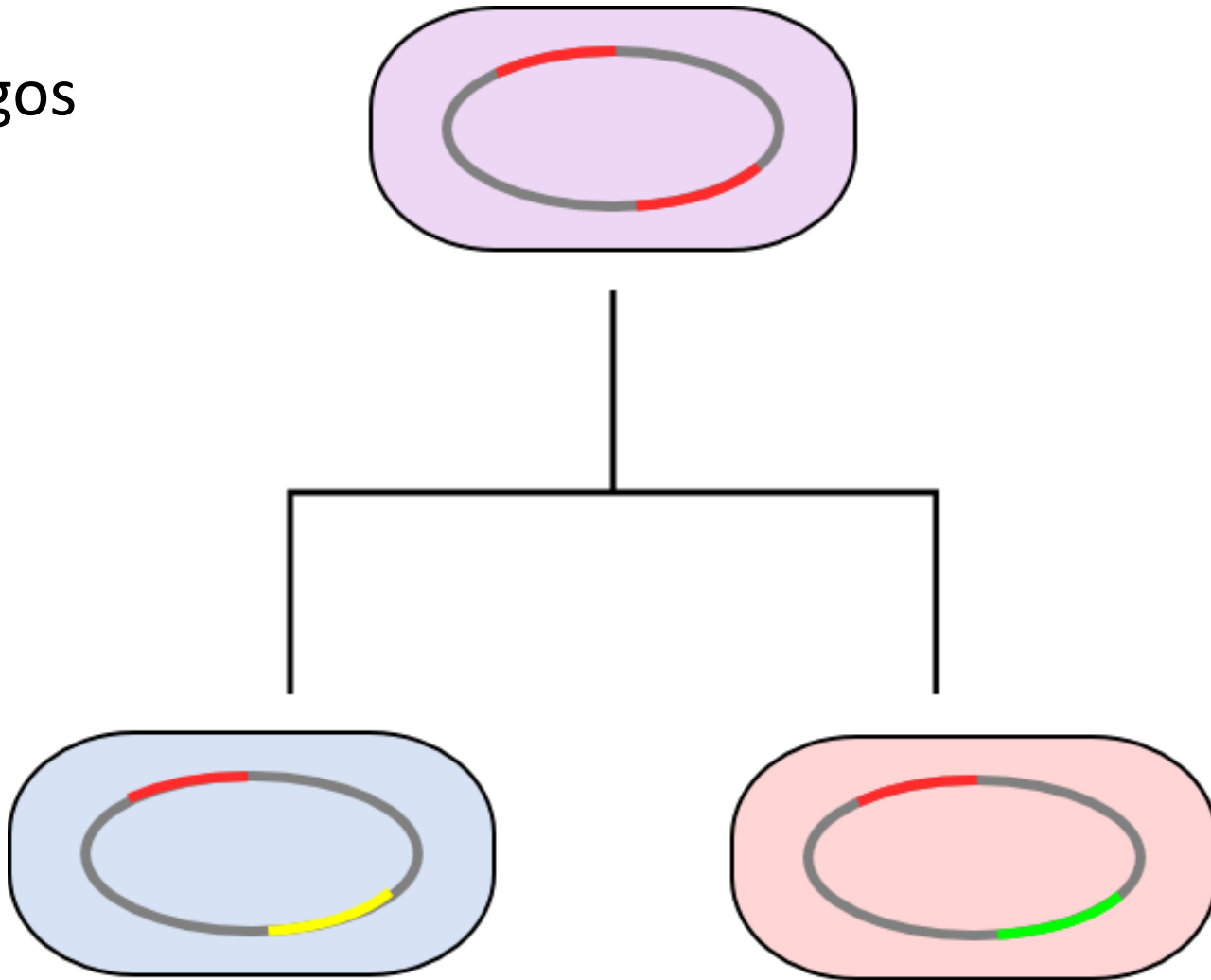
Homologia

- Dois genes que tem um mesmo ancestral são **homólogos**
- Freq. usado erroneamente com o sentido de **similar**
- Similaridade não implica necessariamente em homologia
 - Asas: morcêgo e insetos (convergência)
- Às vezes a similaridade é (ou parece) baixa mas mesmo assim existe homologia
 - Barbatana de baleia e braços em humanos
- Dois tipos de homologia
 - **Ortologia** e **paralogia**

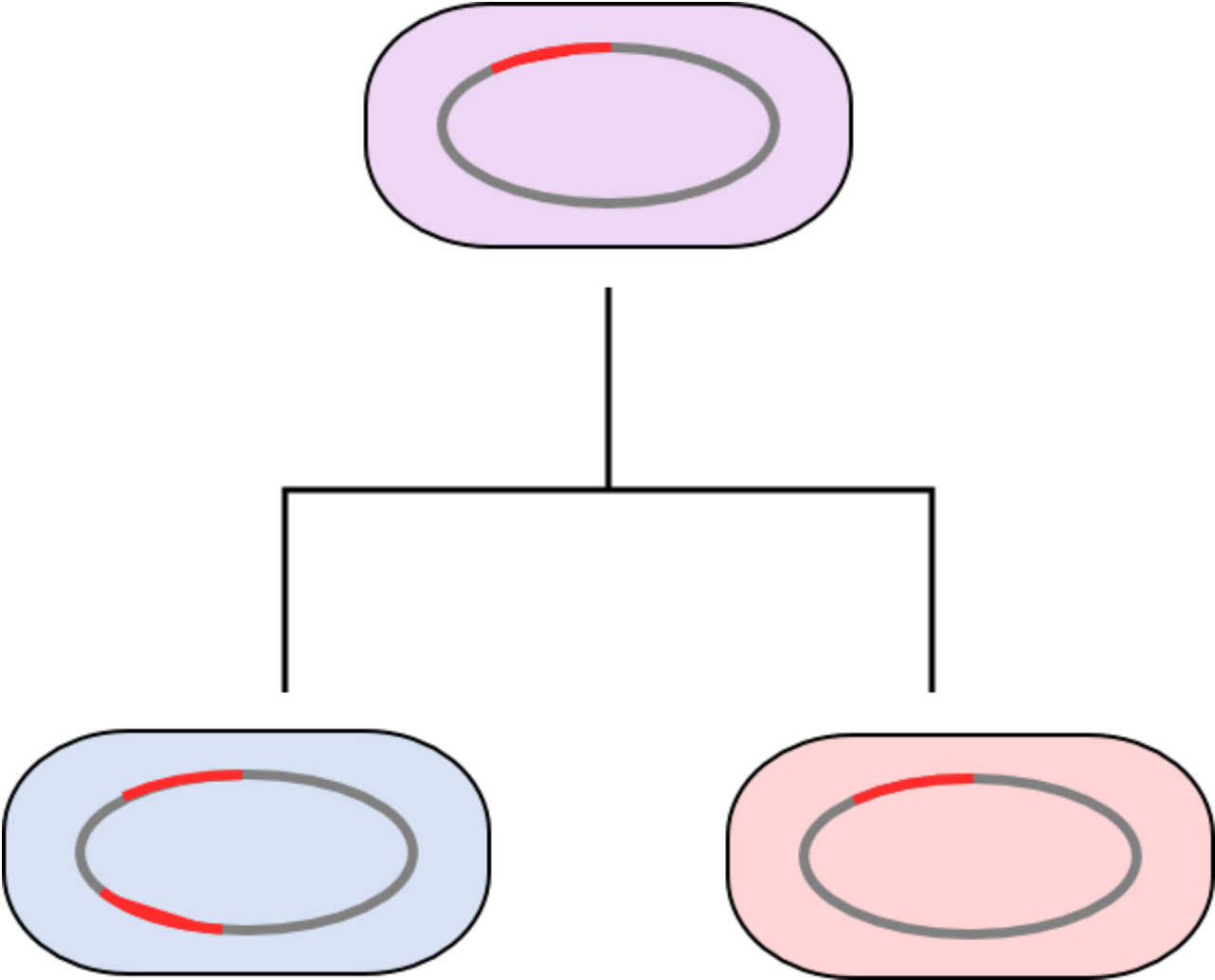
Ortólogos



parálogos



In-parálogos

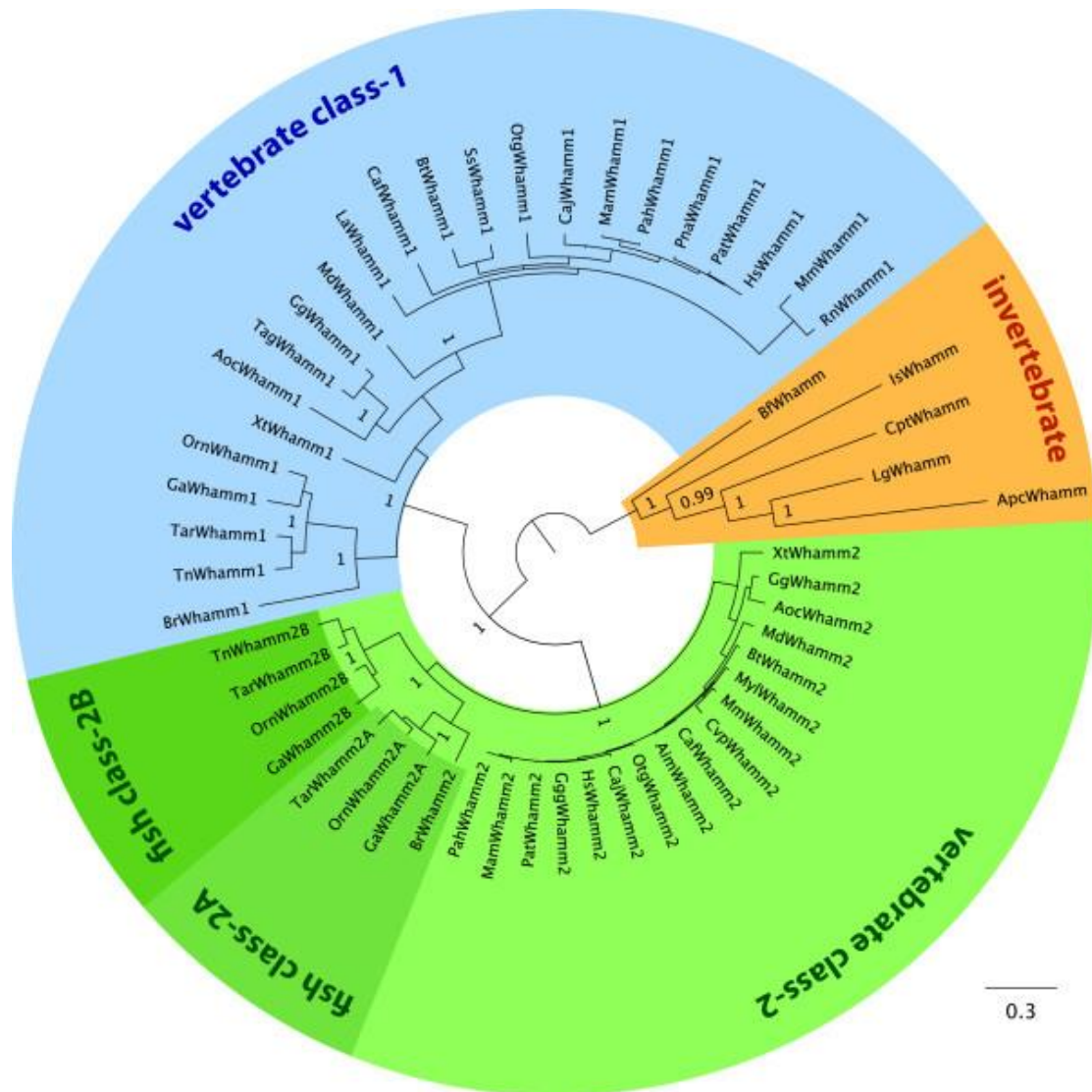


Homologia e função

- Seria bom se proteína homólogas tivessem mesma função
- Geralmente é o caso; mas nem sempre
- Parálogos estão mais sujeitos a desenvolver novas funções
 - Neo-funcionalização
- Na prática
 - Membros de uma mesma família de proteínas são homólogos e em geral tem mesma função
 - Superfamílias e subfamílias

Família de proteínas

- Definição operacional
 - Duas proteínas estão na mesma família se seus genes são homólogos
- ou (mais exigente)
 - Duas proteínas estão na mesma família se seus genes são ortólogos
- Falar em proteínas homólogas é um certo abuso de linguagem: são os genes que são homólogos



Phylogenetic tree of the WHAMM proteins

Kollmar *et al.* *BMC Research Notes* 2012 5:88 doi:10.1186/1756-0500-5-88

Colunas num AM devem ser homólogas

- O gene ancestral comum das sequências no AM também tinha aquela posição

Alinhar DNA ou aminoácidos?

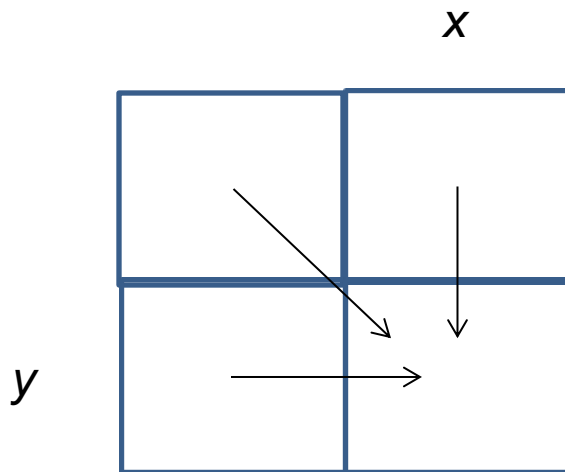
- DNA: mais difícil garantir homologia nas colunas
- DNA é mais sensível, mas a 3ª base de codons não é informativa
- Comparação com aminoácidos permite que proteínas mais distantes possam ser incluídas
 - Há casos em que não dá para alinhar DNA (muita divergência)
- DNA é indicado quando as proteínas são todas idênticas ou quase idênticas
 - Ex: cepas de uma bactéria

Algoritmo para alinhamento múltiplo de sequências

- Programação dinâmica
- Generalização de alinhamento 2-a-2

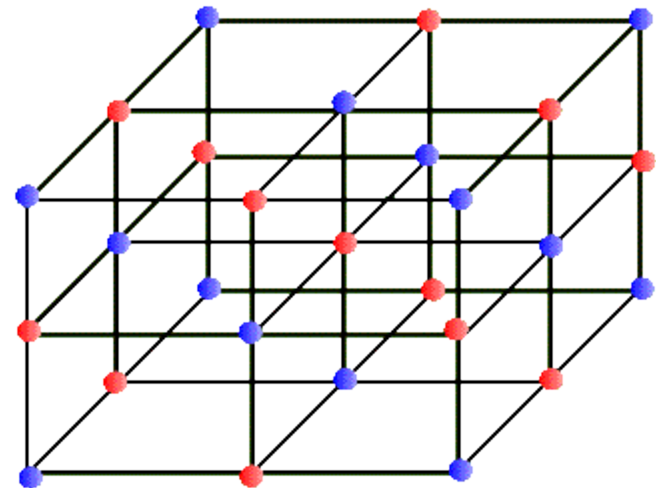
Generalização de PD para AM

2 sequências



$$O(n^2)$$

3 sequências



$$O(n^3)$$

$$\Omega(2^k n^k)$$

Consequência

- Se PD para alinhamentos 2-a-2 já é caro...
- ...para AM é ainda **mais caro!**
- Portanto todos os programas práticos para AM são **heurísticas**
 - Não tem **garantia de otimalidade** (produzem **aproximações**)

Mesmo sendo heurísticas esses programas tem limitações

- As sequências de entrada:
 - Não muito longas (menos do que ~10 kb)
 - Não muitas (menos do que ~500)
 - **esses números variam dependendo do programa e do computador**

Alinhamento progressivo

- Ideia: combinar alinhamentos de pares, iniciando com o par mais similar entre si
- Ir juntando os pares
- Dois estágios
 1. constrói-se uma árvore-guia que determina a hierarquia de similaridade entre os pares
 2. as sequências são adicionadas ao alinhamento num processo guiado pela árvore
- Seria melhor que AM e árvore fossem feitos simultaneamente
 - Muito mais complicado de fazer com rigor

Programas para AM

- **Muscle**

- Edgar, R.C. (2004) *Nucleic Acids Res.* **32**(5):1792-1797
- www.drive5.com/muscle

- **MAFFT**

- Katoh, Misawa, Kuma, Miyata 2002 (*Nucleic Acids Res.* **30**:3059-3066)
- mafft.cbrc.jp/alignment/software/

- **ClustalW/X (antigos) Clustal Omega (novo)**

- Sievers et al. *Molecular Systems Biology* (2011) 7:539
- <http://www.clustal.org/omega/>
- <http://www.ebi.ac.uk/Tools/msa/clustalo/>

- **Outros: Probcons, Cobalt (NCBI), T-coffee**

A Comprehensive Benchmark Study of Multiple Sequence Alignment Methods: Current Challenges and Future Perspectives

Julie D. Thompson , Benjamin Linard, Odile Lecompte, Olivier Poch

Published: March 31, 2011 • DOI: 10.1371/journal.pone.0018093

Article	About the Authors	Metrics	Comments	Related Content
---------	-------------------	---------	----------	-----------------

Download PDF

Print **Share**

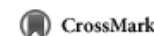
- ▶ Abstract
- Introduction
- Results
- Discussion
- Materials and Methods
- Acknowledgments
- Author Contributions
- References

Reader Comments (0)

Figures

Abstract

Multiple comparison or alignment of protein sequences has become a fundamental tool in many different domains in modern molecular biology, from evolutionary studies to prediction of 2D/3D structure, molecular function and inter-molecular interactions etc. By placing the sequence in the framework of the overall family, multiple alignments can be used to identify conserved features and to highlight differences or specificities. In this paper, we describe a comprehensive evaluation of many of the most popular methods for multiple sequence alignment (MSA), based on a new benchmark test set. The benchmark is designed to represent typical problems encountered when aligning the large protein sequence sets that result from today's high throughput biotechnologies. We show that alignment methods have significantly progressed and can now identify most of the shared sequence features that determine the broad molecular function(s) of a protein family, even for divergent sequences. However, we have identified a number of important challenges. First, the locally conserved regions, that reflect functional specificities or that modulate a protein's function in a given cellular context, are less well aligned. Second, motifs in natively disordered regions are often misaligned. Third, the badly predicted or fragmentary protein sequences, which make up a large proportion of today's databases, lead to a significant number of alignment errors. Based on this study, we demonstrate that the existing MSA methods can be exploited in combination to improve alignment accuracy, although novel approaches will still be needed to fully explore the most difficult regions. We then propose knowledge-enabled, dynamic solutions that will hopefully pave the way to enhanced alignment construction and exploitation in future evolutionary



Subject Areas

- Computer software
- Database searching
- Genome evolution
- Multiple alignment c...
- Sequence alignment
- Sequence analysis
- Sequence databases
- Sequence motif anal...

ADVERTISEMENT



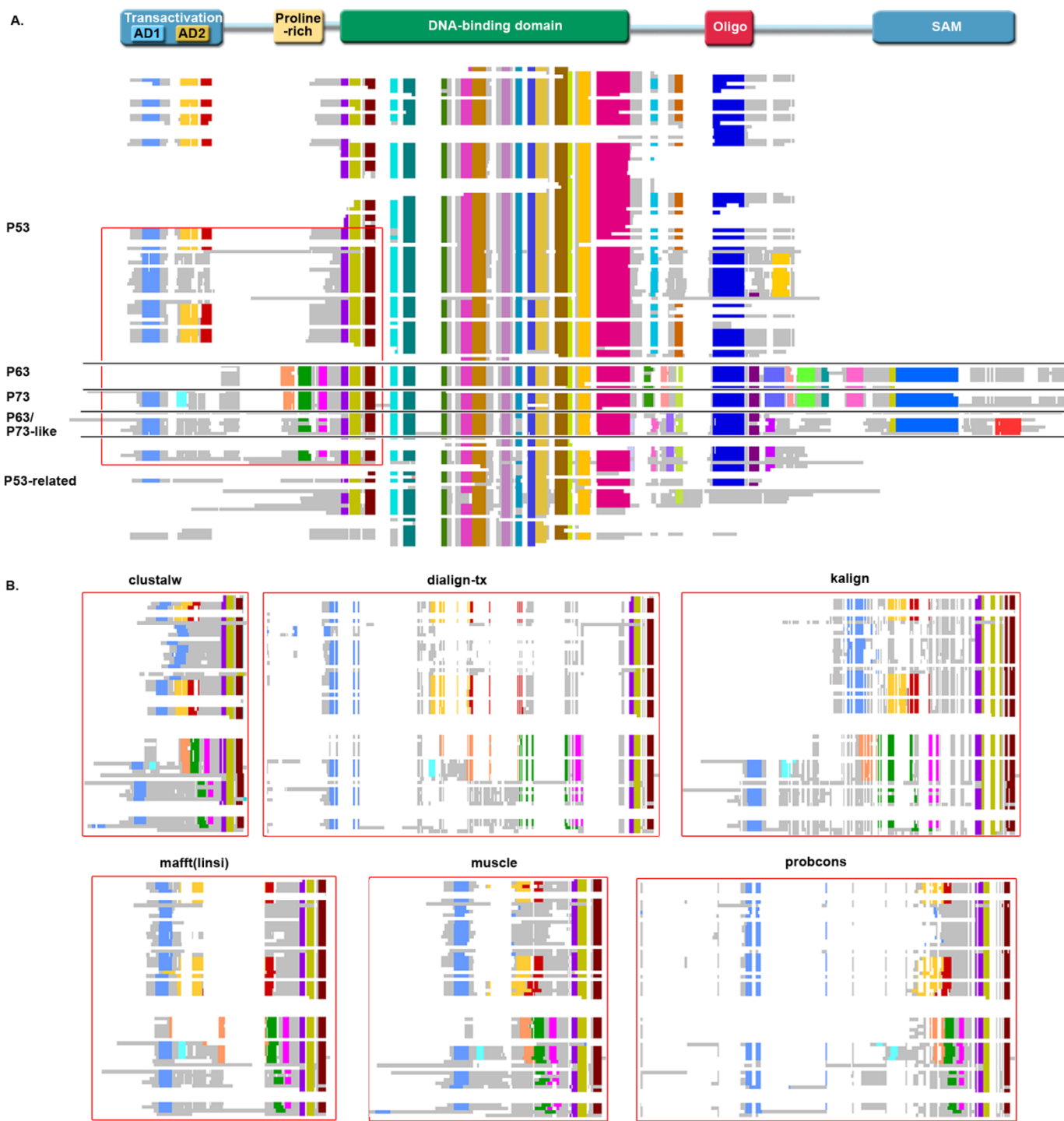


Figure 1: An example benchmark alignment.

(A) Reference alignment of representative sequences of the p53/p63/p73 family, with the domain organization shown above

(B) the alignment (AD: activation domain, Oligo: oligomerization, SAM: sterile alpha motif). Colored blocks indicate conserved regions. The grey regions correspond to sequence segments that could not be reliably aligned and white regions indicate gaps in the alignment. (B) Different MSA programs produce different alignments, especially in the N-terminal region (boxe

(C) regions. The grey regions correspond to sequence segments that could not be reliably aligned and white regions indicate gaps in the alignment. (B) Different MSA programs produce different alignments, especially in the N-terminal region (boxe

(D) gaps in the alignment. (B) Different MSA programs produce different alignments, especially in the N-terminal region (boxe

(E) d in red in A) containing rare motifs and a disordered proline-rich domain.

Esquema
comparativo
de notas

Edição de alinhamentos

Cthe_1566_Clostridium_thermoce	NTVNLIA---	-SIPYQDPFW	KGNLAEYKR-	-----	LLAG	IGLKANVLFQ	
Fisuc_1086_Fibrobacter_succino	GLVNLWF---	-EVPYYNONW	RGDYQELAR-	-----	ILRG	AGFEVNVLFQ	
Metvu_1085_Methanocaldococcus_	MIVPGFG---	-VNPGDILEI	KRILNMFGLK	EGEDYSVLFQ		ISETLYOPLR	
MFS40622_0035	MIVPGFG---	-VNPGDILEI	KRILDMFGLK	EGEDYSVLFQ		ISETLYOPLR	
Metin_0037_Methanocaldococcus_	NIIIPGFG---	-VNPGDILEI	KRILDMFGLK	EGEDYSILFD		ISETLYOPLR	
Csac_2462_Caldicellulosiruptor	YIIPGM---	-INPGDIEEI	KHILDLMKI-	---	DYSVLFQ	ISKTLNSPLM	
Daud_0147_Candidatus_Desulforu	NIIIPGM---	-LMPGDIREI	KHLLAEMGV-	---	EAHVLFQ	ISDVFDTPLM	
Slip_2126_Syntrophothermus_lip	NIIIPGM---	-LYPGDIREI	RHLLLEEMNV-	---	EGTVLFQ	ISDTLDAPLN	
CT1536_nifD_Chlorobium_tepidum	-----	-----	PG	KFKLNLGEGE	---	YNIGG	DAFEIERLLE
Cphamnl_1754_Chlorobium_phaeob	-----	-----	GG	KFKINMLGEGE	---	YNIGG	DAFEIERLLE
MM0722_NifD_Methanosarcina_maz	-----	-----	PT	PFDINIFGEGE	---	YNIGG	DLWEIKPILE
Avin_01390_nifD_Azotobacter_vi	-----	-----	AST	PYDVAIIGD	---	YNIGG	DAWSSRILLE
Moth_0551_Moorella_thermoaceti	-----	-----	ETEKT	PYDICIIGD	---	YNMDW	DLKAIRPLFE
Slip_2127_Syntrophothermus_lip	-----	-----	EEEKT	PYDICLIGE	---	YNMDW	DLOAIRPLFE
Daud_0146_Candidatus_Desulforu	-----	-----	EAERT	PYDVVLVGE	---	YNMDW	DLKVILPLME
Metvu_1084_Methanocaldococcus_	-----	-----	EEEKT	PYDIAIVGE	---	YNMDW	DMAVIKPLFE
MFS40622_0034_Methanocaldococc	-----	-----	EEEKT	PYDIAIVGE	---	YNMDW	DVAVIKPLFE
Metin_0038_Methanocaldococcus_	-----	-----	EEEKT	PYDIAIVGE	---	YNMDW	DLKVIKPLFE
Csac_2463_Caldicellulosiruptor	-----	-----	EDPT	PYDVNIIGE	---	YNIDG	DYWVLEKLFQ
RoseRS_1199_Roseiflexus	-----	-----	EGDF	SRSINIIGD	---	YNIKN	DIRTFEYLFE
Rcas_4041_Roseiflexus	-----	-----	EGDF	SRSINIIGD	---	YNIKN	DIRTFEYLFE
CT1538_nifE_Chlorobium_tepidum	-TEG	-----	IS	KYSINILGE	---	FNLAG	EAWIIREYYE
MM0724_nifE_Methanosarcina_maz	RSESKKPDIK	HSKIEKMEIK	KPKINILGD	---	FNVAG	DVWLVKPLFE	
Avin_01450_nifE_Azotobacter_vi	GSERP	-----	GIR	VHDVNLIGE	---	YNIAG	EFWHVLPPLD
Cthe_1565_Clostridium_thermoce	-----	-----	DDMINVINP	---	WGSD	---	VFYEWFA
Fisuc_1087_Fibrobacter	-----	-----	NDLLNIVAL	---	WGTD	---	YFSEMLA
Ccel_1615_Clostridium_cellulol	-----	-----	NQ	EKSVNLFRR	-WIWDOTKQE		EINEIKRMFO
Mlab_1039_Methanocorpusculum_1	-----	-----		PKTVNILGYT	P-IDFFLSG		DDRRLASFVE
Mlab_1040_Methanocorpusculum_1	-----	-----		TLNFIGN	---	NVPIN	SNSEIYEVLN

Edição de alinhamentos

- Algumas colunas podem não ser informativas
- No olho às vezes é possível ver alinhamentos locais melhores
- Edição manual
- Edição automática

Edição manual de AMs

- **Jalview**

- www.jalview.org

- Waterhouse et al. *Bioinformatics* 2009 **25** (9) 1189-1191

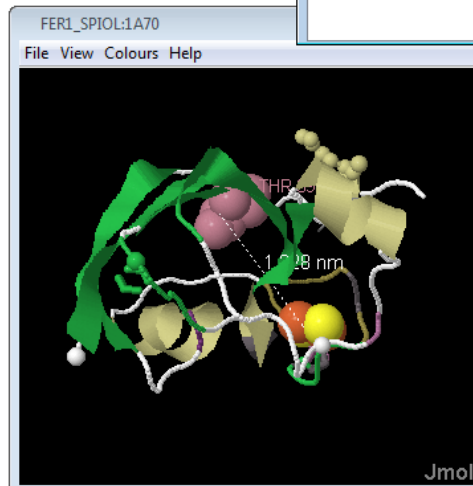
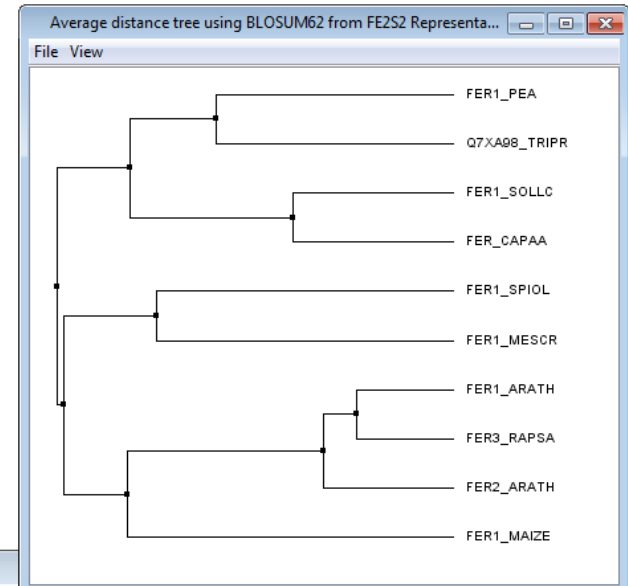
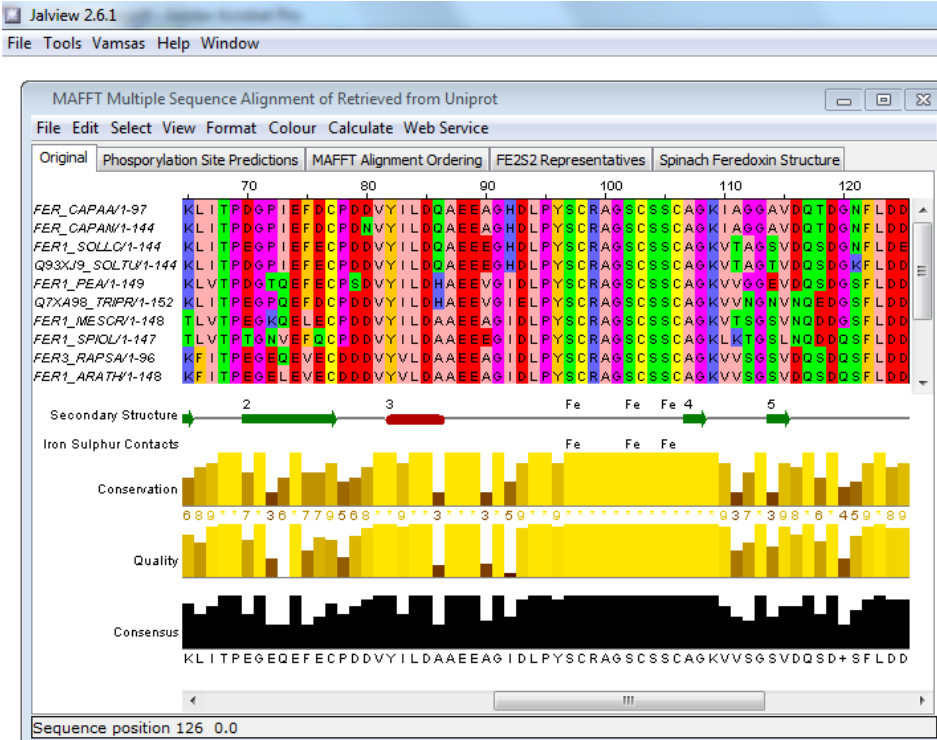
- **Seaview**

- <http://pbil.univ-lyon1.fr/software/seaview.html>

- Gouy M., Guindon S. & Gascuel O. (2010) *Molecular Biology and Evolution* **27(2)**:221-224

JALVIEW

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



SeaView

Version 4.4.2

NEW: seaview drives the **Gblocks** program to select blocks of conserved sites.

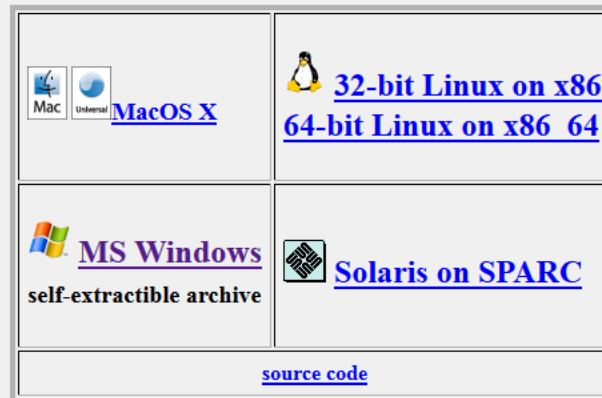
NEW: seaview drives the **Clustal Ω** program to perform multiple sequence alignment.

SeaView is a multiplatform, graphical user interface for multiple sequence alignment and molecular phylogeny.

- SeaView reads and writes various file formats ([NEXUS](#), MSF, CLUSTAL, FASTA, PHYLIP, [MASE](#), Newick) of DNA and protein sequences and of phylogenetic trees.
- SeaView drives programs [muscle](#) or [Clustal Omega](#) for multiple sequence alignment, and also allows to use any external alignment algorithm able to read and write FASTA-formatted files.
- Seaview drives the [Gblocks](#) program to select blocks of evolutionarily conserved sites.
- SeaView computes phylogenetic trees by
 - parsimony, using PHYLIP's [dnapars/protpars](#) algorithm,
 - distance, with [NJ](#) or [BioNJ](#) algorithms on a variety of evolutionary distances,
 - maximum likelihood, driving program [PhyML](#) 3.0.
- SeaView prints and draws phylogenetic trees on screen, SVG, PDF or PostScript files.
- SeaView allows to download sequences from EMBL/GenBank/UniProt using the Internet.

Screen shots of the main [alignment](#) and [tree](#) windows. On-line [help](#) document. Old [seaview version 3.2](#)

Download SeaView



Note for Linux/Unix users: The downloaded archives contain the seaview executable itself, an example data file, a .html file, and 4 other programs (muscle, clustalo, phym1, Gblocks) that seaview drives. These 4 programs and the .html file can either be left in the same directory as seaview, or be put in any directory of your PATH.

Edição automática de AMs

- GBLOCKS

- http://molevol.cmima.csic.es/castresana/Gblocks_server.html
- Castresana, J. (2000) **Molecular Biology and Evolution** 17, 540-552

- GUIDANCE

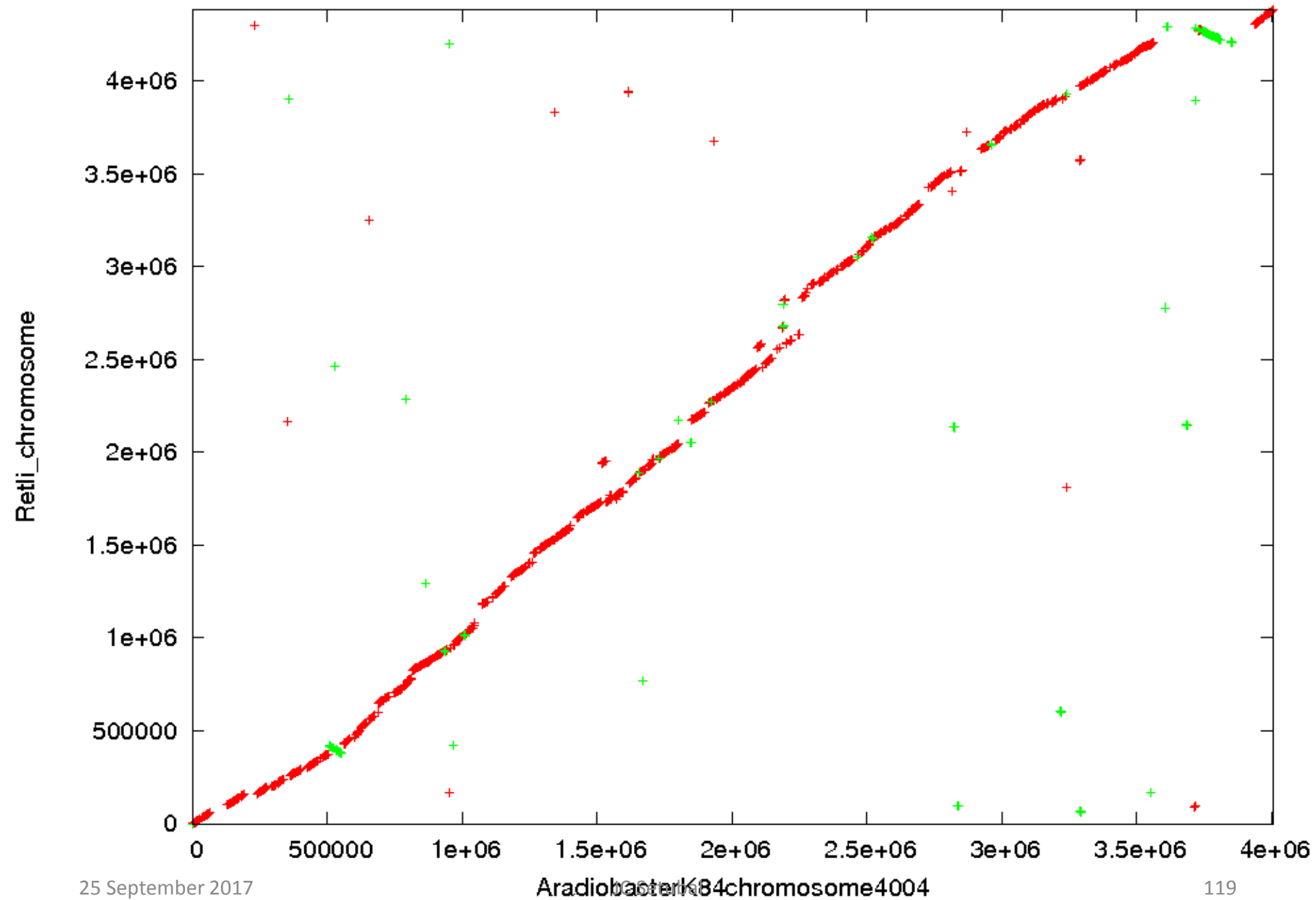
- <http://guidance.tau.ac.il/index.html>
- Penn, O., Privman, E., Ashkenazy, H., Landan, G., Graur, D. and Pupko, T. (2010). **GUIDANCE: a web server for assessing alignment confidence scores.** *Nucleic Acids Research*, 2010 Jul 1; 38 (Web Server issue):W23-W28; doi: 10.1093/nar/gkq443

Formatos de saída

- clustal, FASTA, MSF, NEXUS, PHYLIP
- <http://molecularevolution.org/resources/fileformats/converting>

Alinhamento entre sequências longas

- Cromossomos inteiros
- O cromossomo típico de uma bactéria tem 4 Mbp
- Cromossomo de humanos: 300 Mbp
- Cromossomos e plasmídeos: replicons



BLAST não serve

- Computadores mesmo com dezenas de GB de RAM não dão conta de rodar BLAST para essas entradas
- Problema não é tempo; é **memória RAM**
- Outras abordagens são necessárias

O programa MUMmer

- Delcher AL, Phillippy A, Carlton J, Salzberg SL. **Fast algorithms for large-scale genome alignment and comparison**. *Nucleic Acids Res.* 2002 Jun 1;30(11):2478-83.
- Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. **Versatile and open software for comparing large genomes**. *Genome Biol.* 2004;5(2):R12
- <http://mummer.sourceforge.net>

Como MUMmer funciona

- It finds Maximal Unique Matches
- These are exact matches above a user-specified threshold that are unique
- Exact matches found are clustered and extended (using dynamic programming)
 - Result is approximate matches
- Data structure for exact match finding: **suffix tree**
 - Difficult to build but very fast
- Nucmer and promer
 - Both very fast
 - $O(n + \#MUMs)$, n = genome lengths

Árvore de sufixos para GTATCTAGG

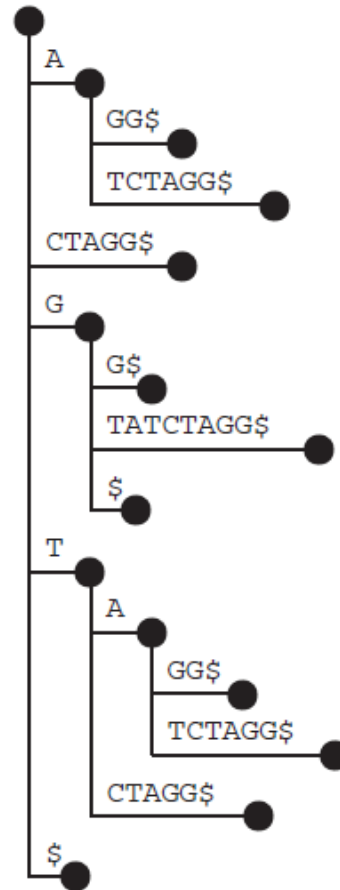


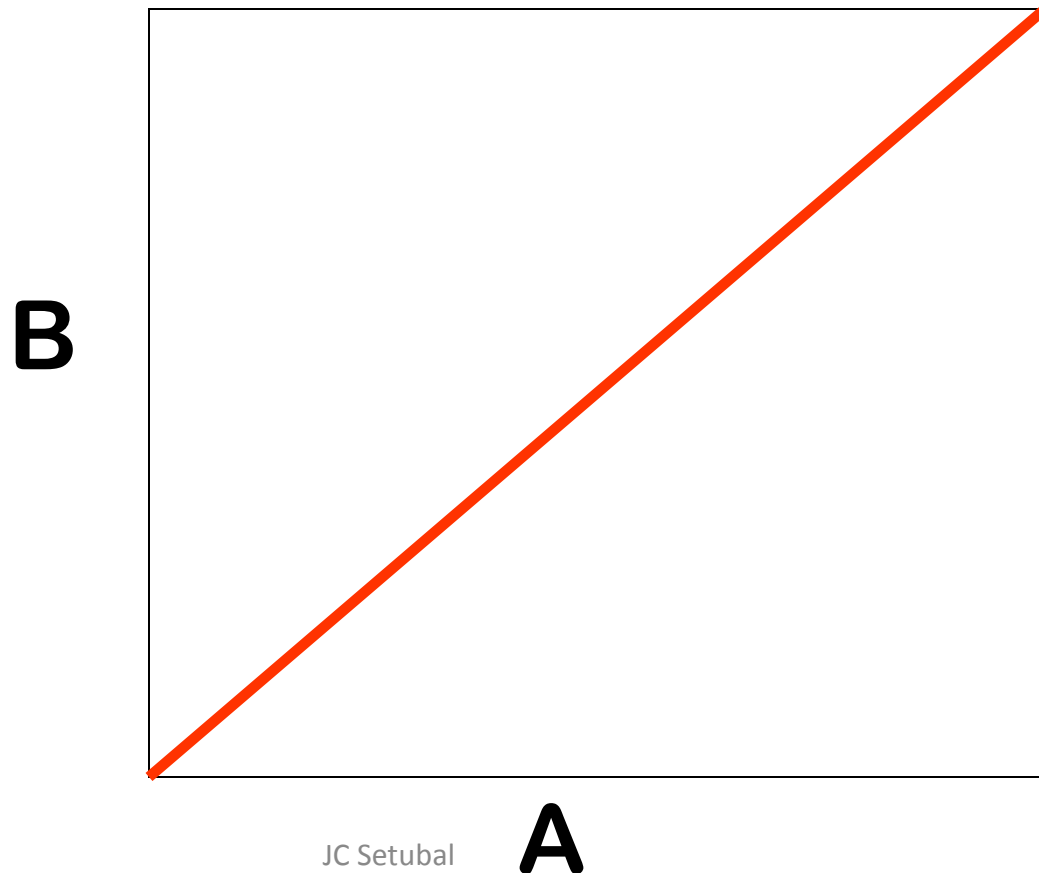
FIGURE 3.19

An example of a suffix tree for string GTATCTAGG. A dollar sign marks the end of the string.

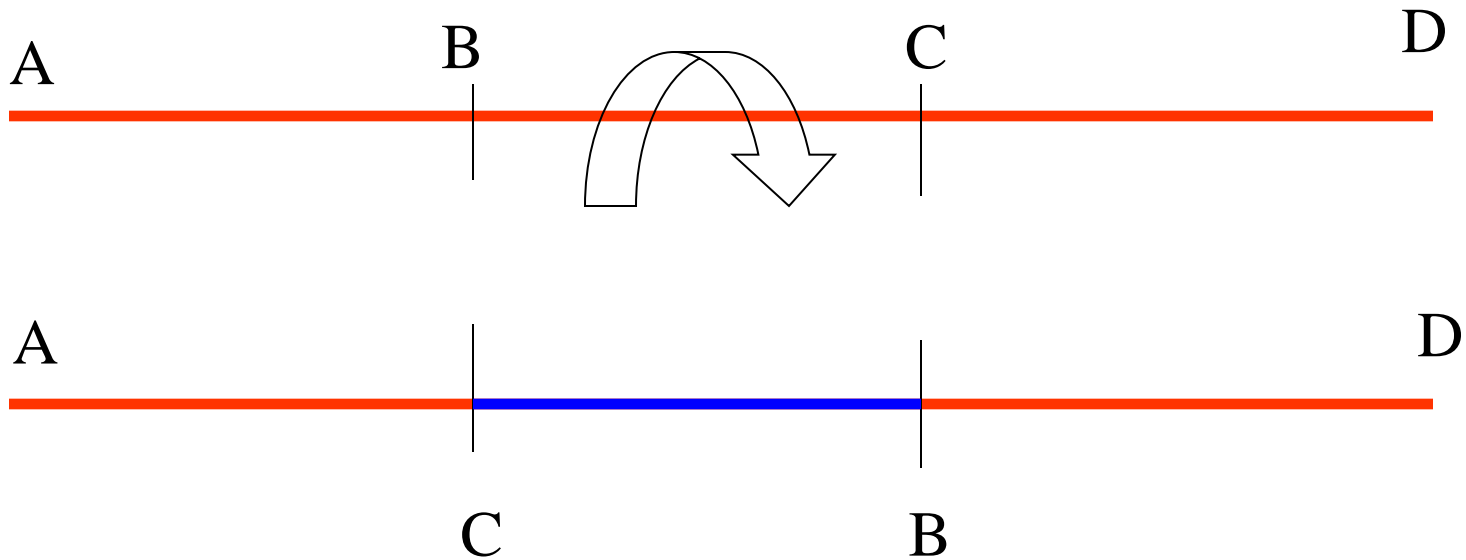
- Alinhamentos de replicons inteiros revelam rearranjos

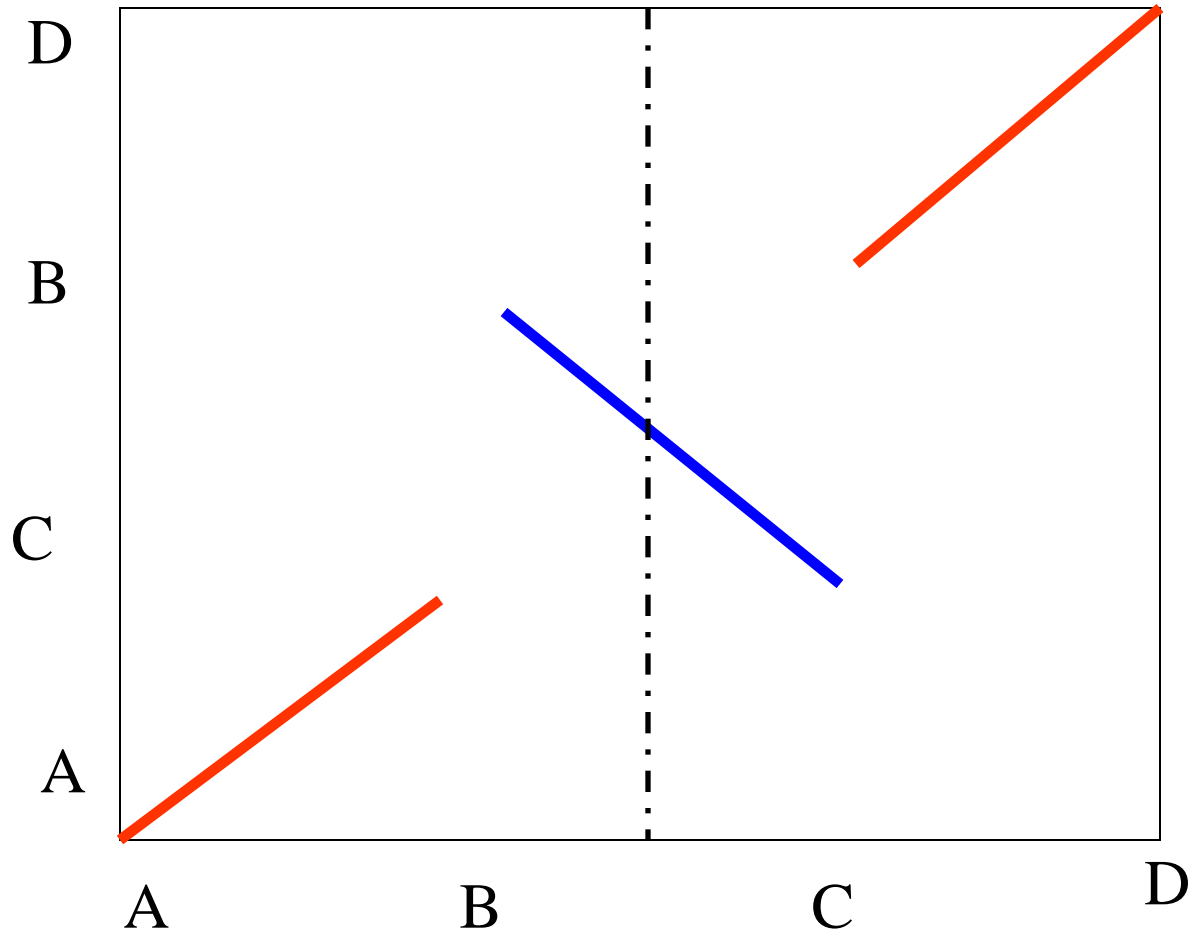
Alinhamentos de pares de replicons completos

Se as sequências fossem idênticas veríamos:

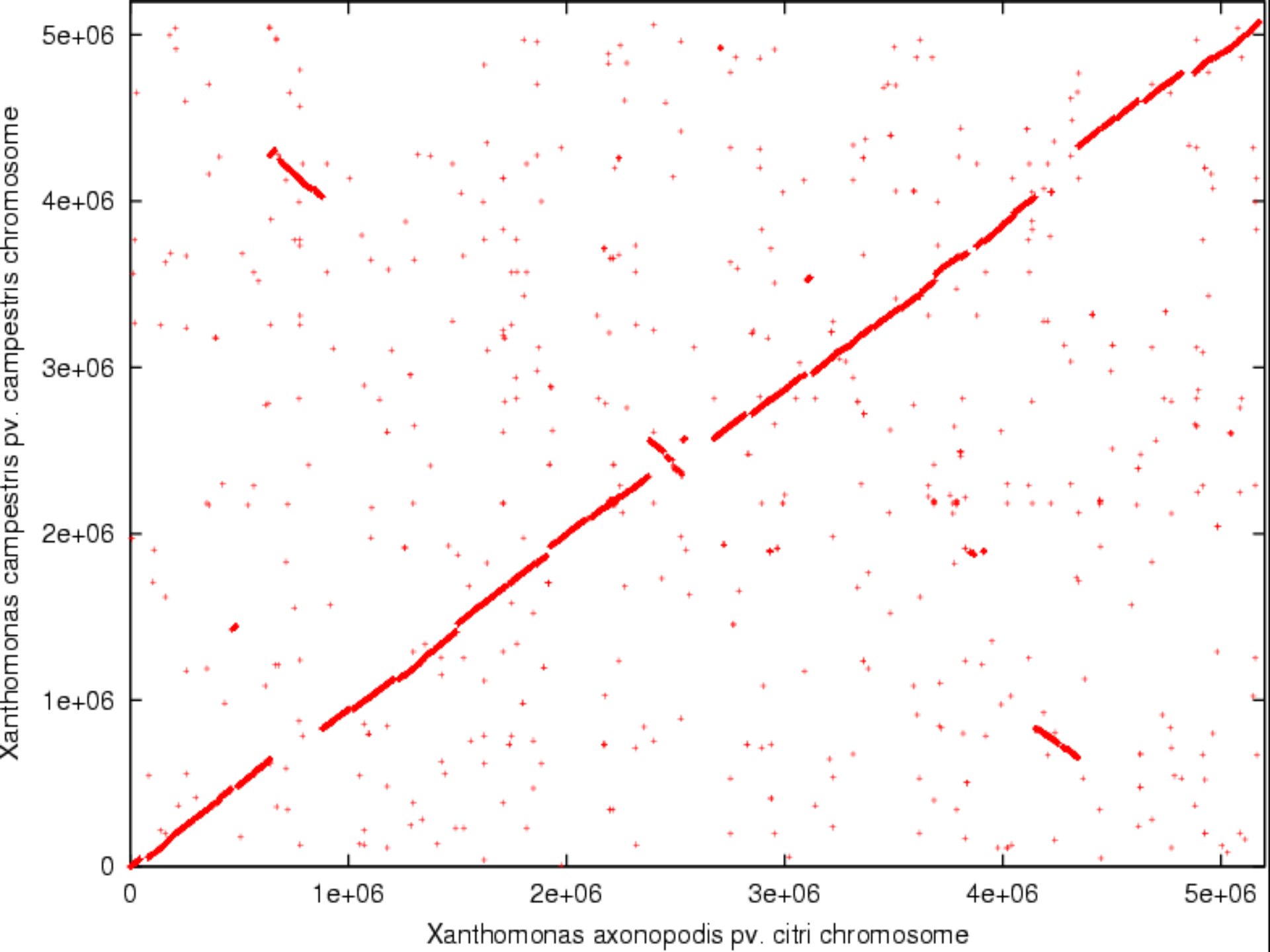


uma *inversão*

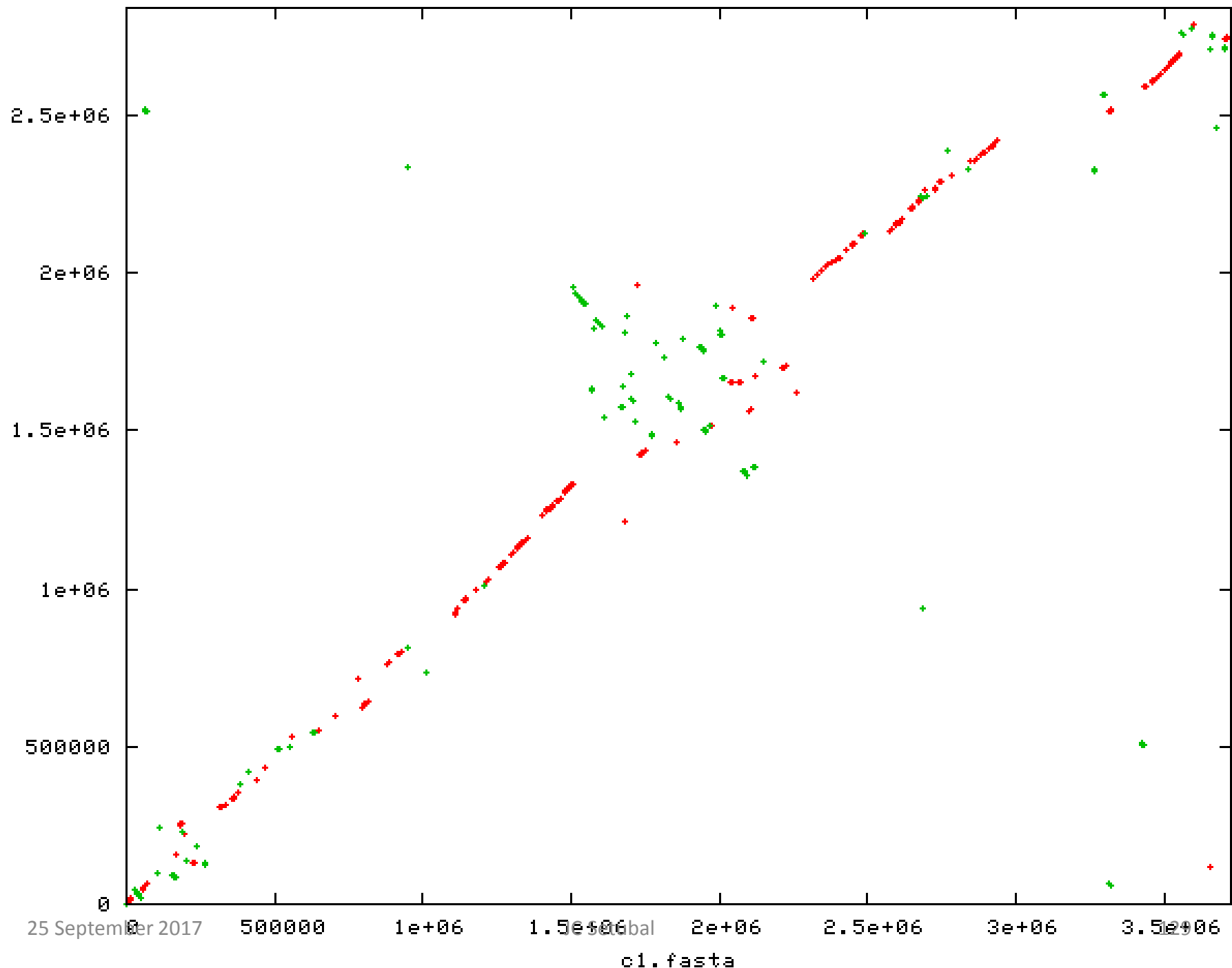




Such inversions seem to happen around
the *origin* or *terminus of replication*



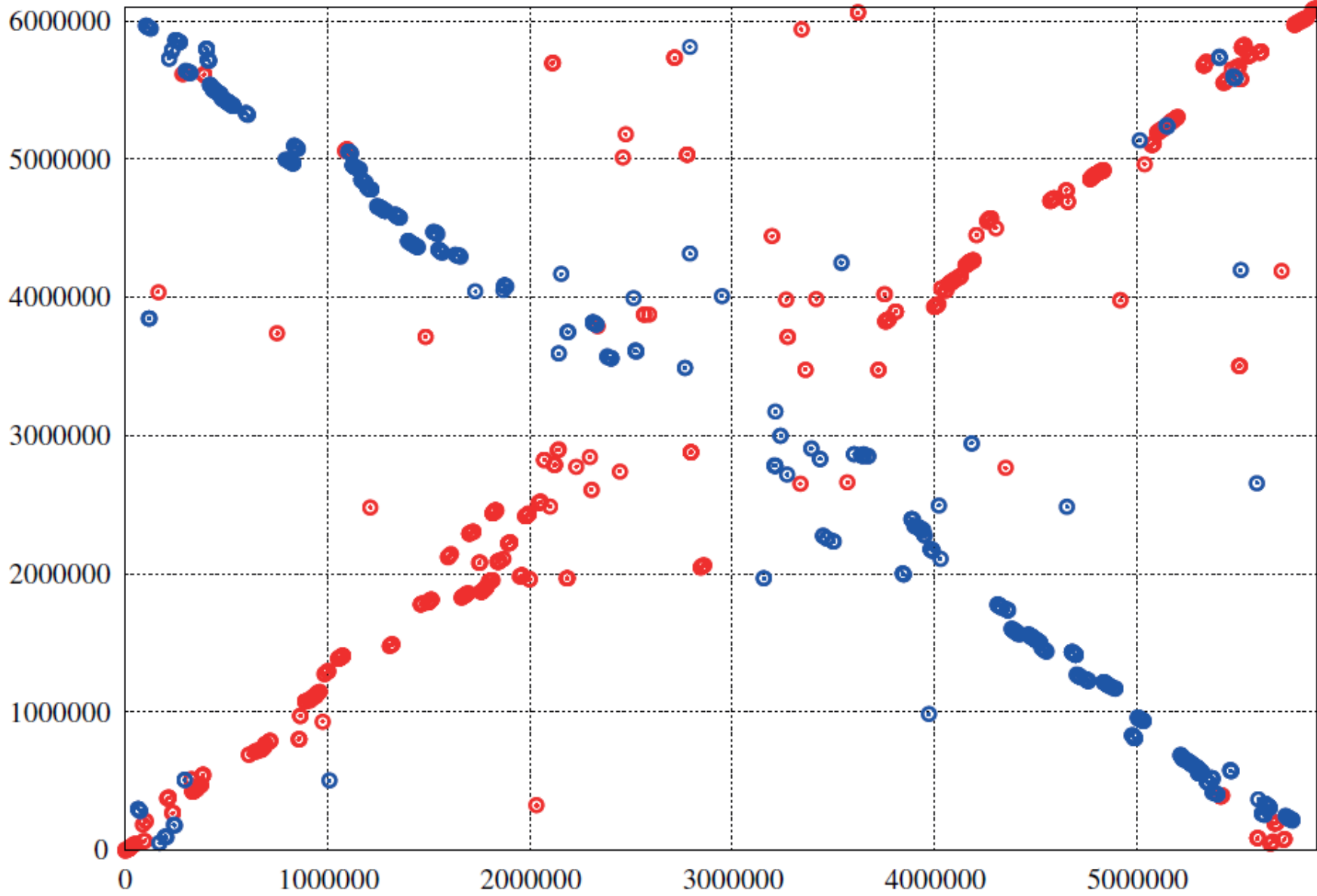
AtuC58-circChrom-1.2



25 September 2017

c1.fasta

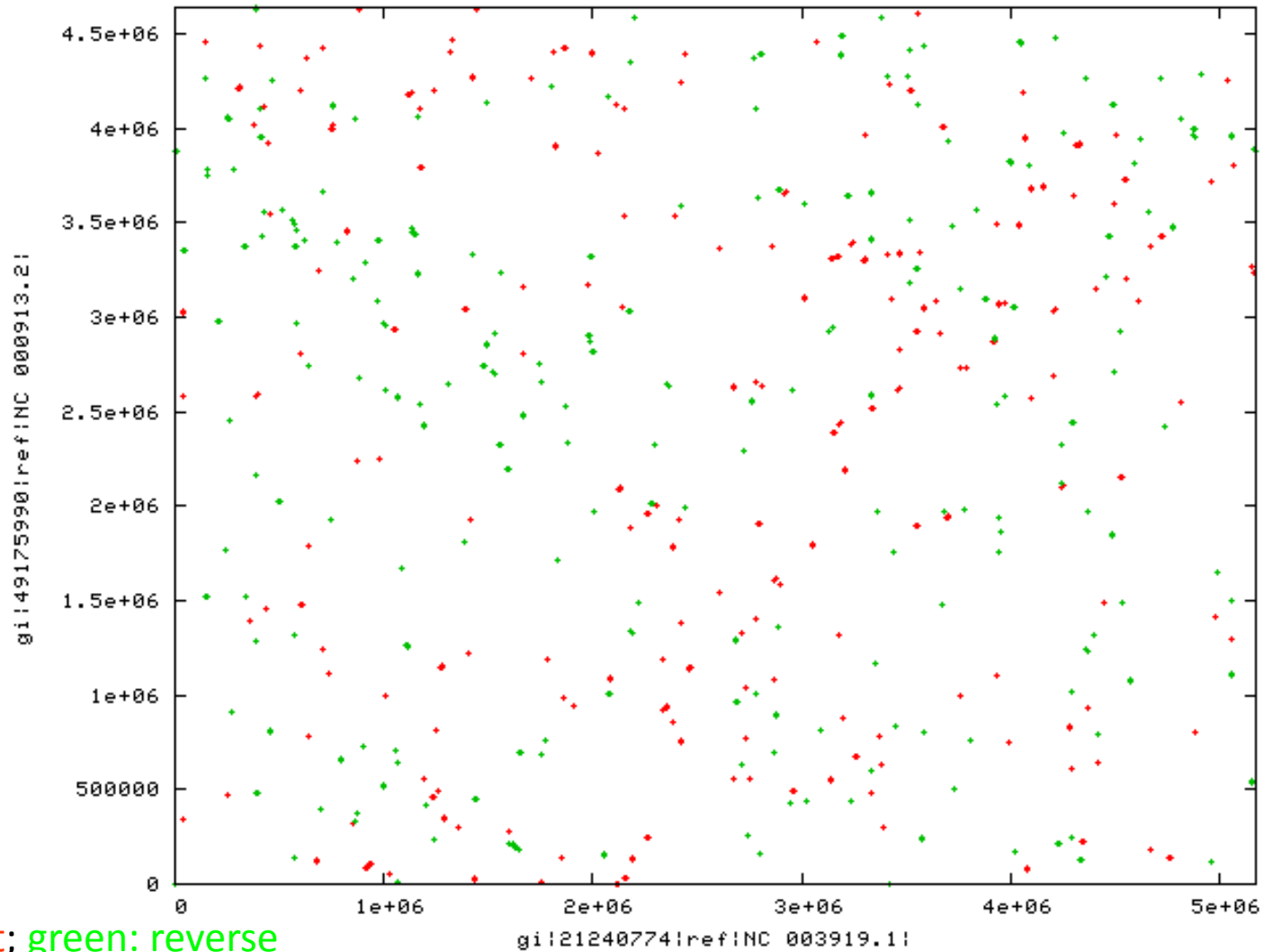
Pseudomonas syringae pv B728a



Pseudomonas entomophila L48

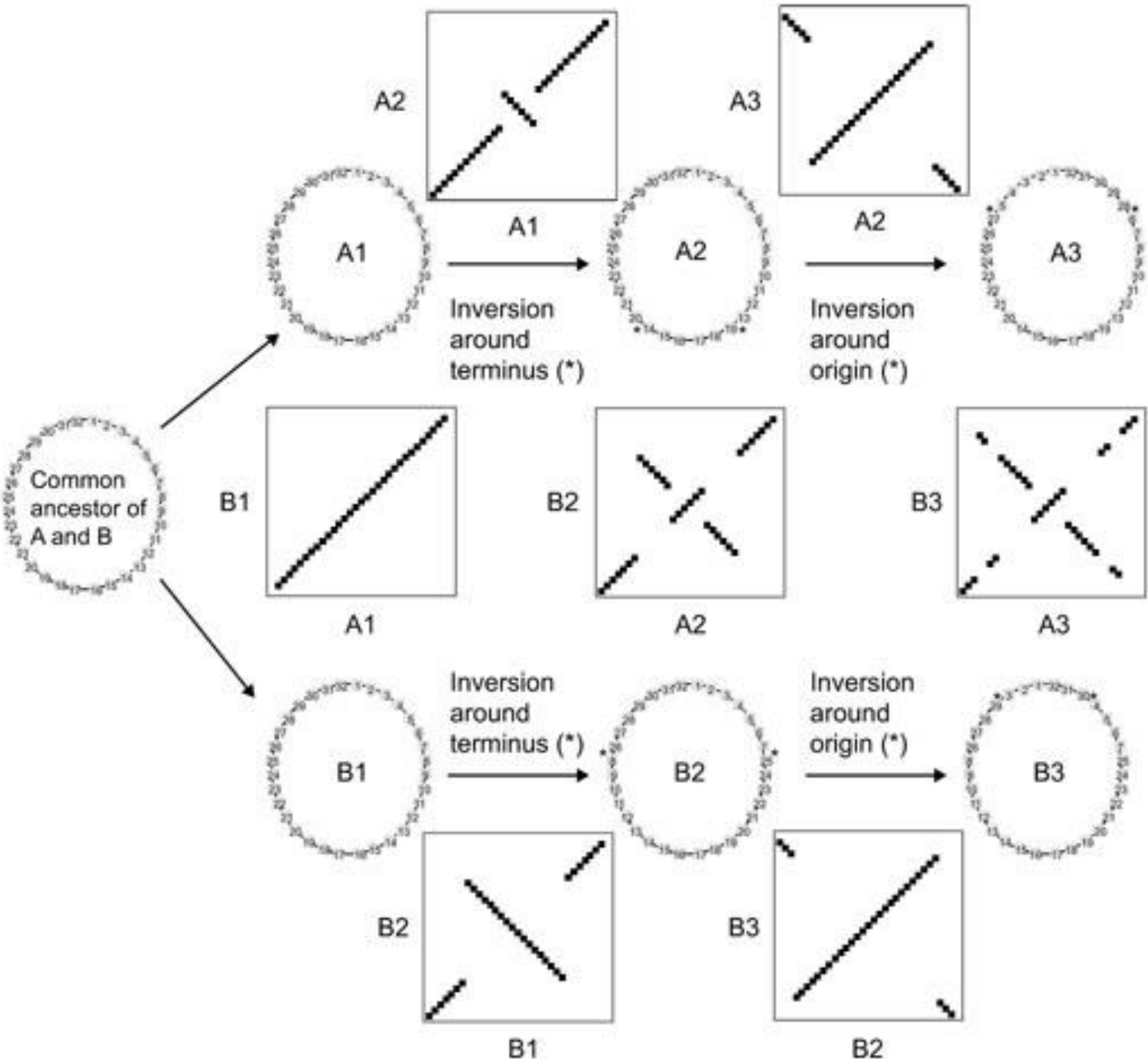
E. coli K12

Promer alignment



Xanthomonas axonopodis pv *citri*

Both are γ proteobacteria!



Eisen JA, Heidelberg JF, White O, Salzberg SL. Evidence for symmetric chromosomal inversions around the replication origin in bacteria. *Genome Biol.* 2000;1(6):RESEARCH0011

Alinhamento múltiplo de sequências longas

- O programa **MAUVE**
- Darling AC, Mau B, Blattner FR, Perna NT. Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res. 2004 Jul;14(7):1394-403.

How MAUVE works

- Seed-and-extend hashing
- Seeds/anchors: Maximal Multiple Unique Matches of minimum length k
- Result: **Local collinear blocks** (LCBs)
- $O(G^2n + Gn \log Gn)$, $G = \#$ genomes, $n =$ average genome length

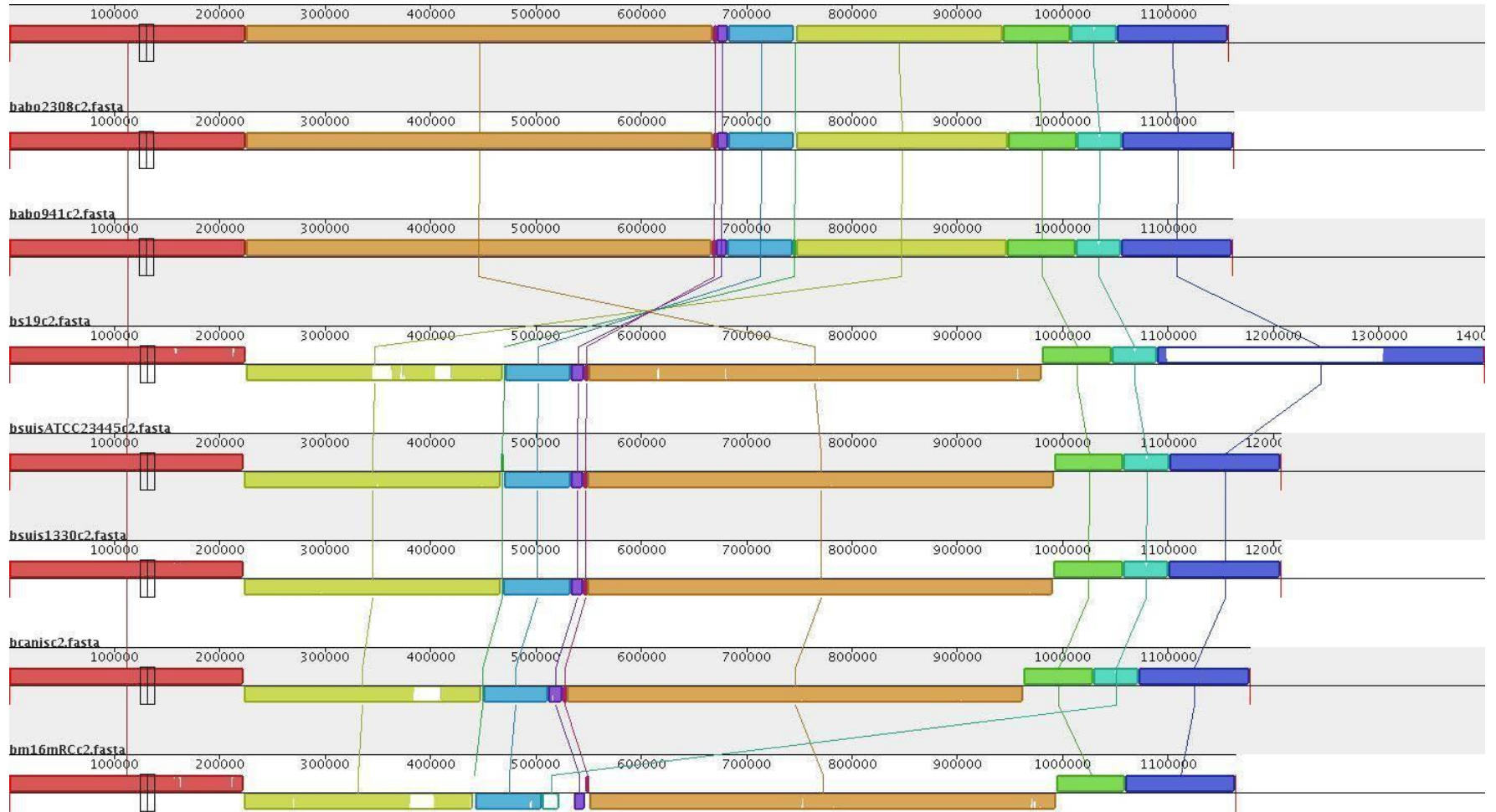
Alignment algorithm

1. Find Multi-MUMs
2. Use the multi-MUMs to calculate a phylogenetic guide tree
3. Find LCBs (subset of multi-MUMs; filter out spurious matches; requires *minimum weight*)
4. Recursive anchoring to identify additional anchors (extension of LCBs)
5. Progressive alignment (CLUSTALW) using guide tree

Brucella: Main chromosome alignment



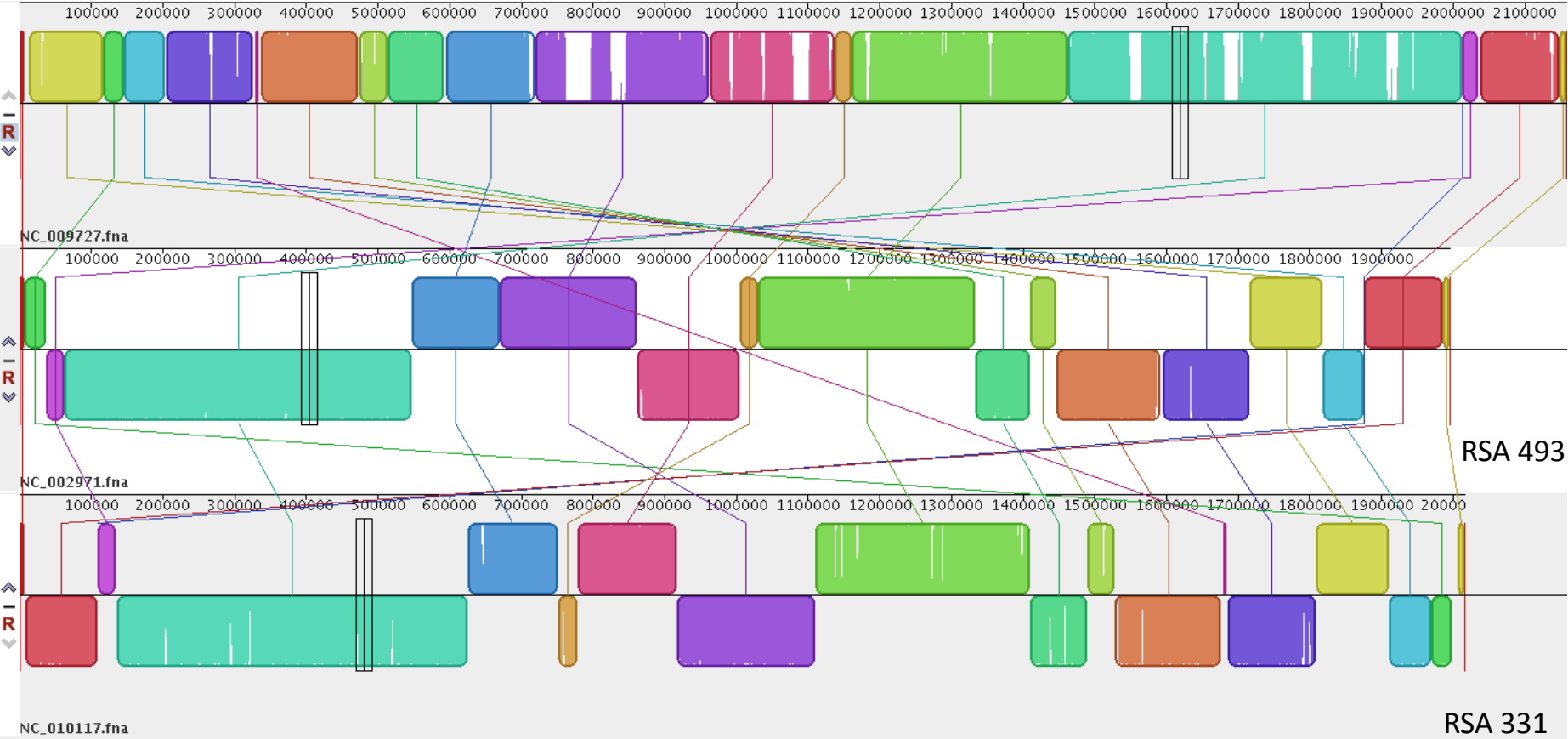
Brucella: Chromosome 2 alignment



bovisc2.fasta

Coxiella: Chromosome alignment

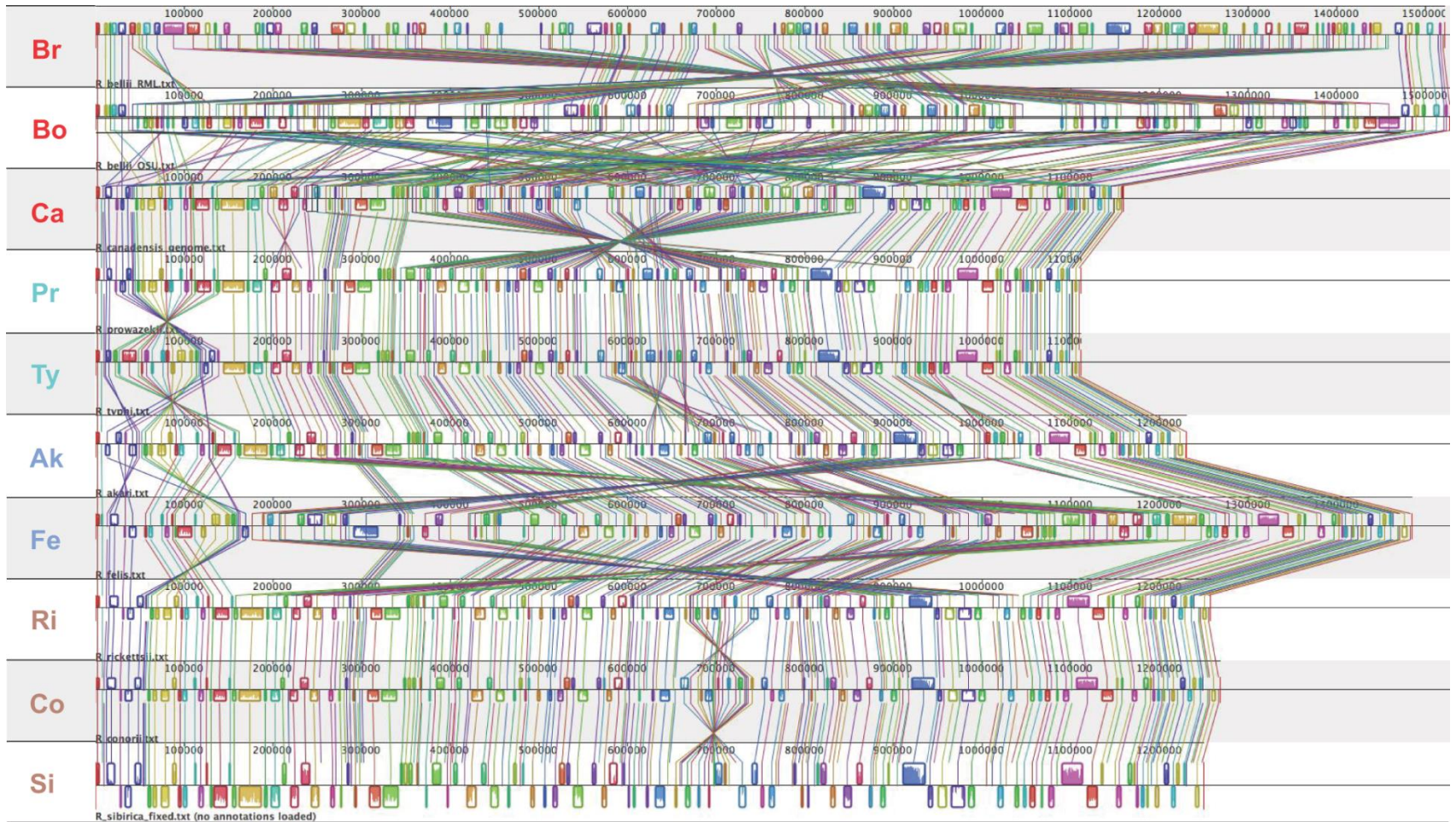
Dugway



RSA 493

RSA 331

Rickettsia



Sumário de comparação de sequências

	Sequências curtas	Sequências longas
2-a-2	Prog. Dinâmica	Mummer
2-a-2 muitas vezes	BLAST, usearch, diamond	Mummer
múltiplo	Muscle, MAFFT	Mauve, MUGSY

Distância genômica

- Ao comparar genomas, muitas vezes é útil poder expressar essa comparação por meio de **um único número**
 - Quando se comparam **pares** de replicons
 - Que podem ser “replicons” concatenados
- Distância pode ser entendida como o **inverso da similaridade**

Distância e similaridade

- São conceitos muito parecidos
- Em particular *distância de edição*
- Como transformar sequência s em sequência t
- Operações
 - **Substituição** do caracter a por b (custo = 1)
 - **Inserção** ou **Remoção** de um caracter (custo = 2)
- O algoritmo de PD já visto resolve esse problema

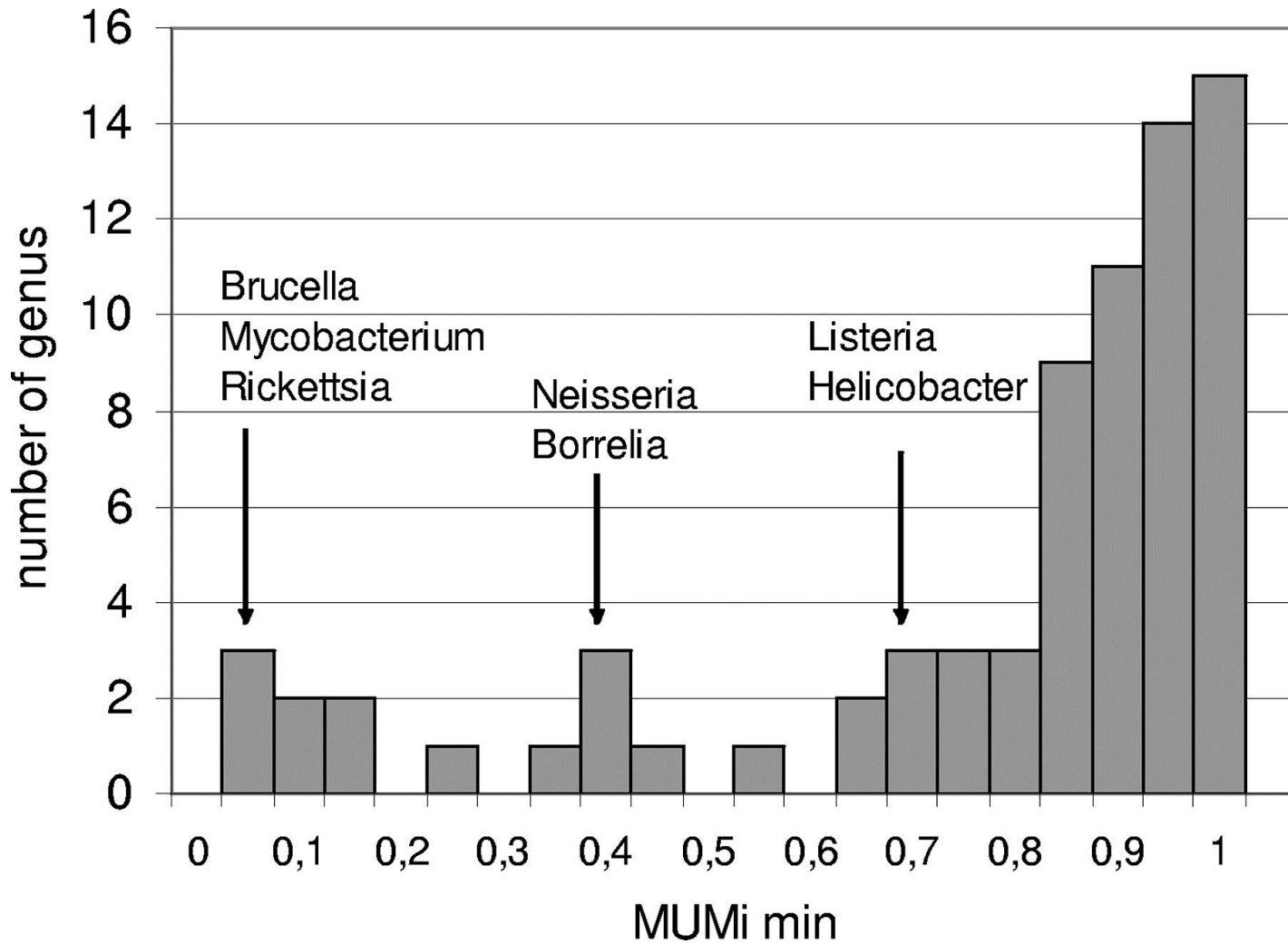
Uma fórmula de distância genômica

- MUMi = MUM index
- Baseado em MUMmer
- Deloger et al. 2009
- $MUMi = 1 - L_{mum}/L_{av}$
- L_{mum} = soma dos comprimentos de todos os MUMs que não tem sobreposição
- L_{av} = comprimento médio dos 2 genomas sendo comparados
- Para obter MUMi, basta rodar MUMmer com um script perl desses autores

How MUMi works

- Identical sequences: **zero**
- Totally different sequences: **1**
- The boundary between **genus** and **species** is around **0.8**

Distribution of all minimal MUMi values per genus.



Marc Deloger et al. J. Bacteriol. 2009;191:91-99

Journal of Bacteriology

Conclusão

- Não dá para comparar distâncias MUMi entre diferentes gêneros

Uma matriz de distâncias genômicas em *Brucella*

Strain	MUMi value ^b					
	83-13	BO2	NF 2653	BO1	<i>B. suis</i>	<i>B. microti</i>
<i>B. neotomae</i> 5K33	0.145	0.168	0.146	0.168	0.017	0.022
<i>Brucella</i> sp. strain 83-13		0.172	0.009	0.175	0.147	0.150
<i>Brucella</i> sp. strain BO2			0.168	0.107	0.169	0.169
<i>Brucella</i> sp. strain NF 2653				0.172	0.147	0.146
<i>B. inopinata</i> BO1					0.169	0.167
<i>B. suis</i> biovar 3 686						0.020

↪^a Only the sequence of *B. microti* was complete, and its two chromosome sequences were concatenated. Concatenation was done by inserting a string with 100 nucleotides between contigs.

↪^b The minimum MUMi value is 0, and the maximum MUMi value is 1.

XacMNT2	0.4076	0.0054	0.0060	0.0166	0.0186	0.0144	0.0169	0.0175	0.0177	0.063
XacNT17	0.4095	0.0005	0.0001	0.0125	0.0221	0.0171	0.0180	0.0135	0.0209	0.065
XacUI6	0.4077	0.0059	0.0065	0.0171	0.0190	0.0149	0.0177	0.0181	0.0182	0.064
XacUI7	0.4076	0.0055	0.0060	0.0173	0.0190	0.0150	0.0177	0.0182	0.0183	0.064
XacX18	0.3933	0.1462	0.1461	0.1379	0.1488	0.1348	0.1368	0.1378	0.1364	0.152
XacX20	0.4086	0.1559	0.1559	0.1476	0.1590	0.1447	0.1469	0.1476	0.1462	0.141
XagCFBP7119	0.4154	0.1611	0.1608	0.1537	0.1650	0.1507	0.1531	0.1535	0.1521	0.155
XalfaCFBP3836	0.1448	0.3896	0.3893	0.3831	0.3930	0.3808	0.3822	0.3826	0.3832	0.392
XalfaF1	0.1413	0.3926	0.3923	0.3876	0.3965	0.3852	0.3866	0.3870	0.3868	0.395
Xalfacm1510	0.1390	0.3895	0.3892	0.3829	0.3935	0.3807	0.3810	0.3824	0.3822	0.391
Xalfacm1637	0.1398	0.3903	0.3901	0.3832	0.3932	0.3814	0.3810	0.3831	0.3823	0.393
XauB11122	0.4191	0.2843	0.2842	0.2767	0.2833	0.2726	0.2749	0.2762	0.2709	0.283
XauB1561	0.4264	0.2948	0.2946	0.2872	0.2984	0.2854	0.2857	0.2874	0.2867	0.294
XauC10535	0.4164	0.2877	0.2876	0.2800	0.2912	0.2777	0.2784	0.2798	0.2798	0.291
XauC1559	0.4412	0.3166	0.3166	0.3101	0.3201	0.3085	0.3079	0.3104	0.3100	0.321
XauC1609	0.4354	0.3036	0.3034	0.2974	0.3070	0.2951	0.2959	0.2972	0.2969	0.308
XauC535	0.4241	0.2953	0.2953	0.2875	0.2989	0.2854	0.2861	0.2872	0.2871	0.298
XauC752	0.4199	0.2997	0.2996	0.2799	0.2913	0.2799	0.2719	0.2799	0.2791	0.294

Largest distance: **0.4506** (*X. perforans* 91-118 and *X. fuscans aurantifolii* C 1559)

For comparison: distance between *Stenotrophomonas maltophilia* and Xac 306: **0.912**

ANI

- **Average Nucleotide Identity** [Goris et al. 2007]
- Baseado em BLAST
- one-way ANI (best hits)
- two-way ANI (reciprocal best hits)
- Typically, 2 species will have **ANI \geq 95%**
- **< 75%** : not to be trusted
- <http://enve-omics.ce.gatech.edu/ani/>

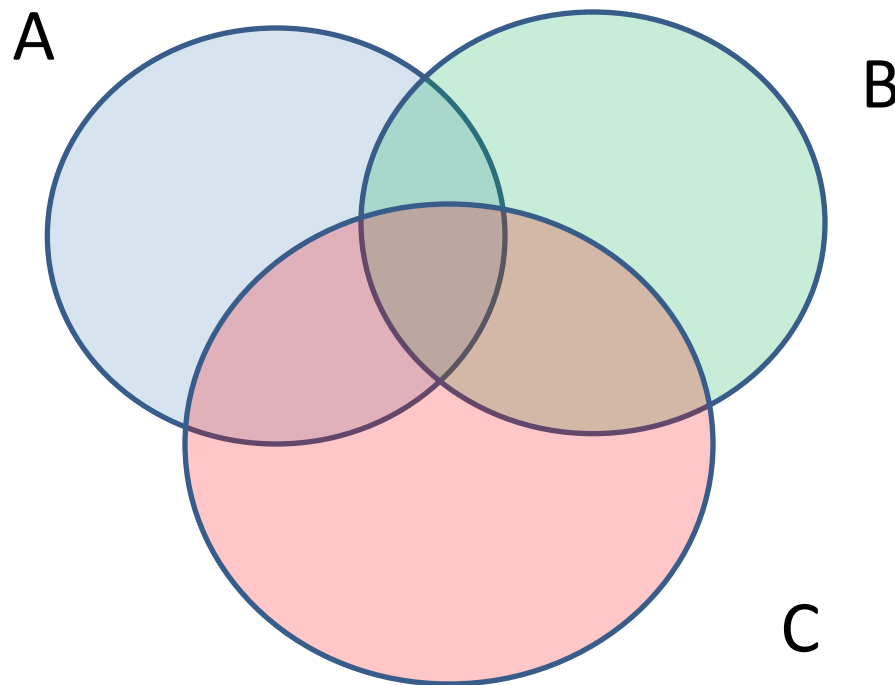
GBDP

- Genome Blast Distance Phylogeny
- rigorous *in silico* replacement for DNA-DNA hybridization wet-lab experiments
- <http://ggdc.dsmz.de>
- Meier-Kolthoff, J.P., et al., *Genome sequence-based species delimitation with confidence intervals and improved distance functions*. BMC Bioinformatics, 2013. 14: p. 60

Comparação de conjuntos de genes

- Given a set of genomes, represented by their 'proteomes' or sets of protein sequences
- Given homologous relationships (as given for example by orthoMCL)
 - Which genes are shared by genomes X and Y ?
 - Which genes are unique to genome Z ?
 - Venn or extended Venn diagrams

3-way genome comparison



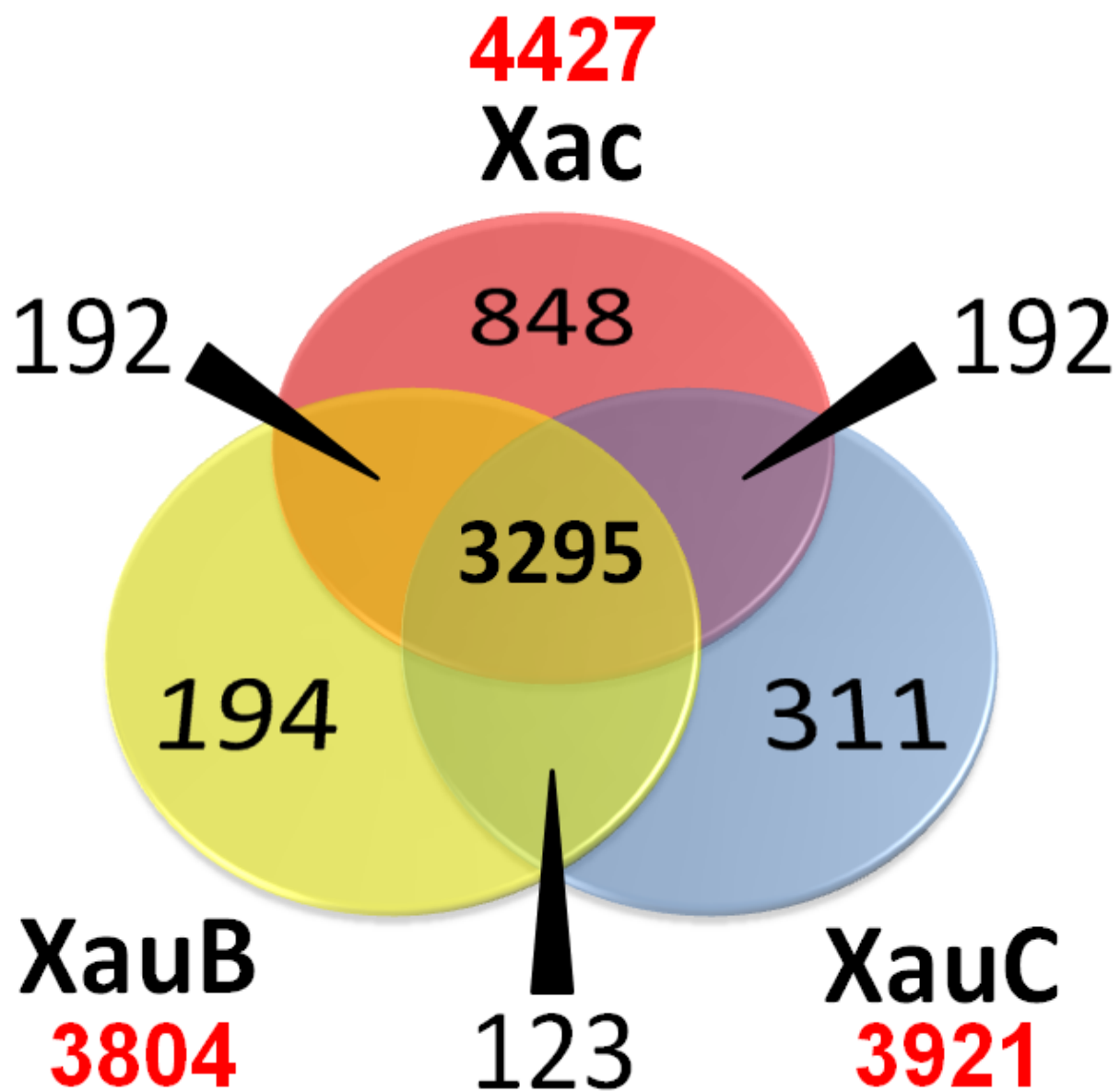
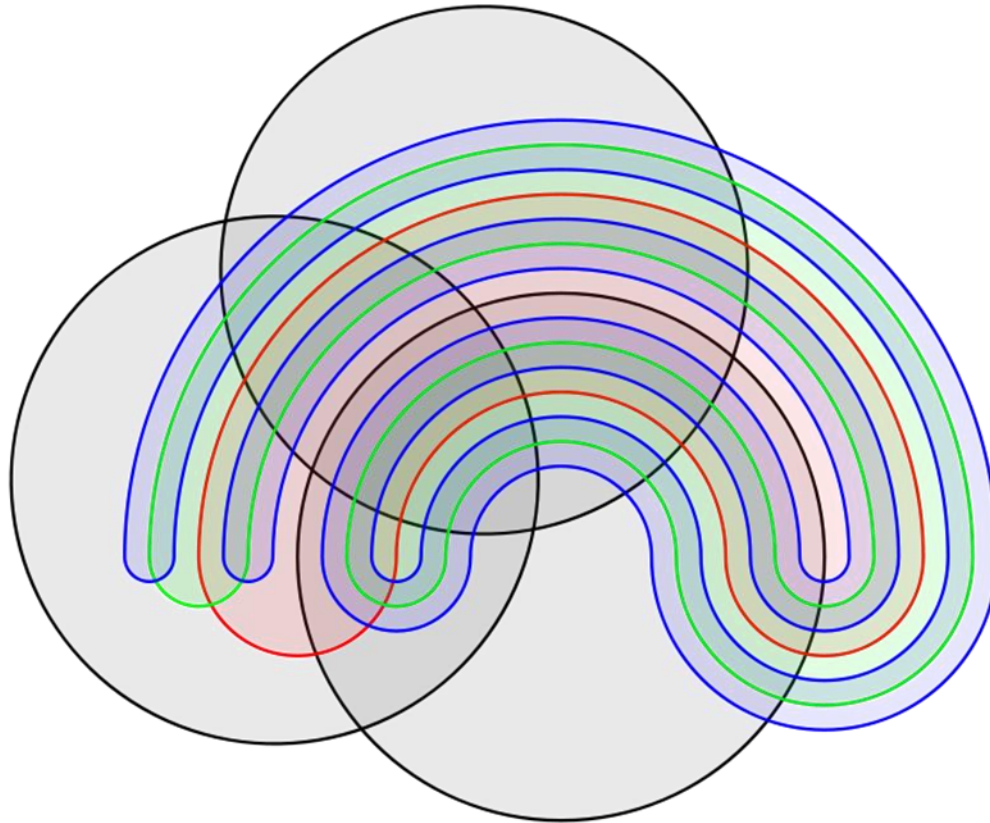


Diagrama de Venn para $n = 6$



Número de comparações é quadrático em n
Número de regiões num diagrama de Venn = 2^n

Source: wikipedia

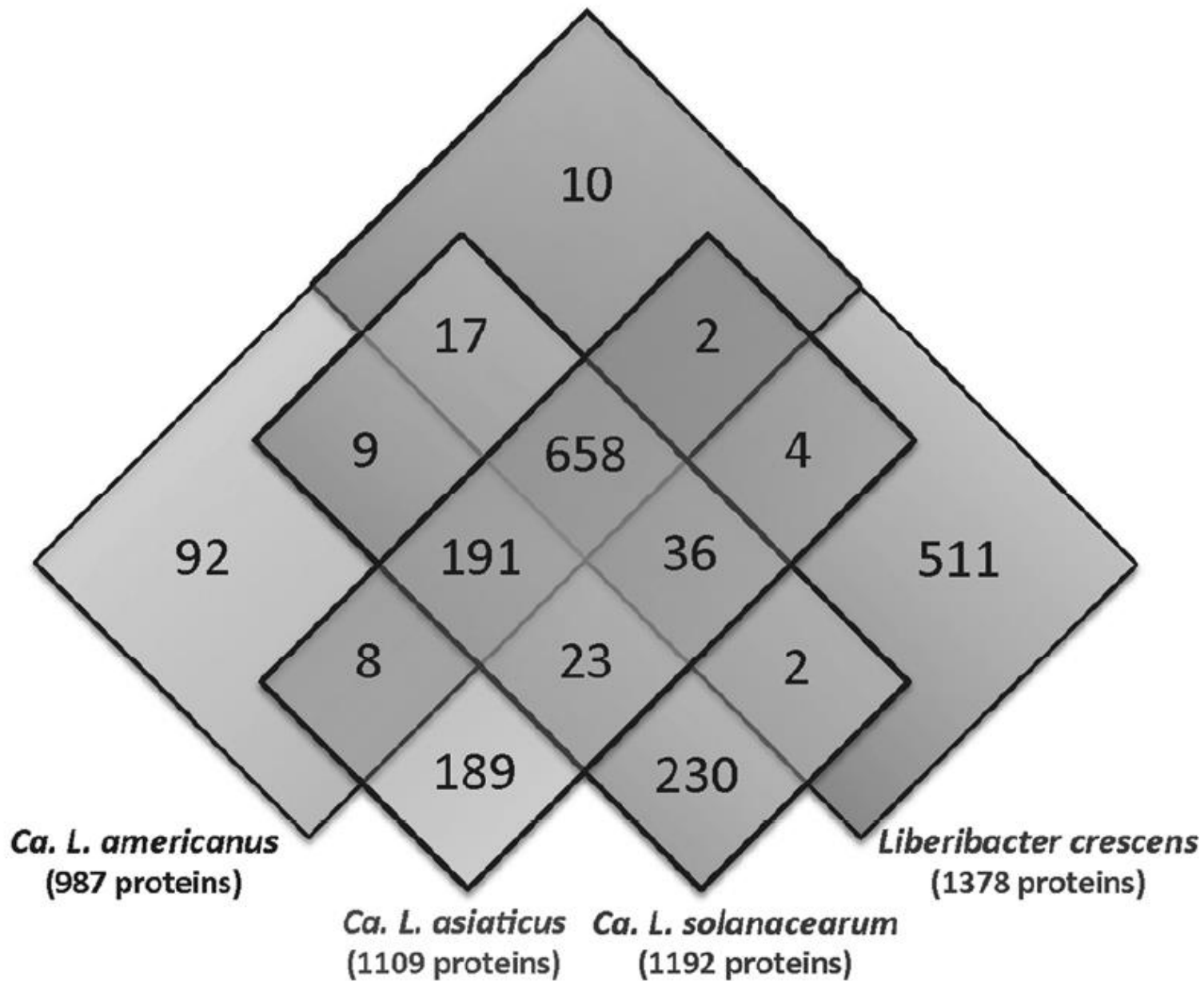
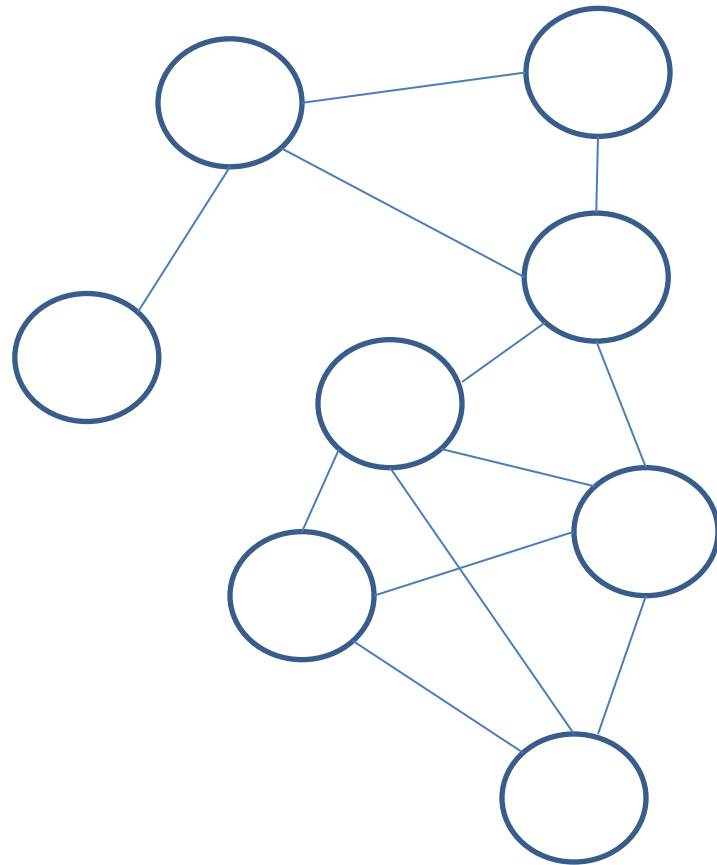
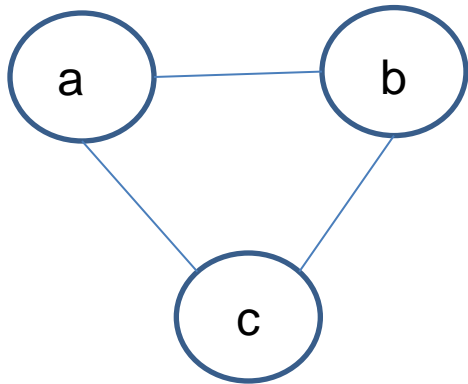


Fig. 2. Venn diagram showing protein-coding gene sharing among the four *Liberibacter* genomes. Numbers below each species are the total number of predicted protein-coding genes for that genome.

Cômputo de famílias de proteínas

1. Verificar as similaridades entre as sequências
 - a) Usando (por exemplo) BLAST + critérios
2. Representar as similaridades num **grafo**
3. Aplicar um **algoritmo de clusterização** sobre o grafo

Clusterização é necessária porque o grafo pode ser complexo

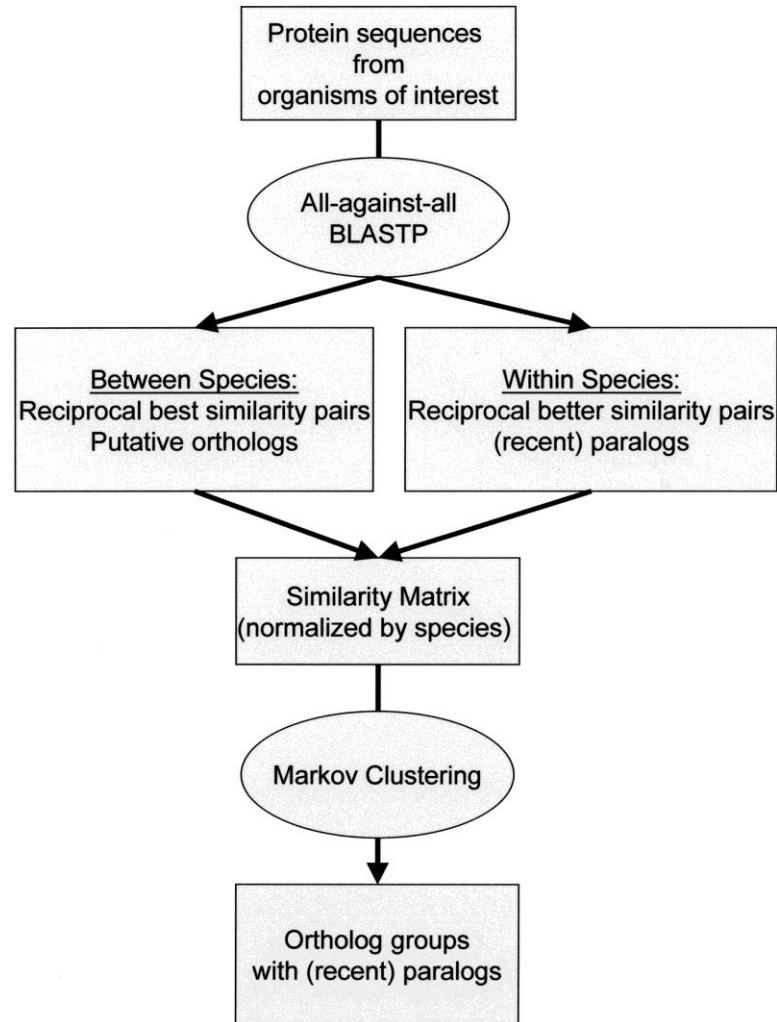


Resultado da clusterização

- Matriz com genes nas colunas (g_j), genomas nas linhas (O_i)
- Cada coluna representa uma **família de genes homólogos**

	g1	g2	g3	g4	...
O1	✓		✓	✓	
O2		✓	✓		
O3	✓		✓		
O4			✓	✓	✓
O5	✓		✓		✓
O6			✓		
...			✓		

orthoMCL pipeline



Li Li et al. *Genome Res.* 2003; 13: 2178-2189

OrthoMCL DB

Ortholog Groups of Protein Sequences

■ [Home](#)

■ [About OrthoMCL](#) ▾

■ [Data](#) ▾

■ [Search](#) ▾

■ [Tools](#) ▾

NEWS

Mar 31, 2011

OrthoMCL-DB Version 5 is released. We have included 150 genomes in this release.

May 31, 2010

OrthoMCL-DB Version 4 is released. There are 138 genomes included in version 4.

Oct 9, 2009

OrthoMCL-DB Version 3 is released. The new dataset includes 128 genomes. New web site features include: (1) a tool to assign your proteins to OrthoMCL groups (see the new tools menu); (2) a mapping from Version 3 groups to Version 2 and 1 is available for searching, allowing you to track changes across versions; (3) the phyletic pattern in the groups result page is configurable, so you can tailor it to the clades you are interested in.

Sep 22, 2009

OrthoMCL-DB pipeline re-engineered. We have completely overhauled how we produce the database, encoding the entire process in a pipeline system. This should dramatically improve our ability to deliver new versions of the database.

Feb 28, 2008

OrthoMCL-DB Version 2 is released. The

Welcome to OrthoMCL DB

Ortholog Groups of Protein Sequences from Multiple Genomes!

Current Release:

Version: **5**

Number of Genomes: **150**

Number of Protein Sequences: **1398546**

Number of Ortholog Groups: **124740**

Search for Groups

- [by IDs, Keyword, or PFam domain](#)
- [by Phyletic Pattern](#)
- [by Phyletic Pattern - Advanced](#)
- [by Group Properties](#)
- [Query History - Groups](#)

Search for Sequences

- [by IDs, Keyword, Taxonomy or PFam domain](#)
- [by BLAST Search](#)
- [Query History - Sequences](#)

Tools

- [Assign your proteins to OrthoMCL groups](#)

OrthoMCL Software

Database Download



Xanthomonas fuscans str. *aurantifolii* Genome Project



Choose genomes to see OrthoMCL families containing genes from them

Include	Exclude	I don't care	
<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<i>Xanthomonas fuscans</i> subsp. <i>aurantifolii</i> str. 11122 (B)
<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<i>Xanthomonas fuscans</i> subsp. <i>aurantifolii</i> str. 10535 (C)
<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<i>Xanthomonas campestris</i> pv. <i>campestris</i> str. 8004
<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<i>Xanthomonas campestris</i> pv. <i>campestris</i> str. ATCC 3391
<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<i>Xanthomonas campestris</i> pv. <i>campestris</i> B100
<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<i>Xanthomonas campestris</i> pv. <i>vesicatoria</i> str. 85-10
<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<i>Xanthomonas axonopodis</i> pv. <i>citri</i> str. 306
<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<i>Xanthomonas oryzae</i> pv. <i>oryzae</i> KACC10331
<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<i>Xanthomonas oryzae</i> pv. <i>oryzae</i> MAFF 311018
<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<i>Xanthomonas oryzae</i> pv. <i>oryzae</i> PXO99A
<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<i>Xanthomonas albilineans</i> GPE PC73
<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<i>Xylella fastidiosa</i> 9a5c
<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<i>Xylella fastidiosa</i> M12
<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<i>Xylella fastidiosa</i> M23 plasmid pXFAS01
<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<i>Xylella fastidiosa</i> Temecula1

check all check all check all(see all families)

Submit Query

Reset

```
=====
>Family 5286 -- Gblocks 331/384 (86%)
21243924 nr alkanesulfonate transporter substrate-binding subunit [Xanthomonas axonopodis pv. citri str. 306]
9810940 nr - nitrate transport protein [Xanthomonas fuscans subsp. aurantifolii str. 11122 (B)]
9929350 nr - nitrate transport protein [Xanthomonas fuscans subsp. aurantifolii str. 10535 (C)]

count_ = 3
=====
>Family 5287 -- Gblocks 346/346 (100%)
21243925 nr oxidoreductase [Xanthomonas axonopodis pv. citri str. 306]
9810930 nr - oxidoreductase [Xanthomonas fuscans subsp. aurantifolii str. 11122 (B)]
9929360 nr - oxidoreductase [Xanthomonas fuscans subsp. aurantifolii str. 10535 (C)]

count_ = 3
=====
>Family 5288 -- Gblocks 442/442 (100%)
21243926 nr nitrilotriacetate monooxygenase component A [Xanthomonas axonopodis pv. citri str. 306]
9810920 nr - nitrilotriacetate monooxygenase component A [Xanthomonas fuscans subsp. aurantifolii str. 11122 (B)]
9929370 nr - nitrilotriacetate monooxygenase component A [Xanthomonas fuscans subsp. aurantifolii str. 10535 (C)]

count_ = 3
=====
>Family 5289 -- Gblocks 356/356 (100%)
21243950 nr avirulence protein [Xanthomonas axonopodis pv. citri str. 306]
9814680 nr - type III secretion system hopX1-like protein [Xanthomonas fuscans subsp. aurantifolii str. 11122 (B)]
9900040 nr - type III secretion system hopX1-like protein [Xanthomonas fuscans subsp. aurantifolii str. 10535 (C)]

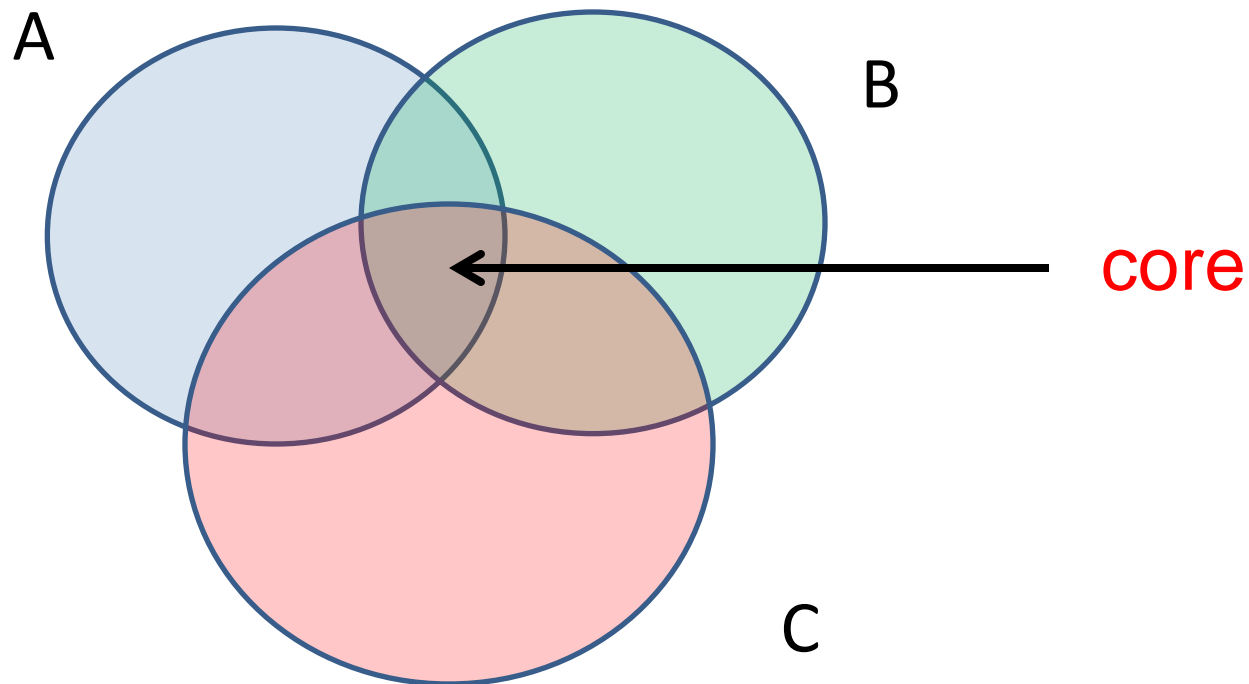
count_ = 3
=====
>Family 5290 -- Gblocks 291/297 (97%)
21243956 nr hypothetical protein XAC3230 [Xanthomonas axonopodis pv. citri str. 306]
9826830 nr - conserved hypothetical protein [Xanthomonas fuscans subsp. aurantifolii str. 11122 (B)]
9923780 nr - conserved hypothetical protein [Xanthomonas fuscans subsp. aurantifolii str. 10535 (C)]

count_ = 3
=====
>Family 5291 -- Gblocks 203/373 (54%)
21243959 nr transposase [Xanthomonas axonopodis pv. citri str. 306]
9803570 nr - transposase [Xanthomonas fuscans subsp. aurantifolii str. 11122 (B)]
9900450 nr - transposase [Xanthomonas fuscans subsp. aurantifolii str. 10535 (C)]

count_ = 3
=====
>Family 5292 -- Gblocks 163/171 (95%)
21243960 nr hypothetical protein XAC3234 [Xanthomonas axonopodis pv. citri str. 306]
9820110 nr - conserved hypothetical protein [Xanthomonas fuscans subsp. aurantifolii str. 11122 (B)]
```



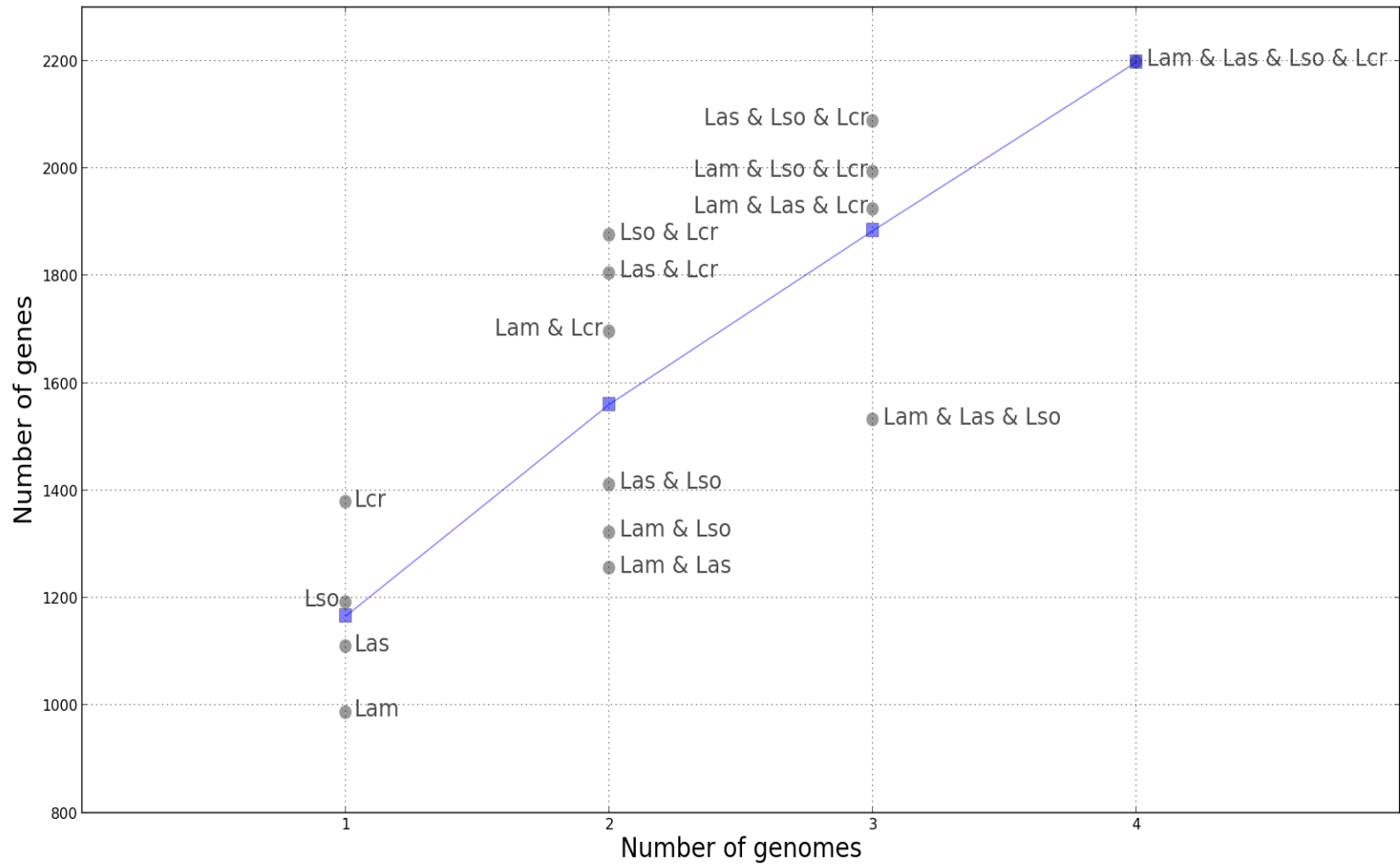
Pan genoma; genoma core e genoma acessório



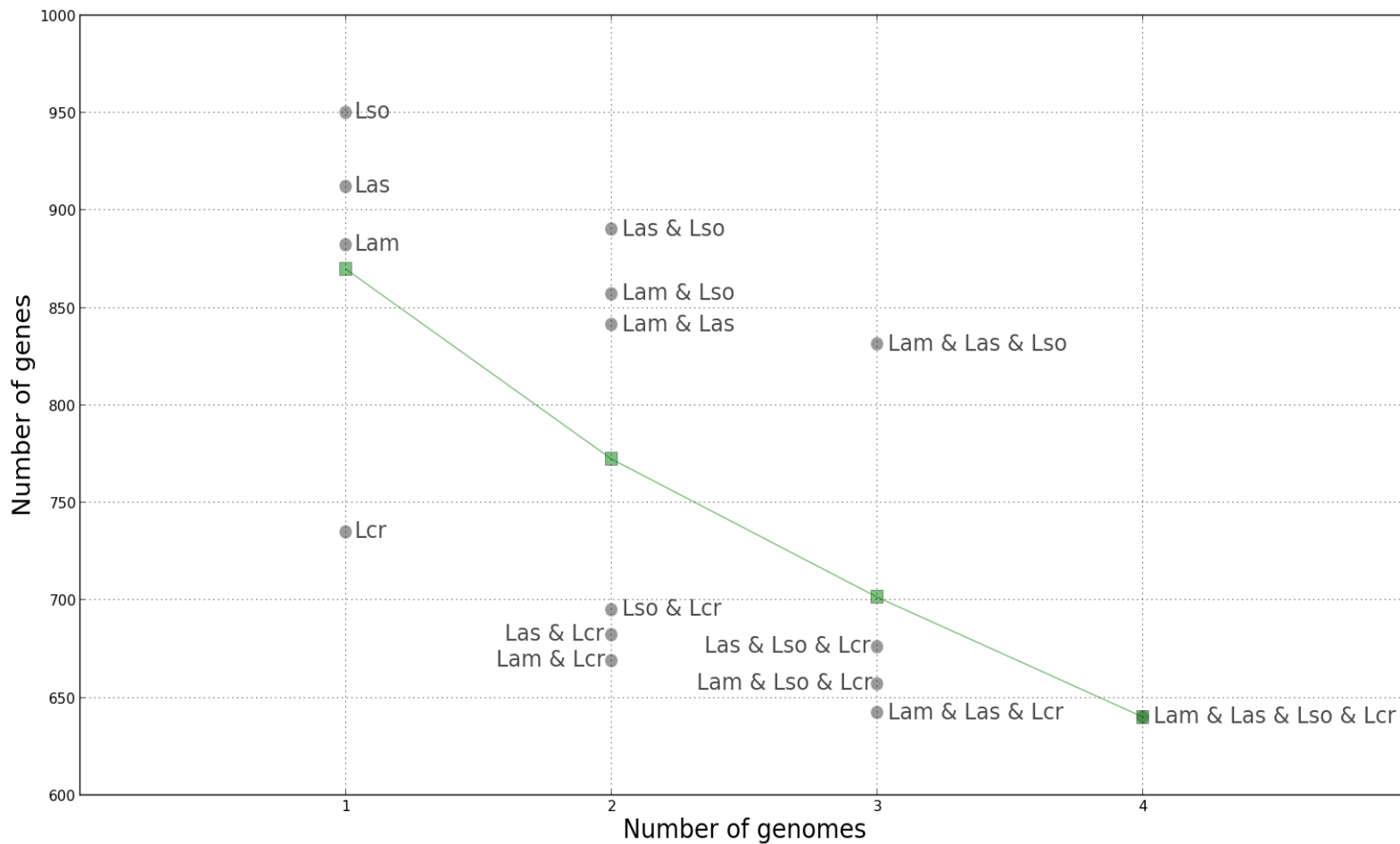
pan: $A \cup B \cup C$

acessório: pan - core

Curva de pan-genoma (n = 4)



Curva de core genoma (n = 4)



O número de genes para x=1 não é o mesmo do gráfico anterior pois os singletons são descartados

Genomas fechados e abertos

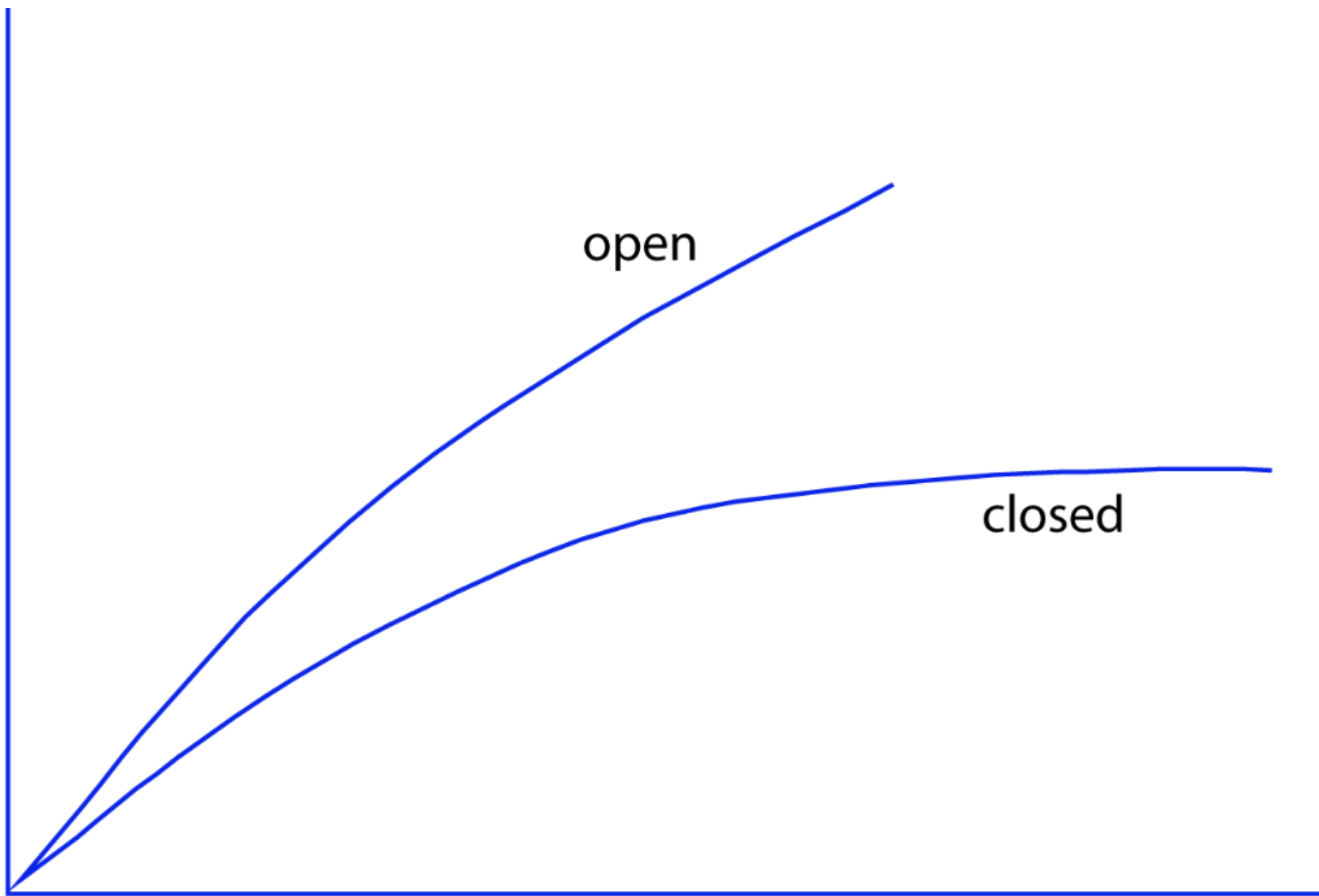
- Fechado
 - O número de genes/famílias do pan-genoma atinge um platô
- Aberto
 - O número de genes/famílias não atinge um platô; cresce sempre que se acrescentam novos genomas

genes

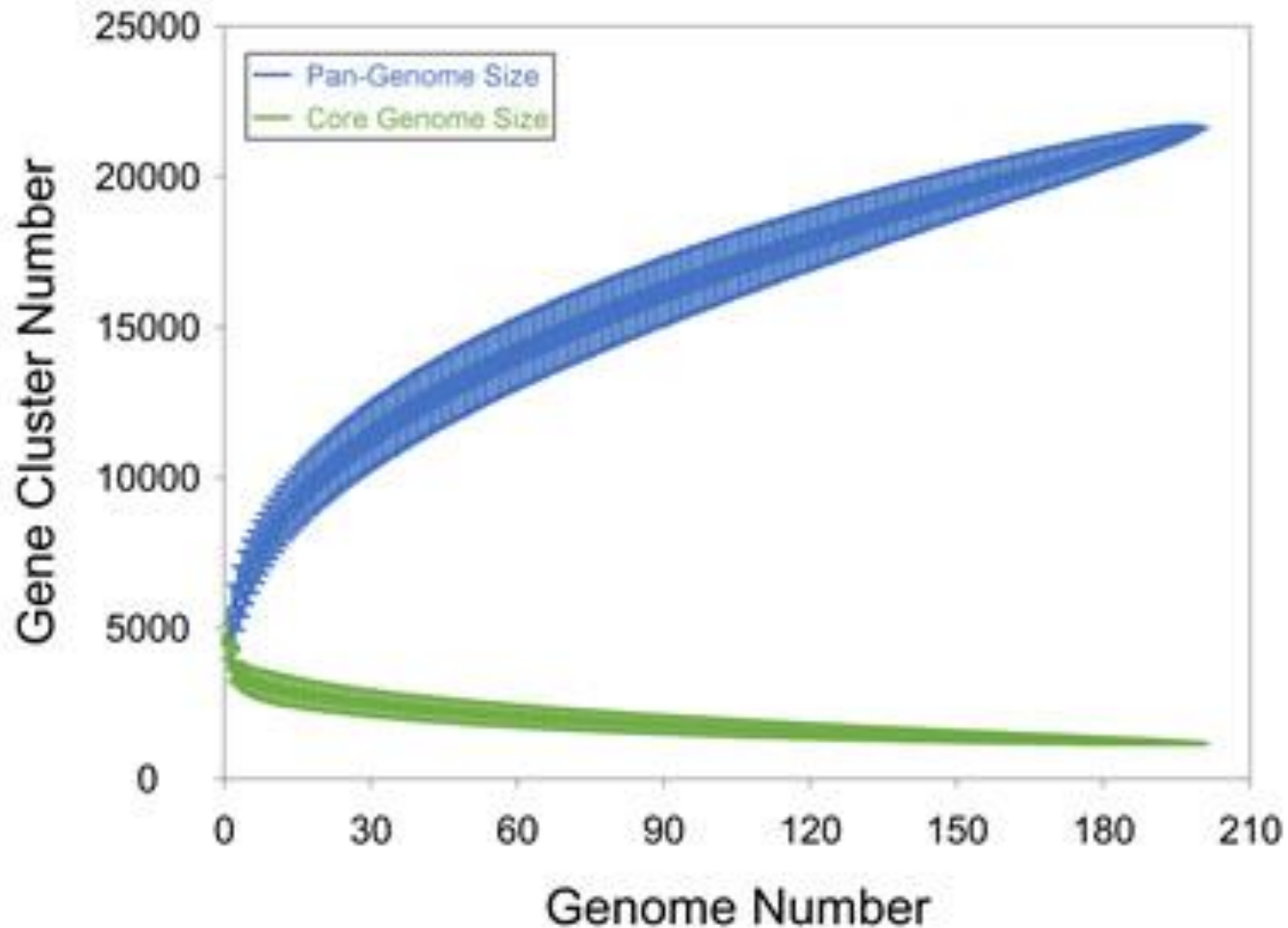
open

closed

genomes



O genoma de *E. coli* é aberto



Bacillus anthracis

- Genoma fechado
- Cerca de 3.000 genes

Ferramentas para análise pan/core

- **Get_Homologues**

- Contreras-Moreira, B. and P. Vinuesa, *GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis*. *Appl Environ Microbiol*, 2013. 79(24): p. 7696-701

- **Roary**

- Page, A.J., et al., *Roary: rapid large-scale prokaryote pan genome analysis*. *Bioinformatics*, 2015. 31(22): p. 3691-3

Conjuntos + contexto

- Como genes compartilhados aparecem em seus respectivos genomas?
- **Filogenômica**
- Busca de **sintenia** = preservação de ordem
- Basta fazer um alinhamento
 - Os “caracteres” a serem alinhados são os genes

Alinhamento de ortólogos obtido no IMG/JGI

Xanthomonas axonopodis pv. *citri* str. 306: NC_003919



Xanthomonas fuscans subsp. *aurantifolii* str. ICPB 11122 xaub_ctg679_0: NZ_ACPX01000220



Xanthomonas axonopodis pv. *malvacearum* GSPB1386 : AHIB01000037



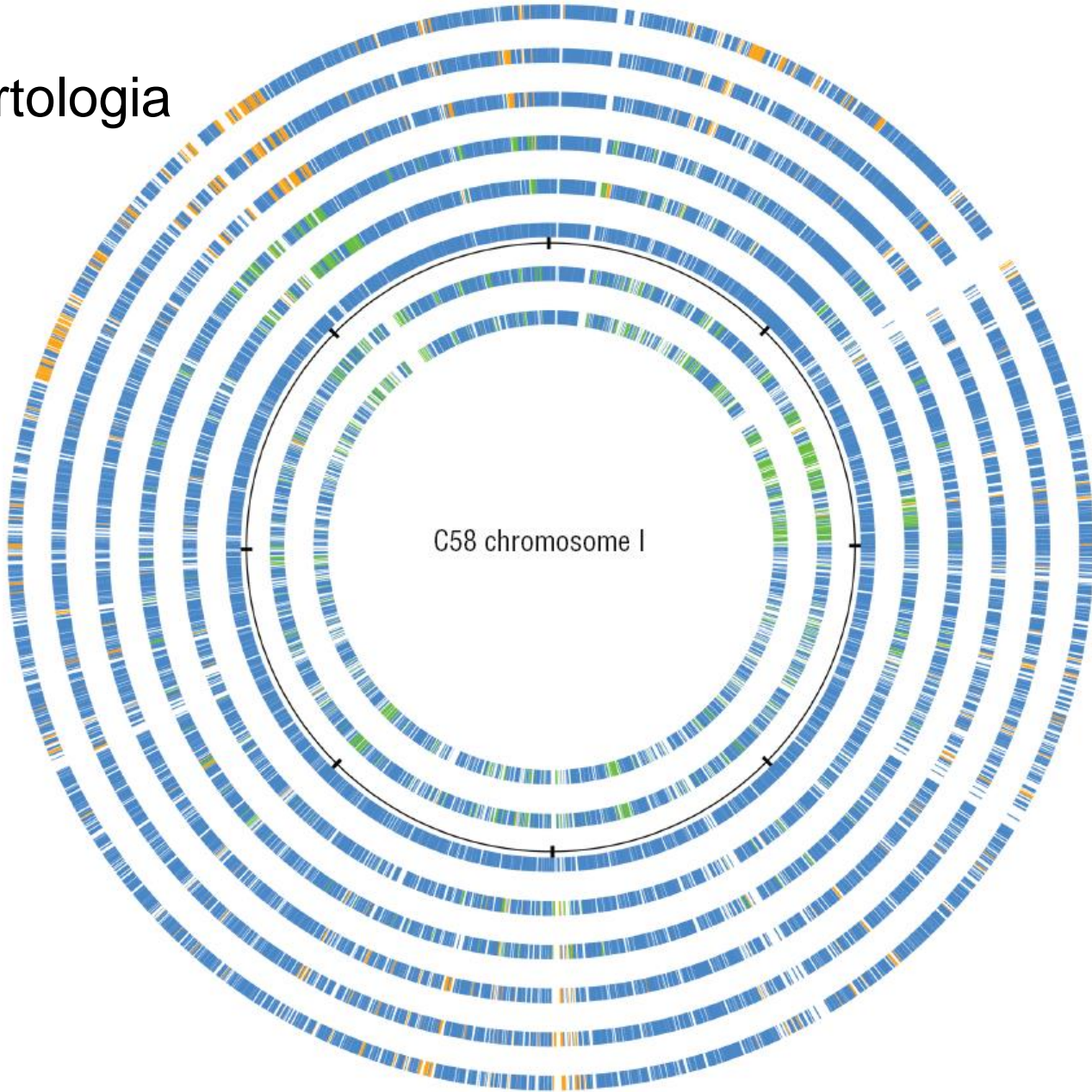
Xanthomonas axonopodis pv. *malvacearum* GSPB2388 : AHIC01000016

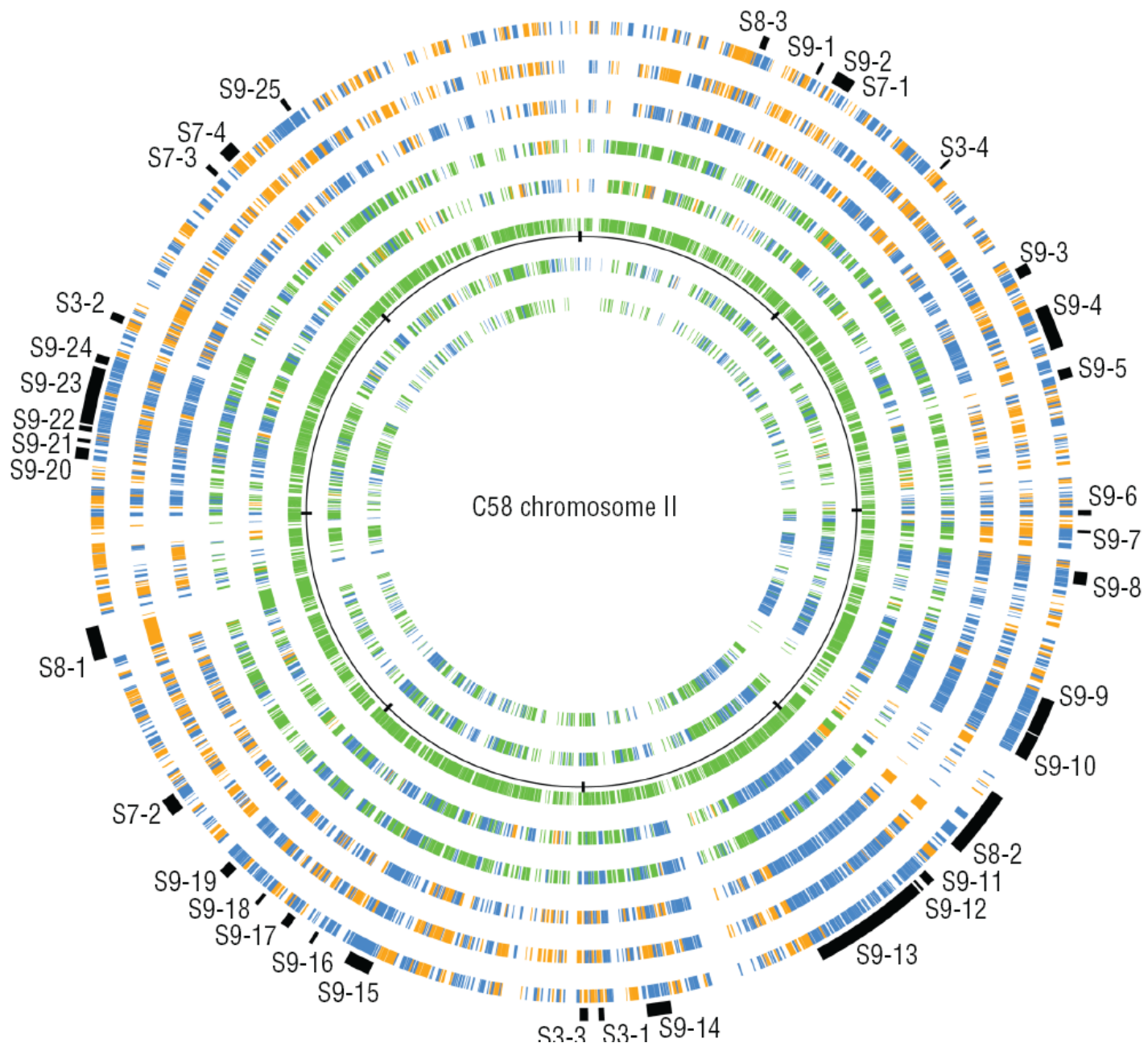


Xanthomonas axonopodis pv. *punicae* LMG 859 : CAGJ01000020



Roda da ortologia





Sumário - 1

- Comparação de sequências “curtas” 2-a-2
 - Alinhamento
 - Sistemas de pontuação
 - Matrizes de substituição
 - Programação dinâmica
 - Significância estatística de alinhamentos
 - BLAST

Sumário - 2

- Comparação de várias sequências ao mesmo tempo
 - Alinhamento múltiplo
 - Programas
- Comparação de sequências “longas”
 - 2-a-2
 - Alinhamento múltiplo

Sumário - 3

- Distâncias genômicas
- Comparação entre conjuntos de genes
- Pan/Core
- Sintenia

Protein family resources



Clusters of orthologous groups (COG, KOG, eggNOG)

KEGG orthologs

Pfam 26.0 (November 2011, 13672 families)

The Pfam database is a large collection of protein families, each represented by **multiple sequence alignments** and **hidden Markov models (HMMs)**. [More...](#)



QUICK LINKS

[SEQUENCE SEARCH](#)

[VIEW A PFAM FAMILY](#)

[VIEW A CLAN](#)

[VIEW A SEQUENCE](#)

[VIEW A STRUCTURE](#)

[KEYWORD SEARCH](#)

[JUMP TO](#)

YOU CAN FIND DATA IN PFAM IN VARIOUS WAYS...

Analyze your protein sequence for Pfam matches

View Pfam family annotation and alignments

See groups of related families

Look at the domain organisation of a protein sequence

Find the domains on a PDB structure

Query Pfam by keywords

Enter any type of accession or ID to jump to the page for a Pfam family or clan, UniProt sequence, PDB structure, etc.

Or view the [help](#) pages for more information

Recent [blog](#) posts

Hide this

[Does my family of interest have a determined 3D protein structure?](#) (posted 9 May 2012)

Two related questions that we are often asked via the Pfam helpdesk is 'Which families have a known three-dimensional structure?' and 'Why is a particular a PDB structure not found in Pfam'. You may think that there are obvious answers to these questions – but as with many things in life the answer is not [...]

[TreeFam is back with a new release!](#) (posted 27 March 2012)

Query by accession

Protein: *PDK1_HUMAN* (Q15118)

1 architecture

1 sequence

0 interactions

1 species

3 structures

Summary

Features

Sequence

Interactions

Structures

TreeFam

Jump to...

enter ID/acc

Summary

PDK1_HUMAN

This is the summary of UniProt entry [PDK1_HUMAN](#) ([Q15118](#)).

Description:	[Pyruvate dehydrogenase [lipoamide]] kinase isozyme 1, mitochondrial EC=2.7.11.2
Source organism:	Homo sapiens (Human) (NCBI taxonomy ID 9606) View Pfam proteome data.
Length:	436 amino acids

Please note: when we start each new Pfam data release, we take a copy of the UniProt sequence database. This snapshot of UniProt forms the basis of the overview that you see here. It is important to note that, although some UniProt entries may be removed *after* a Pfam release, these entries will not be removed from Pfam until the *next* Pfam data release.

Pfam domains

This image shows the arrangement of the Pfam domains that we found on this sequence. Clicking on a domain will take you to the page describing that Pfam entry. The table below gives the domain boundaries for each of the domains.



Source	Domain	Start	End
sig_p	n/a	1	21
low_complexity	n/a	2	41
Pfam A	BCDHK_Adom3	55	221
Pfam A	HATPase_c	268	392

KEYWORD SEARCHAll **SEQUENCE SEARCH**Enter a protein sequence:

Sequence query limits: Protein - 50kb

The PANTHER (**P**rotein **A**nalysis **T**Hrough **E**volutionary **R**elationships) Classification System is a unique resource that **classifies genes by their functions**, using published scientific experimental evidence and evolutionary relationships to predict function even in the absence of direct experimental evidence. Proteins are **classified by expert biologists** according to:

- ✦ [Gene families and subfamilies](#), including annotated phylogenetic trees
- ✦ [Gene Ontology classes](#): molecular function, biological process, cellular component
- ✦ [PANTHER Protein Classes](#)
- ✦ [Pathways, including diagrams](#)

PANTHER is part of the [Gene Ontology Reference Genome Project](#).

PANTHER is supported by a research grant from the National Institute of General Medical Sciences [grant [GM081084](#)] and maintained by the [Thomas lab at the University of Southern California](#).

Quick links[Whole genome function views](#)[Gene expression tools](#)[cSNP tools](#)[Upload multiple gene IDs](#)[Community Curation](#)[My Workspace](#)[HMM scoring](#)[Downloads](#)[Genome statistics](#)[Site map](#)**Newsletter subscription**

Enter your Email:

What can I do on the PANTHER site?

[Guide to getting started](#)**News**

(March 16, 2012)

PANTHER 7.2 is released.

[Click](#) for additional info.**Publications**[How to cite PANTHER](#)

"[PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium.](#)" [Mi, et al.](#)

"[Applications for protein sequence-function evolution data: mRNA/protein expression analysis and coding SNP scoring tools.](#)" [Thomas, et al.](#)

"[PANTHER: a library of protein families and subfamilies indexed by function.](#)" [Thomas, et al.](#)

Search

PANTHER families ▾

Quick links

[Whole genome function views](#)

[Gene expression tools](#)

[cSNP tools](#)

[Upload multiple gene IDs](#)

[Community Curation](#)

[My Workspace](#)

[HMM scoring](#)

[Downloads](#)

[Genome statistics](#)

[Site map](#)

Newsletter subscription

Enter your Email:

PANTHER HMM SEQUENCE SCORING RESULTS ?

The top scoring HMM is reported, along with the E-value (the number of expected false-positive hits expected). If the E-value is less than 1e-3, no hits are reported.

PANTHER Hit: [PROLINE SYNTHETASE CO-TRANSCRIBED BACTERIAL HOMOLOG PROTEIN](#) (PTHR10146)

HMM E-value score: 2.2e-113 ●●● ?

Sequence	Domain	seq-f	seq-t	hmm-f	hmm-t	score	E-value
sequence	1/1	1	234	11	253	387.5	2.2e-113

Alignments of top-scoring domains:

sequence: domain 1 of 1, from 1 to 234: score 387.5, E = 2.2e-113

```

          *->lgvaanLakVlerikaaaakagRdppavrLvAVSKTkPaelileayd
          ++va+nL++V+++++a+ak+ R  + +LvAVSKTkP+e+++eay+
sequence  1  MAVAKNLLAVRAKVAEAVAKSARQ--QQCTLVAVSKTKPVEDLQEAYE  46

          aGqRhFGENYvQEIlleKap1LpdlcpdikWHFIGhLQsNKvkk11.gvpn
          a qRhFGENY+QE1++Kap1Lp  d+kWH+IGh+QsNK+k+l+++vpn
sequence  47 ADQRHFGENYIQELVQKAPLLPK---DVKWHYIGHVQSNKAKPLVrDVPN  93

          ldmvhsvDslklAdklnkaaaklkg1gkplkvlvQVNtsGEesKsGvppe
          l++v++vDs+k+A++lnka  ++  ++++l+v+vQVNts Ee+KsG++ +
sequence  94 LFFVETVDSIKIANALNKASGEF--RSEKLNVMVQVNTSEEEQKSGIDAD  141

          Elpellkhv1kkcpnLellGLMTIGpfdgdlekgnpdFalLaklrkevc
          +el++h+++ c++L+l GLMTIG++++  ++  F +L+++rk+v+
sequence  142 GSVELAQHIVSSCEHLNLTGLMTIGRYGDTTSE----CFDRLVACRKRVA  187

          kklglnpk11ELSMGMSgDfelAIeaGsT1VRvGsaIFGeRdypkpk<-*
          +++g  +  1 LSMGMSgDfelAI  GsT+VRvGs+IFG+R+y +k+
sequence  188 EAIGKAETDLDSMGMSGDFELAI SCGSTHVRVVGSTIFGARNYANKE  234

```

Search

Quick links

[Whole genome function views](#)
[Gene expression tools](#)
[cSNP tools](#)
[Upload multiple gene IDs](#)
[Community Curation](#)
[My Workspace](#)
[HMM scoring](#)
[Downloads](#)
[Genome statistics](#)
[Site map](#)


Newsletter subscription

Enter your Email:

PANTHER FAMILY INFORMATION ?

Family: PROLINE SYNTHETASE CO-TRANSCRIBED BACTERIAL HOMOLOG PROTEIN (PTHR10146)

Subfamilies: [1](#)

PANTHER Links: 

[Tree](#) [MSA](#)

GO Molecular Function: [catalytic activity](#)

GO Biological Process: [metabolic process](#)

↳ [primary metabolic process](#)

↳ [cellular amino acid and derivative metabolic process](#)

↳ [cellular amino acid metabolic process](#)

GO Cellular Component:

PANTHER protein class:

Pathway Categories: No pathway information available

Genes: [45](#)

HMM Length: 273

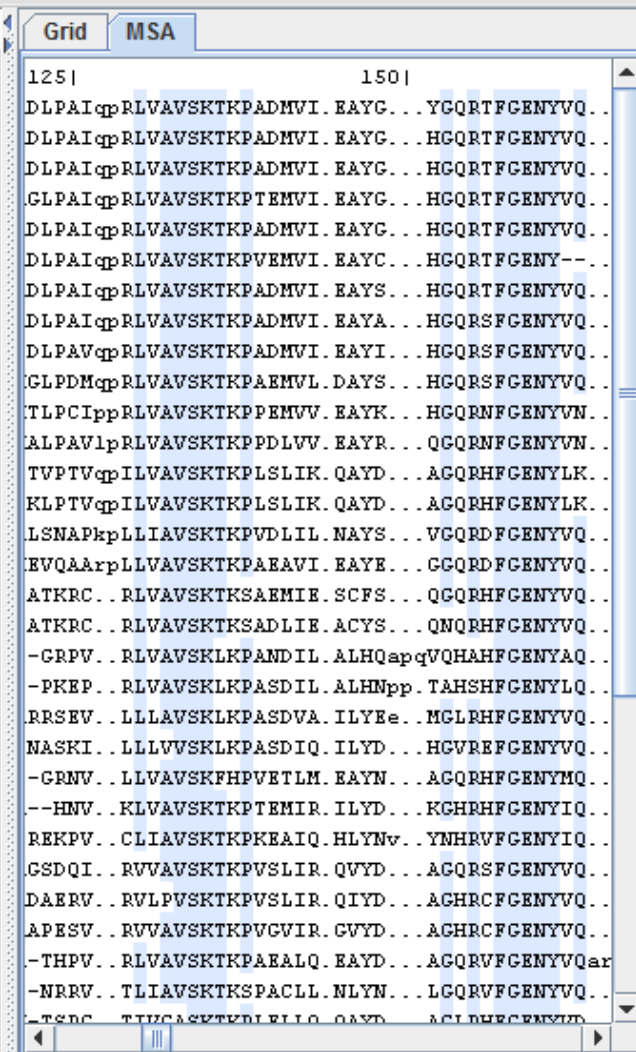
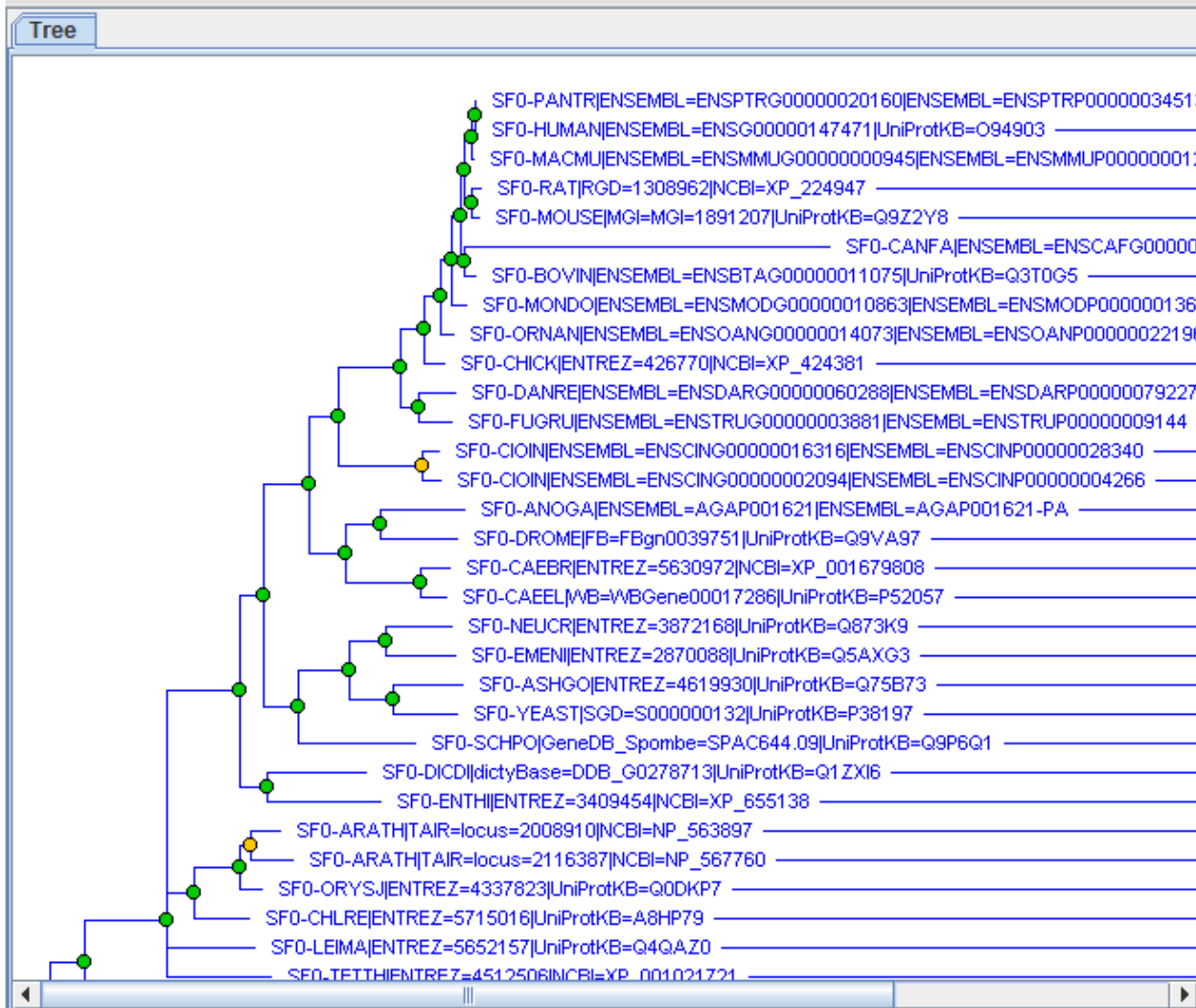
Downloads: [HMM](#) (HMMER format)

GENES ASSIGNED TO THIS FAMILY

Species	Count
Anopheles gambiae	1
Aquifex aeolicus vf5	1
Arabidopsis thaliana	2

Family Name : PROLINE SYNTHETASE CO-TRANSCRIBED BACTERIAL HOMOLOG PROTEIN (PTHR10146)

Tree MSA





Berkeley
Phylogenomics
Group

[HOME](#) | [PUBLICATIONS](#) | [FUNDING](#) | [CONTACT US](#)

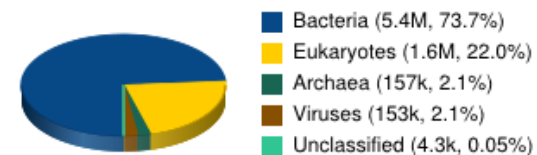
PhyloFacts 3.0.2

PhyloFacts release PF3.0.2 contains 7,337,238 protein sequences from 99,254 unique taxa (including strains) across 92,800 families (25,446 grouped by PFAM domain and 67,354 grouped by multi-domain architecture agreement). [More ...](#)

SEQUENCE ACCESSION SEARCH	Query PhyloFacts by UniProt accession or identifier
ORTHOLOG IDENTIFICATION	PhyloFacts Orthology Group: phylogenetic orthologs
JUMP TO PHYLOFACTS FAMILY	View PhyloFacts family alignments, trees, and annotations
PHYLOFACTS-PFAM SEARCH	Query PhyloFacts by Pfam accession (PhyloFacts-Pfam Project)
GENOME COVERAGE	View coverage of key species in PhyloFacts
STATISTICS	View PhyloFacts coverage statistics
DOWNLOADS	Download PhyloFacts data
CITING PHYLOFACTS	How to cite PhyloFacts

PhyloFacts statistics

7.3M unique proteins across the Tree of Life



PhyloFacts is funded by a grant from the Department of Energy, Division of Biological and Environmental Research ([details](#)).



Phylofacts

query by sequence search

PHOG0274269_00186 – Proline synthetase co-transcribed bacterial protein


PHOG tree:	View tree
Pfam domains:	Ala_racemase_N
Taxonomic distribution:	stramenopiles
PhyloFacts family:	bpg0243724
Alignment:	Global
Number of sequences:	4
Alignment length:	296

PhyloFacts Orthology Group Members

Gene ID	Species	Description	Swiss	GO	EC	KEGG	Lit.
D8LNZ4	Ectocarpus siliculosus (Brown alga)	Putative uncharacterized protein					
B7GC89	Phaeodactylum tricornutum (strain CCAP 1055/1)	Predicted protein					
D0MS28	Phytophthora infestans T30-4	Proline synthetase co-transcribed bacterial protein					
B8BUT1	Thalassiosira pseudonana (Marine diatom)	Predicted protein					

Download As CSV

Resource federation: InterPro

- [InterPro:Home](#)
- [Advanced Search](#)
- [InterProScan](#)
- [BioMart](#) 
- [Help / Documentation](#)
- [About InterPro](#)
- [Release Notes](#)
- [BioMart Manual](#)
- [Tutorial](#)
- [Publications](#)
- [Contributors](#)
- [Web Services](#)
- [Downloads](#)
- [Protein Focus](#)
- [Killer toxin Protein \(KP4\)](#)


EBI > Databases > InterPro

InterPro protein sequence analysis & classification

InterPro is an integrated database of predictive protein signatures used for the classification and automatic annotation of proteins and genomes. InterPro classifies sequences at superfamily, family and subfamily levels, predicting the occurrence of functional domains, repeats and important sites. InterPro adds in-depth annotation, including GO terms, to the protein signatures.

Current release: **37.0 30 April 2012** (see [Release Notes](#) for further details)

Do a sequence search of InterPro, via [InterProScan](#)

Extract large datasets by querying our [BioMart](#) 

You can access our data programmatically, via [Web Services](#)

Use the updated [InterProScan Web Service](#)

News

We are delighted to announce the release of InterProScan 5RC1: the first release candidate of InterProScan version 5. To obtain a copy of this release candidate, please visit [Running InterProScan 5](#) for complete documentation regarding downloading, installing and using InterProScan 5RC1.

A recently published paper describing new developments with the InterPro database is available at Nucleic Acids Research ([doi: 10.1093/nar/qkr948](#))

A paper describing InterPro's approach to Gene Ontology curation has also recently been published and is available at Database ([doi: 10.1093/database/bar068](#))

Not as easy as it may sound...

- Specific protein families may not be consistent across resources
- Most families (MSAs, trees, HMMs) in these resources are not manually curated
 - Domains in Pfam-A are curated
 - TIGRfams are curated
 - HAMAP families are curated

http://www.expasy.org



ExPASy
Bioinformatics Resource Portal

[Home](#) [About](#) [Contact](#)

Query all databases [help](#)

Visual Guidance

Categories

- proteomics
- genomics
- structural bioinformatics
- systems biology
- phylogeny/evolution
- population genetics
- transcriptomics
- biophysics
- imaging
- IT infrastructure
- drug design

Resources A..Z

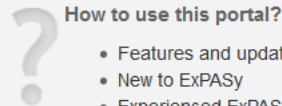
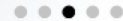
Links/Documentation

ExPASy is the **SIB Bioinformatics Resource Portal** which provides access to scientific databases and software tools (i.e., *resources*) in different areas of life sciences including proteomics, genomics, phylogeny, systems biology, population genetics, transcriptomics etc. (see [Categories](#) in the left menu). On this portal you find resources from many different SIB groups as well as external institutions.

Featuring today

SuperTree

Build phylogenetic supertrees
[\[details\]](#)



How to use this portal?

- Features and updates
- New to ExPASy
- Experienced ExPASy users: what is different

Popular resources

- UniProtKB
- SWISS-MODEL
- STRING
- PROSITE

Latest News

UniProt Knowledgebase release 2013_08 - 2013-07-24

UniProtKB/Swiss-Prot Release of 24-Jul-2013 contains 540,732 sequence entries...[More](#).
UniProtKB/TrEMBL Release of 24-Jul-2013 contains 41,451,118 sequence entries...[More](#)

Protein Spotlight: the intricacy of a smell - 2013-07-22

We all need a nose. Inside this part of an animal's body lie millions of olfactory receptors awaiting smells that they will send on to the brain. In turn, our brain will say whether the smell is good, or bad...[More](#).

[\[More news\]](#) [\[SIB news\]](#)



Query all databases help

Visual Guidance

Categories

- proteomics
- genomics**
- sequence alignment
- similarity search
- characterisation/annotation
- structural bioinformatics
- systems biology
- phylogeny/evolution
- population genetics
- transcriptomics
- biophysics
- imaging
- IT infrastructure
- drug design

Resources A..Z

Links/Documentation

- SIB resources
- External resources - *(No support from the ExpASY Team)*

Tools

- Alignment tools** • Four tools for multiple alignments • [\[more\]](#)
- boxshade** • MSA pretty printer • [\[more\]](#)
- ClustalW** • Multiple sequence alignment • [\[more\]](#)
- ClustalW - PBIL** • Multiple sequence alignment program • [\[more\]](#)
- ClustalW2** • Multiple sequence alignment program • [\[more\]](#)
- Codon Suite** • codon-based sequence analysis • [\[more\]](#)
- Decrease redundancy** • Sequence redundancy reduction • [\[more\]](#)
- DIALIGN** • Local multiple sequence alignment • [\[more\]](#)
- GENIO/logo** • RNA/DNA & Amino Acid Sequence Logos • [\[more\]](#)
- Kalign - EBI** • Fast and accurate multiple sequence alignment • [\[more\]](#)
- Kalign - SBC** • Fast and accurate multiple sequence alignment • [\[more\]](#)
- LALIGN** • Pairwise alignment • [\[more\]](#)
- MADAP** • clustering for genome annotation data • [\[more\]](#)
- MAFFT - CBRC** • Multiple sequence alignment • [\[more\]](#)
- MAFFT - EBI** • Multiple sequence alignment • [\[more\]](#)
- MaxAlign** • Gap removal from alignments • [\[more\]](#)
- Multialin** • Multiple sequence alignment • [\[more\]](#)
- MUSCLE** • Multiple alignment server • [\[more\]](#)
- Newick Utilities** • high-throughput phylogenetic tree processing • [\[more\]](#)
- Phylogibbs** • regulatory sites discovery • [\[more\]](#)
- SIBsim4** • spliced sequence alignment • [\[more\]](#)
- T-Coffee** • sequence and structure multiple alignments • [\[more\]](#)
- T-Coffee - EBI** • Multiple sequence alignment program • [\[more\]](#)
- T-Coffee - WUR** • Multiple sequence alignment program • [\[more\]](#)
- WebLogo** • Sequence logos • [\[more\]](#)

proteômica

Categories

proteomics

protein sequences and identification
mass spectrometry and 2-DE data
protein characterisation and function
families, patterns and profiles
post-translational modification
protein structure
protein-protein interaction
similarity search/alignment

genomics

structural bioinformatics

systems biology

phylogeny/evolution

population genetics

transcriptomics

biophysics

imaging








IT infrastructure

drug design

Resources A..Z

Links/Documentation

Databases

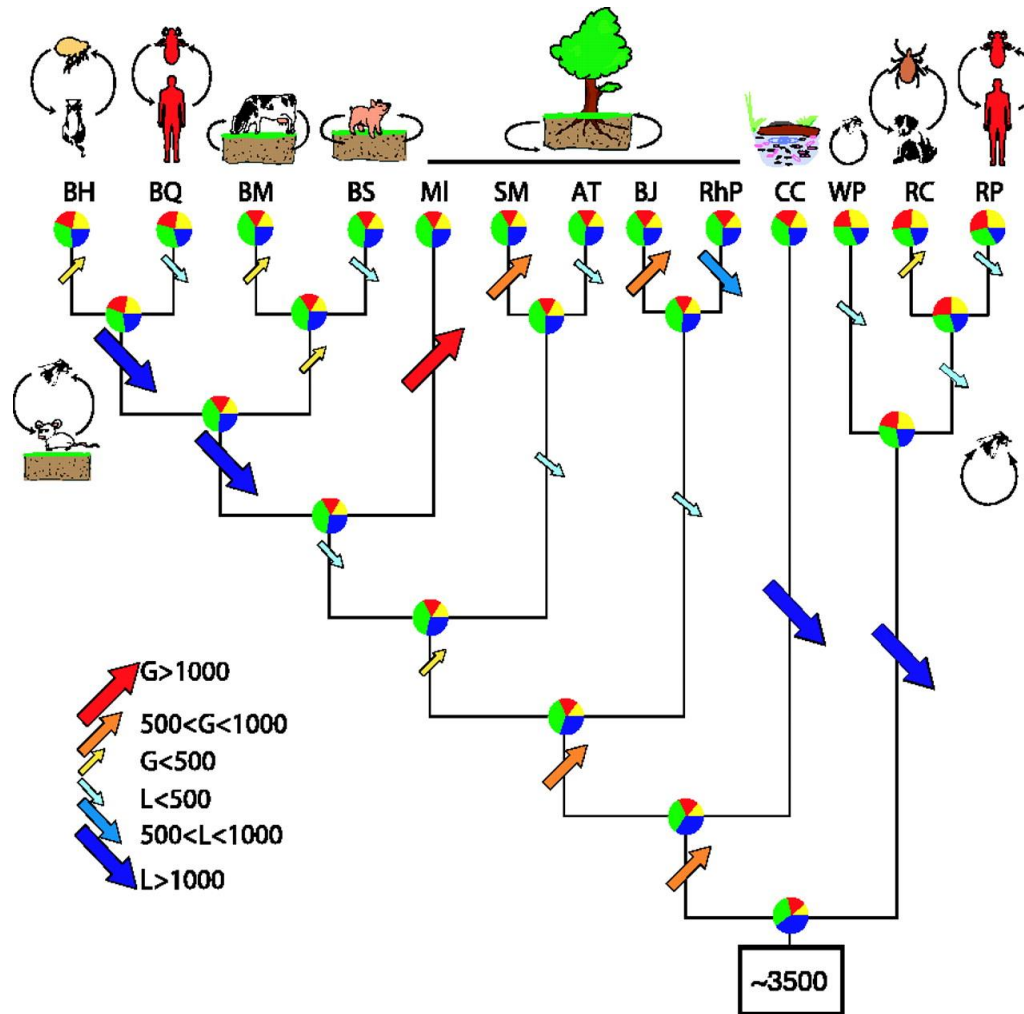
 [neXtProt](#) • human proteins • [\[more\]](#)
 [PROSITE](#) • protein domains and families • [\[more\]](#)
 [STRING](#) • protein-protein interactions • [\[more\]](#)
 [SWISS-MODEL Repository](#) • protein structure homology models • [\[more\]](#)
 [UniProtKB](#) • functional information on proteins • [\[more\]](#)
 [UniProtKB/Swiss-Prot](#) • protein sequence database • [\[more\]](#)
 [ViralZone](#) • portal to viral UniProtKB entries • [\[more\]](#)

 [EMBNet services](#) • bioinformatics tools, databases and courses • [\[more\]](#)
 [ENZYME](#) • enzyme nomenclature • [\[more\]](#)
 [GlycoSuiteDB](#) • glycan database • [\[more\]](#)
 [GPSDB](#) • gene and protein synonyms • [\[more\]](#)
 [HAMAP](#) • UniProtKB family classification and annotation • [\[more\]](#)
 [MetaNetX](#) • Metabolic Network Repository & Analysis • [\[more\]](#)
 [MIAPEGelDB](#) • MIAPE document edition • [\[more\]](#)
 [MyHits](#) • protein domains database and tools • [\[more\]](#)
 [PANDITplus](#) • protein families and domains resources • [\[more\]](#)
 [PaxDb](#) • protein abundance database • [\[more\]](#)
 [Prolune](#) • Popular science articles (in French) • [\[more\]](#)
 [Protein Model Portal](#) • structural information for a protein • [\[more\]](#)
 [Protein Spotlight](#) • Informally written reviews on proteins • [\[more\]](#)
 [SugarBind](#) • pathogen sugar-binding • [\[more\]](#)
 [SWISS-2DPAGE](#) • proteins on 2-D and SDS PAGE maps • [\[more\]](#)
 [SwissSidechain](#) • non-natural amino-acid sidechains • [\[more\]](#)
 [SwissVar](#) • variants in UniProtKB entries • [\[more\]](#)
 [TCS](#) • interaction specificity in two-component systems • [\[more\]](#)
 [UniMES \(UniProt metagenomic samples\)](#) • UniProt Metagenomic and Environmental Sequences • [\[more\]](#)
 [UniParc \(UniProt sequence archive\)](#) • UniProt sequence archive • [\[more\]](#)
 [UniPathway](#) • metabolic pathways for the UniProtKB • [\[more\]](#)
 [UniRef \(UniProt sequence clusters\)](#) • UniProtKB sequence clusters • [\[more\]](#)
 [World-2DPAGE Constellation](#) • set of 2DPAGE resources • [\[more\]](#)
 [World-2DPAGE Repository](#) • gel-based proteomics data • [\[more\]](#)

Tools

 [SWISS-MODEL Workspace](#) • structure homology-modeling • [\[more\]](#)
 [SwissDock](#) • protein ligand docking server • [\[more\]](#)
 [2ZIP](#) • Prediction of leucine zipper domains • [\[more\]](#)
 [3of5](#) • find user-defined patterns in protein sequences • [\[more\]](#)
 [AACompIdent](#) • protein identification by aa composition • [\[more\]](#)
 [AACompSim](#) • amino acid composition comparison • [\[more\]](#)
 [Agadir](#) • Prediction of the helical content of peptides • [\[more\]](#)
 [ALF](#) • simulation of genome evolution • [\[more\]](#)
 [Alignment tools](#) • Four tools for multiple alignments • [\[more\]](#)
 [AllAll](#) • protein sequences comparisons • [\[more\]](#)
 [APSSP](#) • Advanced Protein Secondary Structure Prediction • [\[more\]](#)
 [Ascalaph](#) • Molecular modeling software • [\[more\]](#)
 [big-PI](#) • predict GPI modification sites • [\[more\]](#)
 [Biochemical Pathways](#) • Biochemical Pathways • [\[more\]](#)
 [BLAST](#) • sequence similarity search • [\[more\]](#)
 [BLAST \(UniProt\)](#) • BLAST search on the UniProt web site • [\[more\]](#)
 [BLAST - NCBI](#) • Biological sequence similarity search • [\[more\]](#)
 [BLAST - PBIL](#) • BLAST search on protein sequence databases • [\[more\]](#)
 [Blast2Fasta](#) • Blast to Fasta conversion • [\[more\]](#)
 [boxshade](#) • MSA pretty printer • [\[more\]](#)
 [CFSSP](#) • Protein secondary structure prediction • [\[more\]](#)
 [ChloroP](#) • chloroplast transit peptides & cleavage sites • [\[more\]](#)
 [Click2Drug](#) • Directory of computational drug design tools • [\[more\]](#)
 [ClustalO \(UniProt\)](#) • Align two or more protein sequences • [\[more\]](#)
 [ClustalW](#) • Multiple sequence alignment • [\[more\]](#)
 [ClustalW - PBIL](#) • Multiple sequence alignment program • [\[more\]](#)
 [ClustalW2](#) • Multiple sequence alignment program • [\[more\]](#)
 [Coiled-Coils prediction](#) • Prediction of coiled coils regions • [\[more\]](#)
 [COILS](#) • Prediction of Coiled Coil Regions in Proteins • [\[more\]](#)
 [ColorSeq](#) • Color Protein Sequence • [\[more\]](#)
 [Compute pI/MW](#) • theoretical pI and Mw computation • [\[more\]](#)
 [CPHmodels](#) • Protein homology modeling • [\[more\]](#)
 [CSS-Palm](#) • Prediction of palmitoylation sites in proteins • [\[more\]](#)
 [DAS-TMfilter](#) • Prediction of transmembrane regions • [\[more\]](#)
 [Decrease redundancy](#) • Sequence redundancy reduction • [\[more\]](#)
 [DIALIGN](#) • Local multiple sequence alignment • [\[more\]](#)
 [DictyOGlyc](#) • GlcNAc O-glycosylation sites in D.discoideum • [\[more\]](#)
 [DisEMBL](#) • Prediction of disordered protein regions • [\[more\]](#)

Fig. 4. Net gene loss or gain throughout the evolution of the {alpha}-proteobacterial species



Boussau, Bastien et al. (2004) Proc. Natl. Acad. Sci. USA 101, 9722-9727