



Comparação de sequências

aula 3

João Carlos Setubal 2020

Alinhamento múltiplo

- Queremos alinhar mais do que n = 2 sequências
- *n* pode variar de 3 a milhares
- Por que haveria interesse em fazer tais alinhamentos?

Motivação mais geral

- Representante da situação em que semilaridade entre 2 sequências pode ser apenas coincidência
- Mas similaridade entre 10 ou 20 ou 100 sequências (ou seja, todas com todas) é muito mais difícil que seja coincidência

Motivação mais concreta

Para construir filogenias é necessário criar AMs

Alinhamento múltiplo

```
C: NKNKGVVWTKETLFEYLLNPKKYIPGTKMVFAGLKKADERADLIKYIEVESAKSL

Y: NIKKNVLWDENNMSEYLTNPKKYIPGTKMAFGGLKKEKDRNDLITYLKKACE---

A: NKSMAVNWEEKTLYDYLLNPKKYIPGTKMVFPGLKKPQDRADLIAYLKEGTA---

D: NKAKGITWNEDTLFEYLENPKKYIPGTKMIFAGLKKPNERGDLIAYLKSATK---

H: NKNKGIIWGEDTLMEYLENPKKYIPGTKMIFVGIKKKEERADLIAYLKKATNE--

M: NKNKGITWGEDTLMEYLENPKKYIPGTKMIFAGIKKKGERADLIAYLKKATNE--

* : * :..: :** ******** * *:** :* *** *:*
```

JC Setubal

Como dar notas para alinhamentos múltiplos?

Soma de pares (aminoácidos)

- Numa coluna determinada, podemos separar todos os pares de aminoácidos
 - da linha 1 com linha 2, da linha 1 com linha 3, 1 com 4, etc
 - depois: da linha 2 com linha 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

BLOSUM62

Ш	-	_	
ىل			
ш			

L/L: 4

L/I: 2

$$\nabla$$

L/V: 1

I/I: 4

V

I/V: 3

V/V: 4

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

Nota da coluna

```
C: ----SDIPAGDYEKGKKVYKQRCLQCHVVDSTAT-KTGPTLHGVIGRTSGTVSGFDYSAA

Y: ----TEFKAGSAKKGATLFKTRCLQCHTVEKGGPHKVGPNLHGIFGRHSGQAEGYSYTDA

A: MASFDEAPPGNPKAGEKIFRTKCAQCHTVEKGAGHKQGPNLNGLFGRQSGTTPGYSYSAA

D: -----GVPAGDVEKGKKLFVQRCAQCHTVEAGGKHKVGPNLHGLIGRKTGQAAGFAYTDA

H: ------GDVEKGKKIFIMKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAAGFSYTDA

** : * :: . * ***.*: . * ***.*:*: * . *: *: *: *
```

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

As cores representam famílias de aminoácidos

Cthe 1566 Clostridium thermoce VLDT-DLOVV DDVDSLVSEF Fisuc 1086 Fibrobacter succino VFDG-DLYVV LTGCVGGLIG DDVPSLVNEY RD-RYDP-ELISV DDIEAFIRAA Metvu 1085 Methanocaldococcus VTTCSSETIG DDTEAFTRAA MFS40622 0035 LVEGLENLVA RYDP-DLISV Metin 0037 Methanocaldococcus RYDP-DLISV VTTCSSETIG Csac 2462 Caldicellulosiruptor RYSP-TVIGV DDIEAFIKEA RYRP-DLITV VTTCSSEIIG Daud 0147 Candidatus Desulforu DDMVSFIKVA RYWP-GLIGV RARLSREIGE Slip 2126 Syntrophothermus lip VTTCSSEIMG CT1536 nifD Chlorobium tepidum DDVHAVAREM KEKLGD LFHP-KAIAI FSTCPVGLIG KEKLGD Cphamn 1754 Chlorobium phaeob IFRP-KAIGI FSTCPVGLIG DDVHAVAREM MM0722 NifD Methanosarcina maz CATCPVGLIG DDIEAVSREA EKEHG-LFPLNKGISV GAELS Avin 01390 nifD Azotobacter vi OSECPIGLIG DDIESVSKVK Moth 0551 Moorella thermoaceti DDVOAVARSV FTTCTTGLIG Slip 2127 Syntrophothermus lip FTTCTTGLIG DDVOGVAROV EKEVG LKASCLEAFE Daud 0146 Candidatus Desulforu ERETG Metvu 1084 Methanocaldococcus EFPEAKGIII YATCTTGLIG DNLGAVAKKV EEKIG LEKACLEAAA MFS40622 0034 Methanocaldococc EFPOAKGIII YATCTTGLIG EEKIG Metin 0038 Methanocaldococcus EFPEAKGIFI YATCPTALIG DNLEAVARKV EEKIK Csac 2463 Caldicellulosiruptor EFPEAKAVFI YATCPTALIG DDLEAVAKKA SKAIG RoseRS 1199 Roseiflexus **EFPNAKAVFV** YNTCSTALIG DDGRDVAKOA EAIIG DDGRDVAKOA Rcas 4041 Roseiflexus **EFPDAKAVFV** YNTCSTALIG EAIIG LLOSTIEASA LYKSLIELID DDIDAVCKKV AKEKG CT1538 nifE Chlorobium tepidum YSTCIIGLIG MM0724 nifE Methanosarcina maz YSTCIVGIIG DDLEAVCKTA Avin 01450 nifE Azotobacter vi AERFG YNTCVPALIG DDVDAVCKAA Cthe 1565 Clostridium thermoce LTTCAAGIIG DDVESVCNEA EEELG Fisuc 1087 Fibrobacter LROTIRDAKE RFNP-KAIFI GMACATAIIG **EPEV**G EDIDSIAEEM Ccel 1615 Clostridium cellulol LVDSLNEVNS RYNP-KIIAV LTNCCADIIG DDVEGCIEGL PDEIR-Mlab 1039 Methanocorpusculum 1 LGTPVPALIG LLNKILOECA SHHP-KFVAI CDISGIATEV FDTTK Mlab 1040 Methanocorpusculum l

Exercício

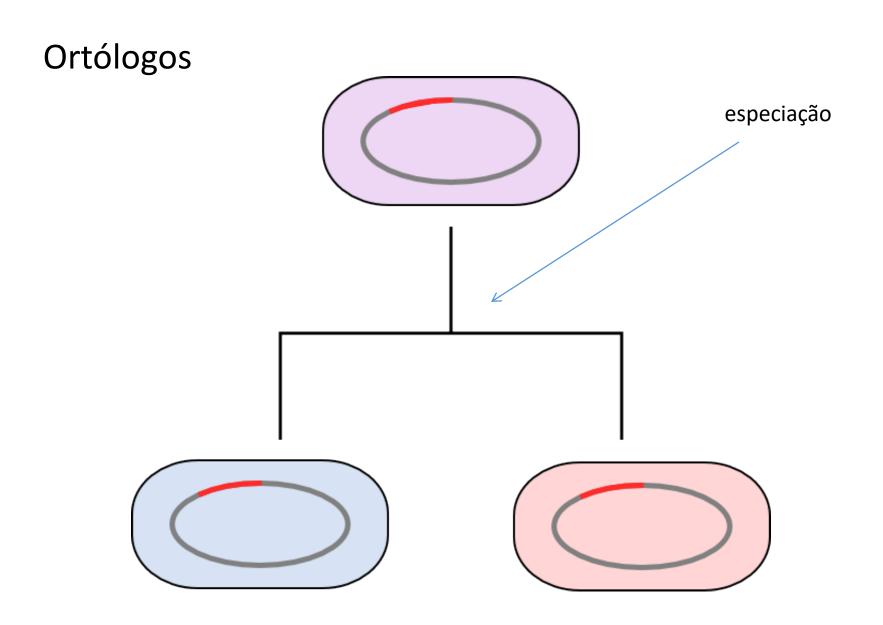
- Compare os aminoácidos de mesmas cores do slide anterior com as famílias apresentadas na aula 1 de comparações
- Compare também com as notas desses grupos de aminoácidos na matriz BLOSUM62

Sequências de entrada

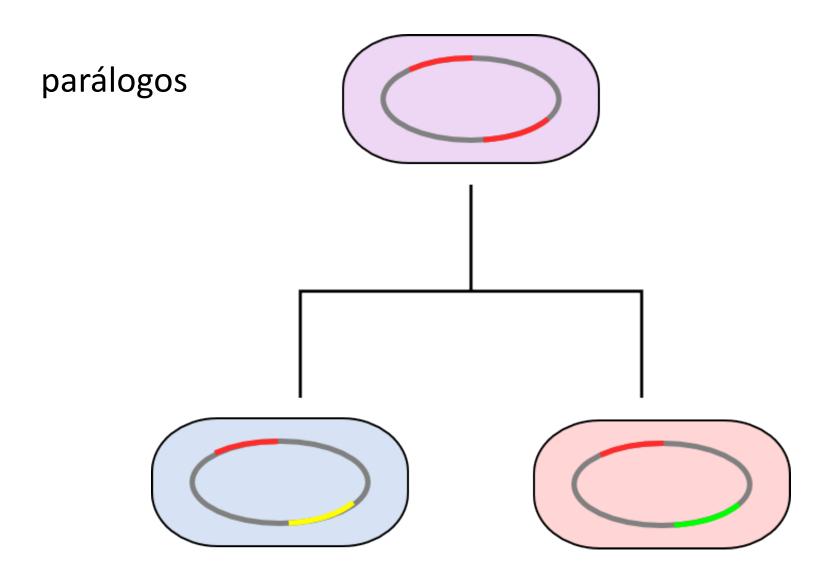
- Dois conceitos importantes
 - Homologia
 - Família

Homologia

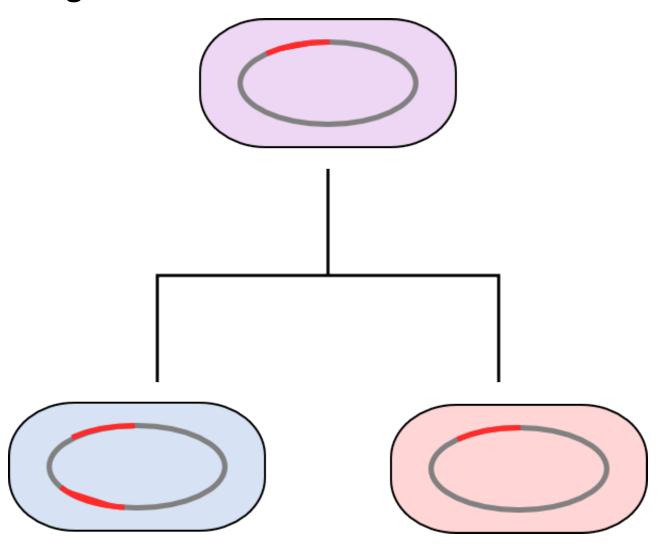
- Dois genes que tem um mesmo ancestral são homológos
- Freq. usado <u>erroneamente</u> com o sentido de <u>similar</u>
- Similaridade n\u00e3o implica necessariamente em homologia
 - Asas: morcêgo e insetos (convergência)
- Às vezes a similaridade é (ou parece) baixa mas mesmo assim existe homologia
 - Barbatana de baleia e braços em humanos
- Dois tipos de homologia
 - Ortologia e paralogia



JC Setubal



In-parálogos



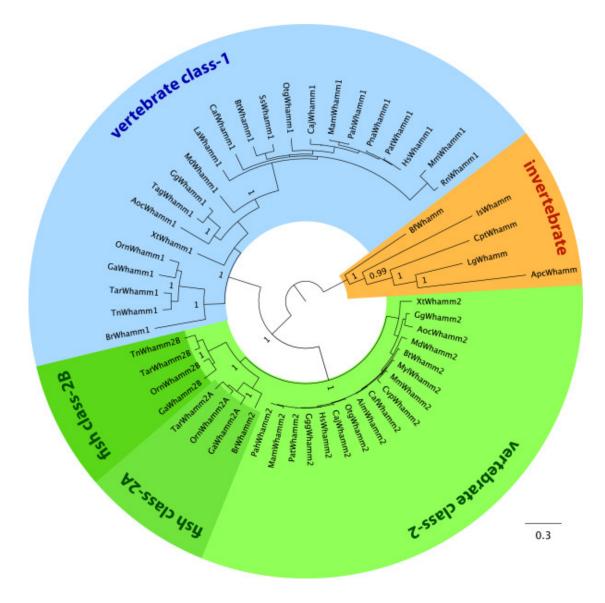
Homologia e função

- Seria bom se proteínas 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
 - Mas existem os conceitos de 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 <u>ortólogos</u>
- Falar em proteínas homólogas é um certo abuso de linguagem: são os genes que são homólogos

Exemplo de subfamílias. Nesta figura são definidas 3 subfamílias (azul, verde, laranja), e 3 subsubfamílias (dentro da subfamília verde)



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

 Uma coluna homóloga significa que o gene ancestral comum das sequências no AM também tinha a posição correspondente a essa coluna

Alinhar DNA ou aminoácidos?

- DNA: mais difícil garantir homologia nas colunas
- DNA é mais sensível, mas a terceira 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 sequências de proteínas são todas idênticas ou quase idênticas
 - Como seria o caso na comparação de proteínas de cepas de uma espécie de bactéria

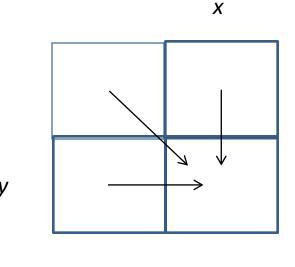
Algoritmo para alinhamento múltiplo de sequências

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

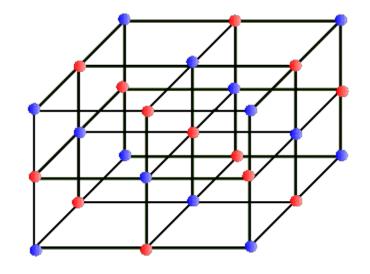
Generalização de PD para AM

2 sequências

3 sequências



 $O(n^2)$



$$O(n^3)$$
 $\Omega(2^k n^k)$

Omega significa que o custo é pelo menos esse

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

- Essas limitações vão variar de programa para programa, e dependendo de onde o programa é rodado
- A grosso modo, as sequências de entrada não podem ser:
 - muito longas (não mais do que algo como 10 kb)
 - nem muitas (não mais do que algo como 1000)

Alinhamento progressivo

- é a heurística que está na base de vários programas de AM
- 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
 - as sequências são adicionadas ao alinhamento num processo guiado pela árvore
- Seria melhor que AM e árvore fossem feitos simultaneamente
 - Mas é muito mais complicado de fazer com rigor

Programas para AM

Muscle

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

MAFFT

- Katoh, Misawa, Kuma, Miyata 2002 (Nucleic Acids Res. 30:3059-3066)
- http://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

Para ilustrar as complexidades de avaliação de alinhamentos múltiplos

 Artigo do próximo slide procurou comparer diferentes programas de AM entre si



Subject Areas

For Authors

About Us

Search

h Q

advanced search



8,544

32

141SAVES

5 SHARES

RESEARCH ARTICLE

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	bout the Authors	Metrics	Comments	Related Content

Abstract

Introduction

Results

Discussion

Materials and Methods

Acknowledgments

Author Contributions

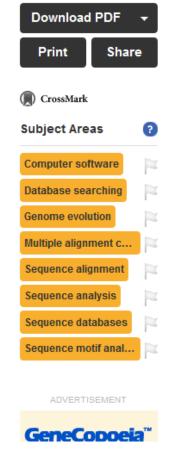
References

Reader Comments (0)

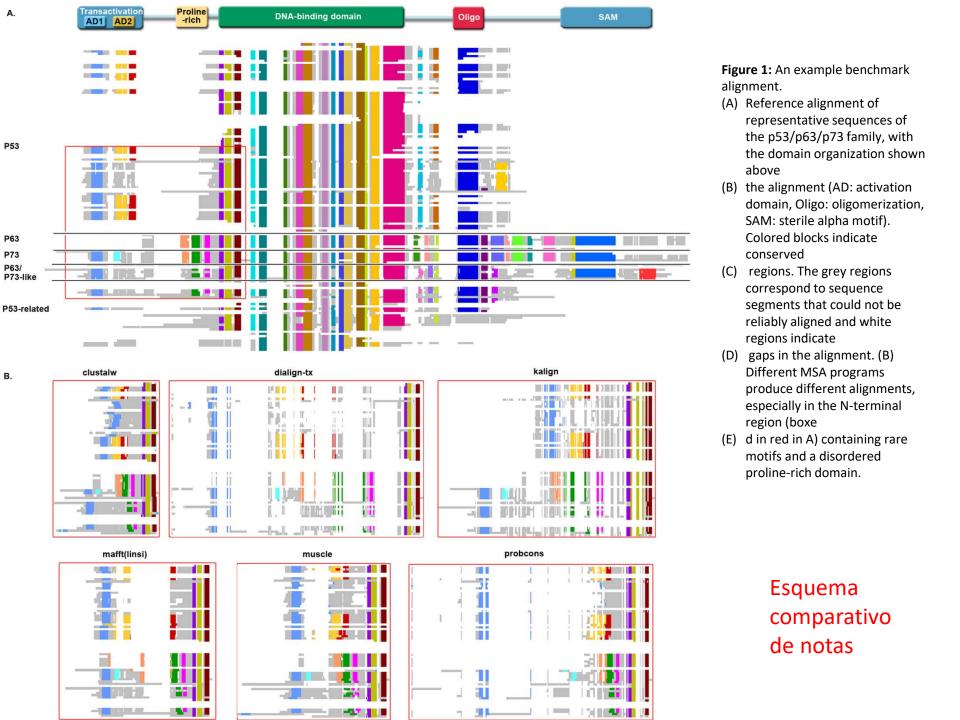
Figures

Abstract

Multiple comparison or alignmentof 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 alignmentmethods 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



10.1371/journal.pone.0018093



Alinhamentos múltiplos particularmente importantes podem ser editados manualmente

- A premissa é que um especialista será capaz de identificar alterações no AM que fazem mais sentido biológico
- O especialista em geral tem uma noção da estrutura das proteínas que estão alinhadas, o que nenhum programa de AM tem
- Algumas colunas podem não ser informativas, e deveriam ser removidas
 - sempre de acordo com o especialista!

O trecho indicado pela flecha poderia ser alvo de edição, para que os Ds ficassem alinhados

Cthe_1566_Clostridium_thermoce Fisuc 1086 Fibrobacter succino	NTVNLIA GLVNLWF	-SIPYQDPFW -EVPYYNONW	KGNLAEYKR- RGDYOELAR-	LLAG	IGL <mark>K</mark> ANVLFG AGFEVNVLFG
Metvu 1085 Methanocaldococcus	MIVPGFG		KRILNMFGLK	EGEDYSVLFD	ISETLYOPLR
MFS40622 0035	MIVPGFG	-VNPGDILEI	KRILDMFGLK	EGEDYSVLFD	ISETLYOPLR
Metin_0037_Methanocaldococcus_	NIIPGFG	-VNPGDILEI	KRILDMFGLK	EGEDYSILFD	OPLR
Csac 2462 Caldicellulosiruptor	YIIPGM	- INPGDIEEI	KHILDLMKI-	DYSVLFD	ISKIL VM
Daud_0147_Candidatus_Desulforu	NIIPGM	-LMPGDLREI	KHLLAEMGV-	EAHVLFD	
Slip_2126_Syntrophothermus_lip	NIIPGM	-LYPGDIREI	RHLLEEMNV-	EG <mark>TVLF</mark> D	ISDTLDAPL
CT1536_nifD_Chlorobium_tepidum				<mark>YN</mark> IGG	DAFEIERLLE
Cphamn1_1754_Chlorobium_phaeob		<mark>GG</mark>	KFKINMLGE-	<mark>YN</mark> IGG	DAFE I ERLLE
MM0722_NifD_Methanosarcina_maz		PT		<mark>YN</mark> IGG	DLWEIKPILE
Avin_01390_nifD_Azotobacter_vi		AST		<mark>YN</mark> IGG	DAWSSRILI
Moth_0551_Moorella_thermoaceti				Y <mark>NMD</mark> W	DLKAIRP A
Slip_2127_Syntrophothermus_lip		EEE <mark>KT</mark>			DIA
Daud_0146_Candidatus_Desulforu		EAERT	PYDVVLVGE-	YNMDW	PLME
Metvu_1084_Methanocaldococcus_		EEE <mark>KT</mark>		Y <mark>N</mark> MDW	
MFS40622_0034_Methanocaldococc		EEE <mark>KT</mark>		Y <mark>N</mark> MDW	DVAVIKPLFE
Metin_0038_Methanocaldococcus_				YNMDW	DL <mark>K</mark> VI <mark>KP</mark> LFE
Csac 2463 Caldicellulosiruptor		EDPT	PYDVNIIGE-	YNIDG	DYWVLEKLFT
RoseRS_1199_Roseiflexus				YNIKN	DIRTFEYLFE
Rcas 4041 Roseiflexus		EGDF	SRSINIIGD-	YNIKN	DIRTFEYLFE
CT1538_nifE_Chlorobium_tepidum	- T EG	IS	KYSINILGE-		EAWIIREYYE
MM0724_nifE_Methanosarcina_maz		HSKIEKMEIK	KPKINILGD-	FNVAG	DVWLVKPLFE
Avin_01450_nifE_Azotobacter_vi	GSERP	G <mark>IR</mark>	VHDVNLIGE-	YNIAG	EFWHVLPLLD
Cthe 1565 Clostridium thermoce			DDMINVINF-	(<mark>WGSD</mark> -)	VFYEWFA
Fisuc_1087_Fibrobacter		NO	NDLLNIVAL -	WGTD	YFSEMLA
Ccel_1615_Clostridium_cellulol		<mark>NQ</mark>	EKSVNLFLRR	-WIWDOTKOE	EINEIKRMFQ
Mlab_1039_Methanocorpusculum_1			PKTVNILGYT TLNFIGN-	PIDFFLSG NVPIN	DDRRLASFVE SNSEIYEVLN
Mlab_1040_Methanocorpusculum_1			171111-1 <mark>G</mark> IN-	INVPIN	
9/15/2020	J. C. S	etubal			33

Edição manual de AMs

Jalview

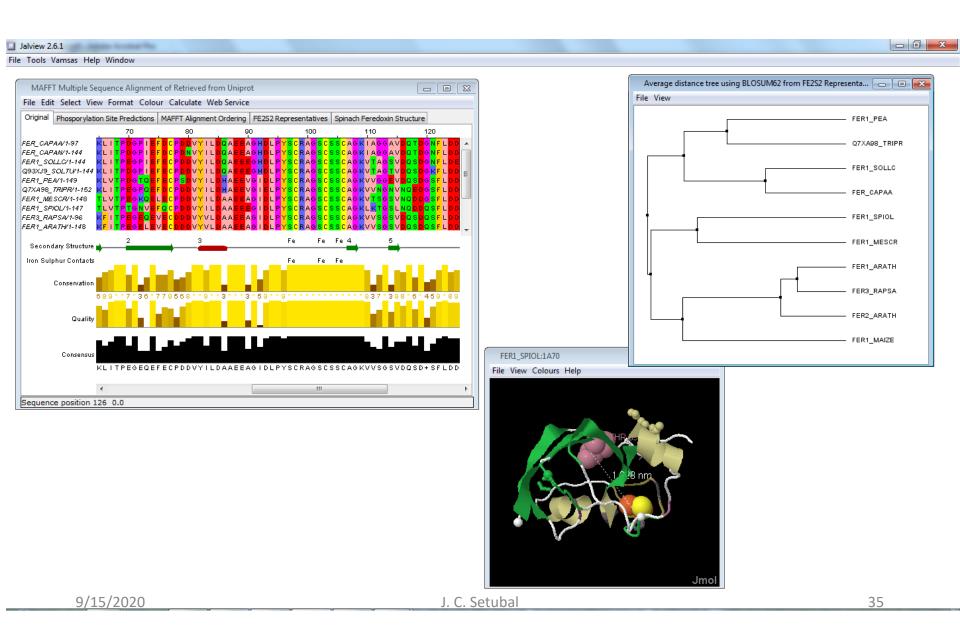
- www.jalview.org
- Waterhouse et al. *Bioinformatics* 2009 **25** (9) 1189-1191

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

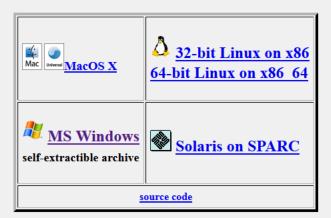
NEW: seaview drives the **Gblocks** program to select blocks of conserved sites. NEW: seaview drives the **Clustal** Ω program to perform multiple sequence alignment.

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
 - o parsimony, using PHYLIP's dnapars/protpars algorithm,
 - o distance, with NJ or BioNJ algorithms on a variety of evolutionary distances,
 - o 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, phyml, 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

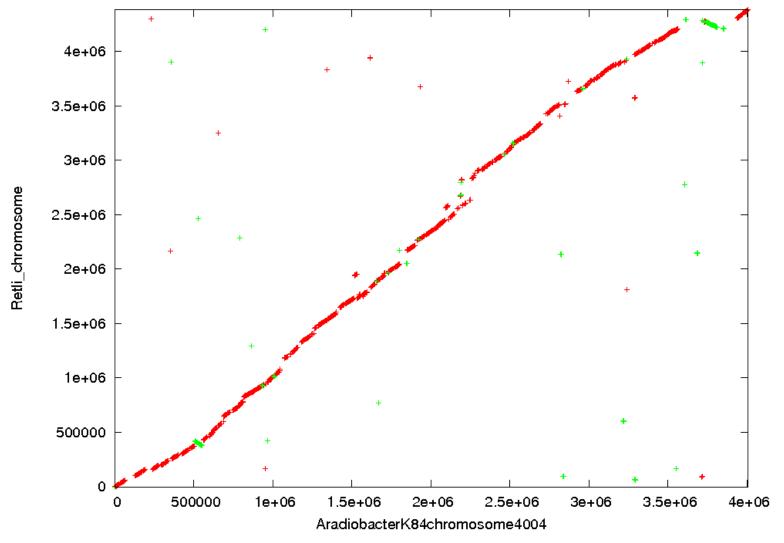
Existem diferentes formatos de alinhamentos múltiplos

- clustal, FASTA, MSF, NEXUS, PHYLIP
- Portanto é preciso cuidado quando se usa a saída de um programa de AM como entrada para um outro programa; os 2 programas tem que estar de acordo quanto ao formato!
- É em geral simples de se converter de um formato para outro
- http://molecularevolution.org/resources/fileformats/converting

Alinhamento entre sequências longas

- Por exemplo, cromossomos inteiros
- O cromossomo típico de uma bactéria tem 4 Mbp
- Cromossomo de humanos: algo como 300 Mbp

Este é um dotplot representando o alinhamento entre os cromossomos de duas bactérias: *Agrobacterium radiobacter* e *Rhizobium etli*



BLAST não serve para isso

- Mesmo computadores com centenas de GBytes 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
- Nucmer é para comparação de nucleotídeos
- Promer faz tradução nos 6 quadros de leitura de ambas as sequências (a la tblastx)

Árvore de sufixos para GTATCTAGG

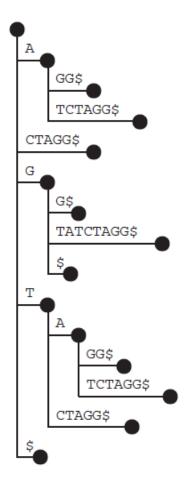


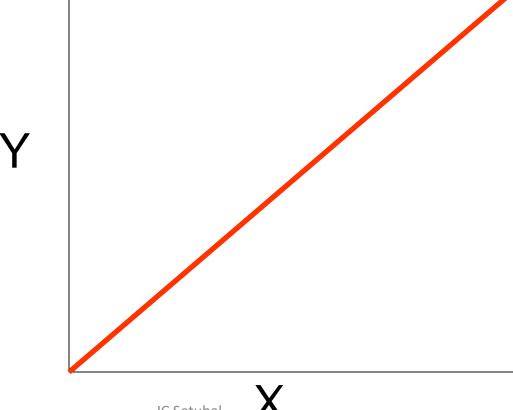
FIGURE 3.19

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

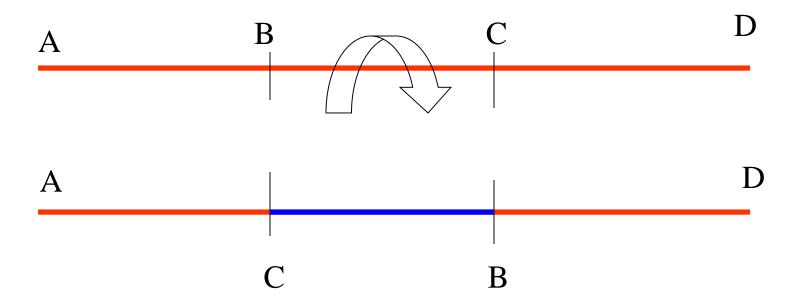
 Alinhamentos de cromossomos podem revelar rearranjos genômicos

Alinhamentos de cromossomos

Se as sequências (X e Y) fossem idênticas, veríamos isto num dotplot:

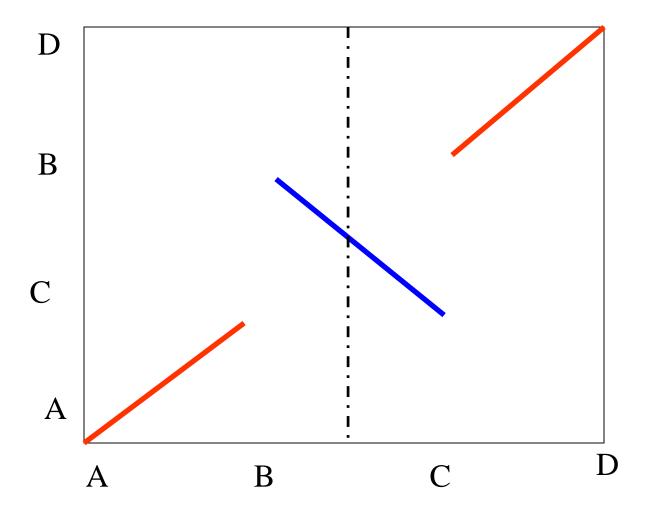


Suponha agora que houve uma inversão no cromossomo X



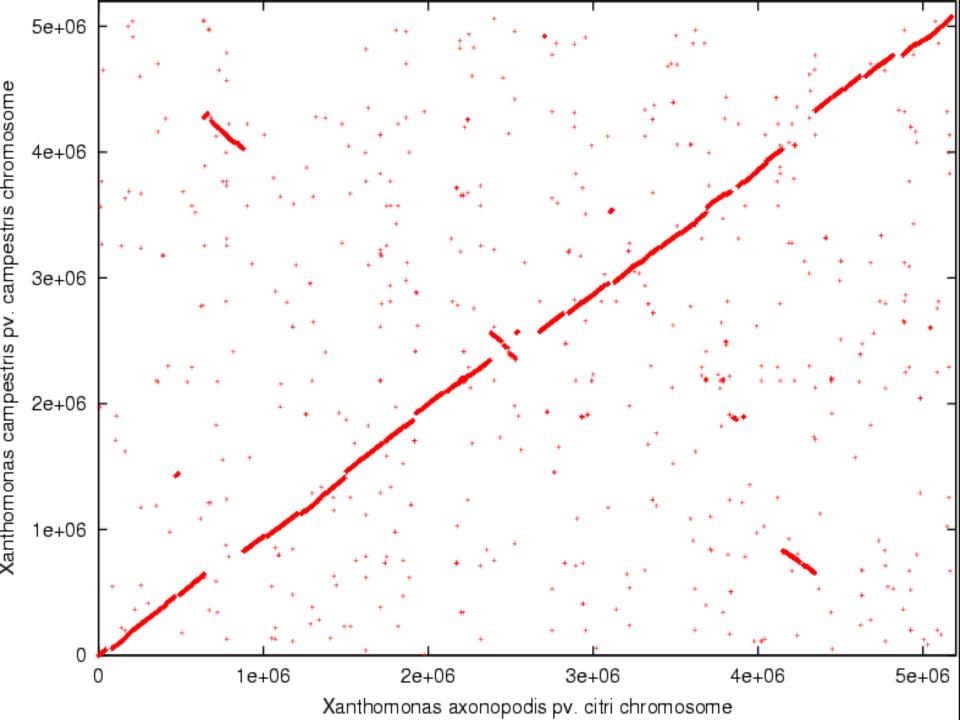
As letras são apenas rótulos para identificar posições ao longo das sequências

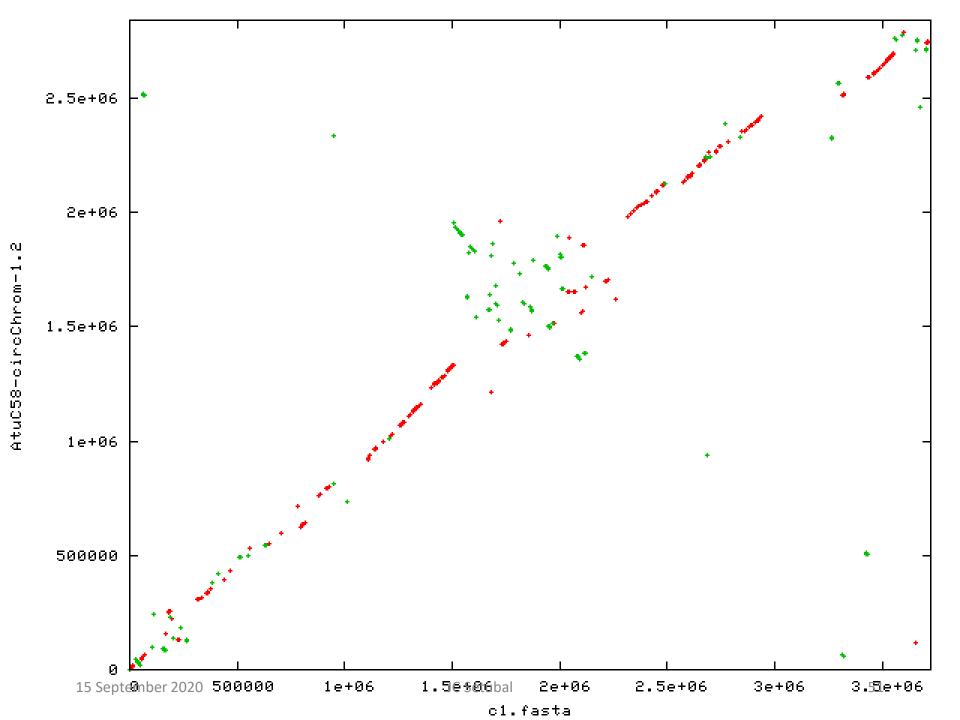
O dotplot entre X com inversão e Y ficaria assim

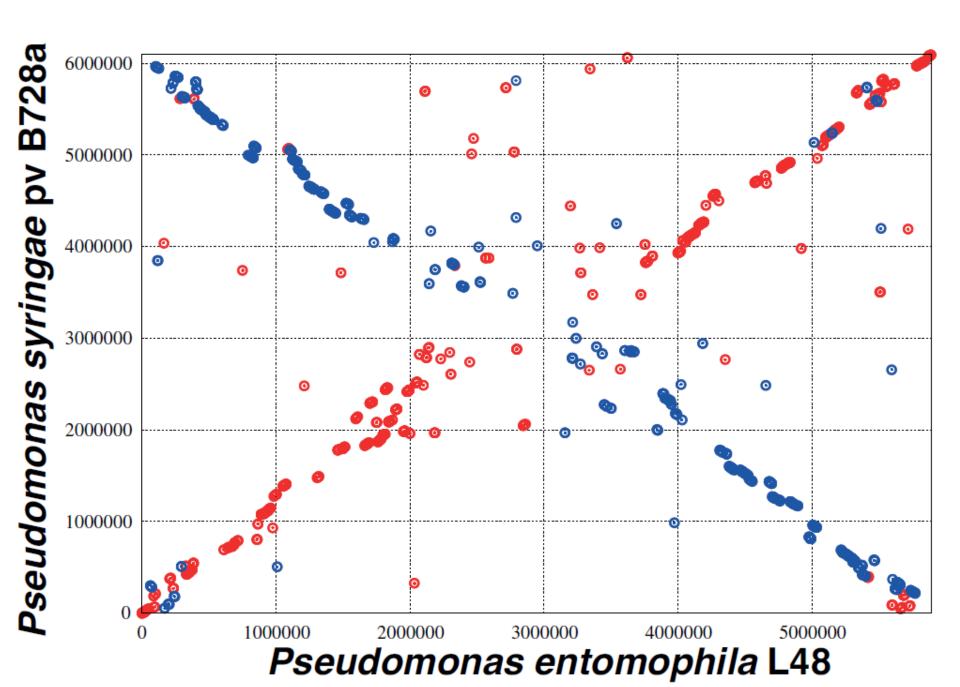


Such inversions seem to happen around the origin or terminus of replication

Vamos ver agora alguns exemplos de alinhamentos reais de cromossomos de diferentes bactérias

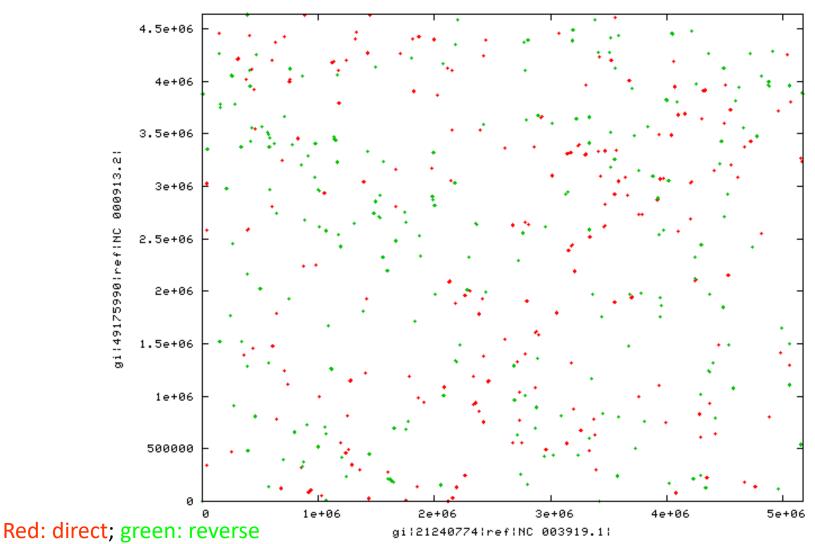






E. coli K12

Promer alignment

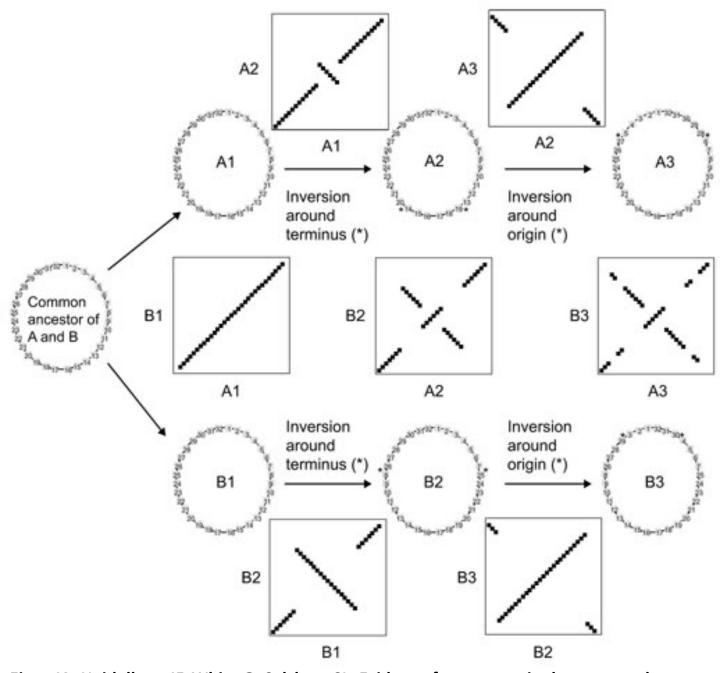


Xanthomonas axonopodis pv citri

Both are γ proteobacteria!

Observações

- Todos os alinhamentos menos o ultimo apresentam um padrão em X
- como explicar?
- J. Eisen et al. (2000) propuseram um modelo para explicar esse padrão
- Esse modelo supõe repetidas inversões ocorrendo nos cromossomos de espécies descendentes de um mesmo ancestral



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

E o último alinhamento?

- Aquele que alinhou Xanthomonas citri com Escherichia coli
- Resultou numa nuvem de pontos
- de acordo com o modelo, a explicação seria
 - houve tantas inversões nesses 2 cromossomos, que o sinal da diagonal se perdeu totalmente