Universidade de São Paulo
Instituto de Química

Comparação de sequências

aula 2

João Carlos Setubal

2021

Queremos descobrir sequências aparentadas

- Aparentadas = ancestral comum = homólogas
- Para inferir que duas sequências são homólogas, precisamos obter um alinhamento entre elas que seja biologicamente relevante
- Obter o alinhamento ótimo por si só não nos informa sobre parentesco
  - Alinhamentos de nota máxima (ótimos) não necessariamente correspondem a alinhamentos biologicamente relevantes
- Como fazer?
  - Para responder, temos que abordar o tema bancos de sequências

Bancos de sequências

- Situação típica
  - Tenho uma sequência-consulta (*query*)
  - 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
- O banco mais importante é o **GenBank**
  - Mantido pelo *National Center for Biotechnological Information*
  - **NCBI**
  - http://www.ncbi.nlm.nih.gov

NCBI
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.

II 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

06 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

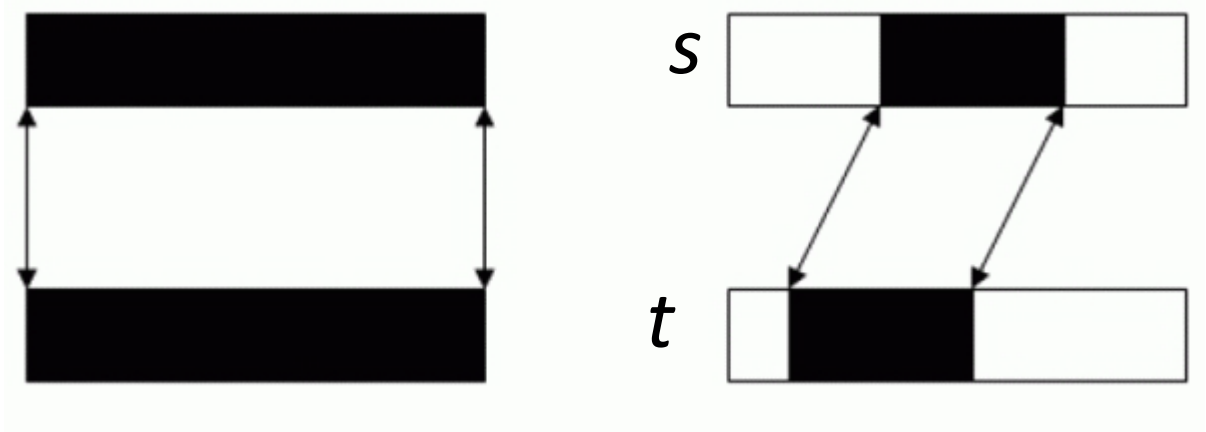
More...

Outro banco: UniProt http://www.uniprot.org/

Estatística de alinhamentos

- Com um banco, temos uma “população” de sequências; será nessa população que farei a busca com minha sequência-consulta
- Essa população é de milhões de sequências; isto quer dizer que, dependendo da sequência consulta, há uma probabilidade não muito pequena de que minha busca resulte em “hits” que são fruto do acaso
- Para entender por quê, é preciso apresentar o conceito de *alinhamento local*

Alinhamento global e local

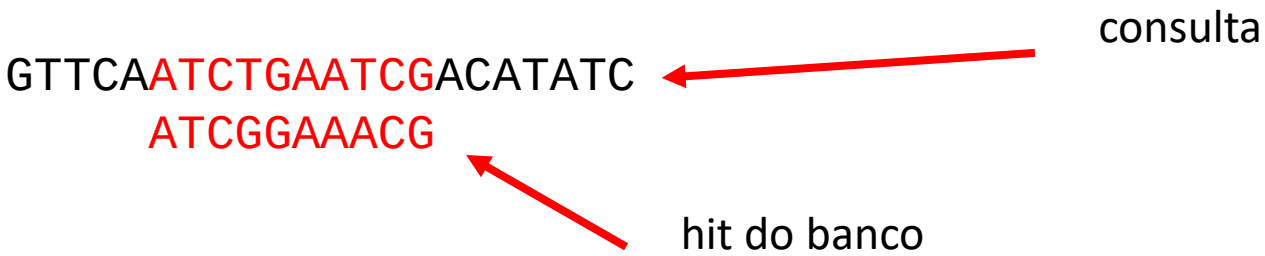


Global: toda a sequência *s* é alinhada com toda a sequência *t*

Local: um trecho de *s* é alinhado com um trecho de *t*, tal que esse alinhamento é “bom”, quando comparado com outros possíveis alinhamentos locais entre *s* e *t*

Alinhamento local é o modo de busca em bancos de sequências

- Exemplo
  - sequência-consulta: GTTCAATCTGAATCGACATATC
  - possível alinhamento local:



- Suponha agora que a sequência-consulta tem apenas 1 base, por exemplo, ‘G’
- Com tal sequência-consulta, basicamente todos os alinhamentos resultantes da busca serão devidos ao acaso

- Ninguém faria uma busca com uma consulta de apenas 1 base
- Mas agora generalize essa ideia para 2 bases, 3 bases, etc
- Não deve ser difícil de ver que para sequências-consulta curtas, a probabilidade de alinhamentos ao acaso é alta, e que essa probabilidade diminui quanto maior for a sequência-consulta
- Esta observação está na base da teoria estatística de alinhamentos

Alinhamentos estatisticamente significativos

- Teoria de Karlin e Altschul
- Permitiu uma quantificação da significância estatística de alinhamentos
- Essa quantificação se chama e-value

e-value

- e-value (expect value) exprime a significância estatística de um alinhamento
- $e\text{-}value = E = Kmne^{-\lambda S}$
- *m* e *n* são os tamanhos das sequências
- *S* é a pontuação (nota, score) do alinhamento
- *K* e λ são parâmetros
- Um banco de sequências pode ser tratado como uma longa sequência de tamanho *n* (concatenando-se as sequências entre si)
- 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
  - Probabilidades são sempre números no intervalo [0,1]
- Conversão de e-values para p-values: $P = 1 - e^{-E}$

Alinhamentos estatisticamente significativos e alinhamentos biologicamente relevantes

- Lembre-se que ao fazer buscas em bancos de sequências estamos em busca de sequências do banco que sejam aparentadas ou homólogas da nossa sequência-consulta
  - um alinhamento entre sequências homólogas é um alinhamento biologicamente relevante
- Como em geral é impossível saber se existe o parentesco, iremos inferir o parentesco por meio da significância estatística
- **Diremos que os alinhamentos estatisticamente significativos são biologicamente relevantes**
  - isto é uma aproximação, sem garantias! mais detalhes à frente

Na literatura

- É comum o termo “homology search”
- Esse termo se refere exatamente ao que estamos apresentando aqui
- Mas o mais correto seria dizer “similarity search”, pois a homologia é sempre uma inferência a partir da similaridade e da significância estatística

Detalhes importante sobre e-values

- E-value depende do tamanho do banco (veja a fórmula apresentada acima)
- Por esse motivo, não se pode comparar diretamente e-values obtidos de consultas a bancos diferentes, ou mesmo consultas ao mesmo banco feitas em tempos muito distintos (por exemplo, separadas por anos)
- 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

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.

Para quem quiser se aprofundar

http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html

Programação dinâmica é cara

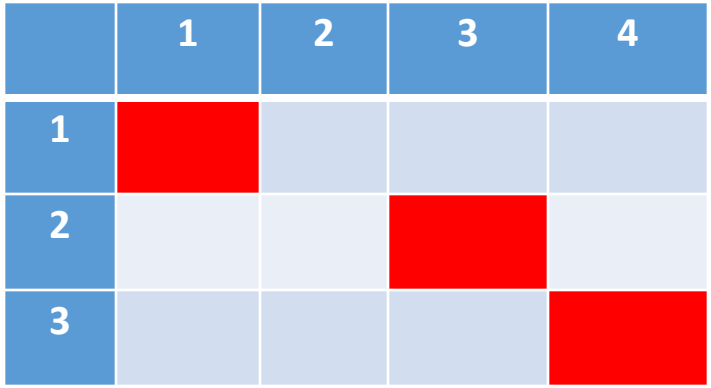
- Especialmente quando
  - Comparação contra muitas sequências
    - Buscas em banco de dados
  - Comparação de muitas sequências entre si
    - Todas contra todas
- Alternativa: BLAST
- Basic Local Alignment Search Tool
- Este é o programa básico de busca em bancos de sequências
- Veja que “alinhamento local” faz parte do nome

BLAST

- Altschul et al., 1990, 1997
- Cerca de 75 mil citações (em outubro de 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 os alinhamentos ótimos são de fato encontrados
  - Reporta e-values
  - (É possível fazer cálculo de e-values também com PD)

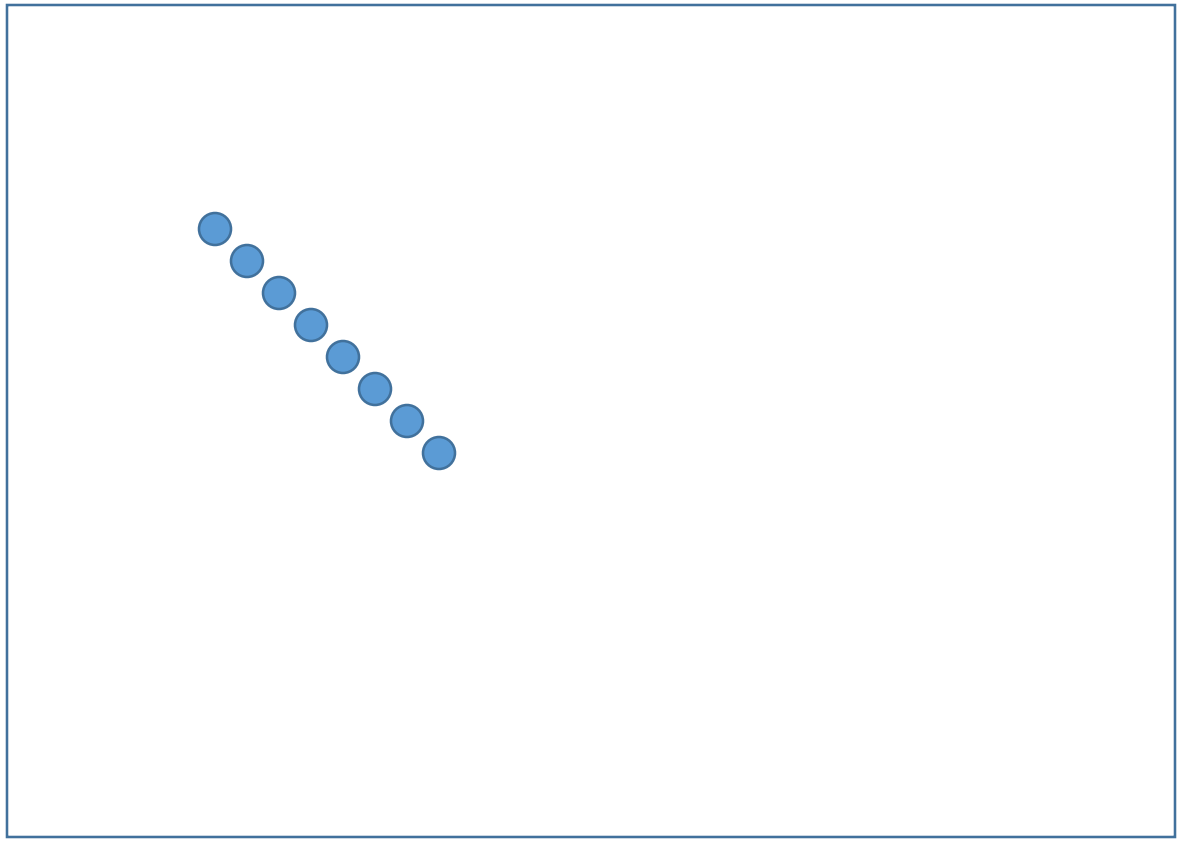
É útil pensar na matriz de PD como se fosse um dotplot

	j	0	1	2	3	4
i		t	G	A	T	C
0	s	0	-2	-4	-6	-8
1	G	-2	1	-1	-3	-5
2	T	-4	-1	0	0	-1
3	C	-6	-3	-2	-1	1

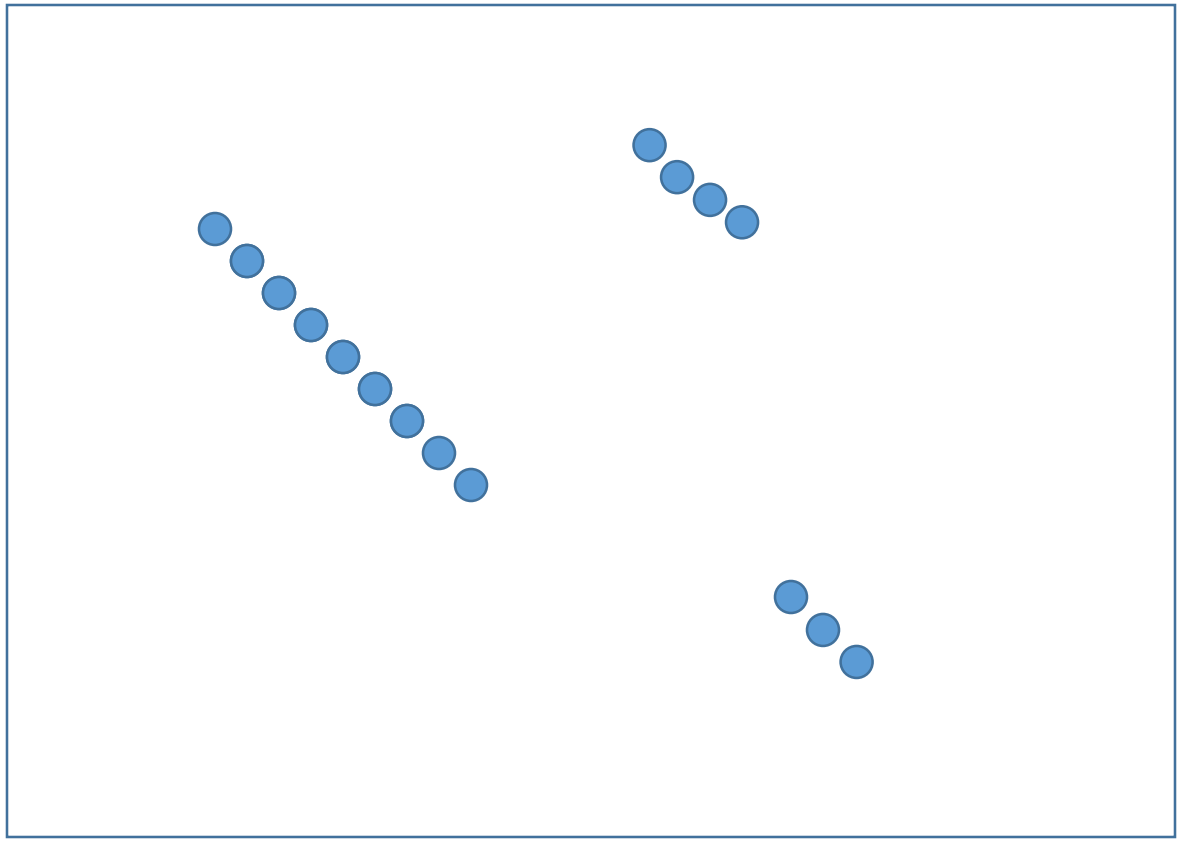


marcamos em vermelho as células onde a base de *s* é igual à base de *t*

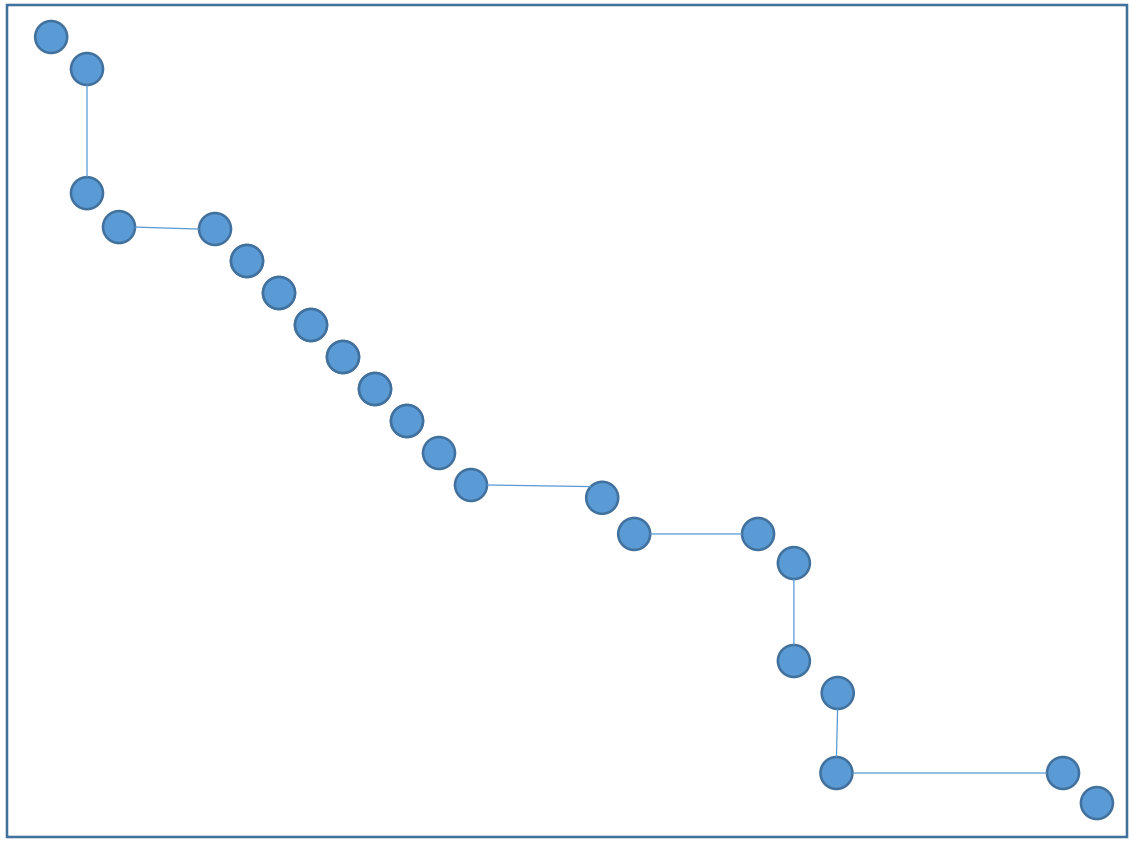
Um alinhamento local pode ser visto como um trecho de células consecutivas num dotplot



Pode haver vários alinhamentos locais “bons”



Um alinhamento global poderia ter esta cara num dotplot



Exercício

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

Resposta do exercício: 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

   Valor ← 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

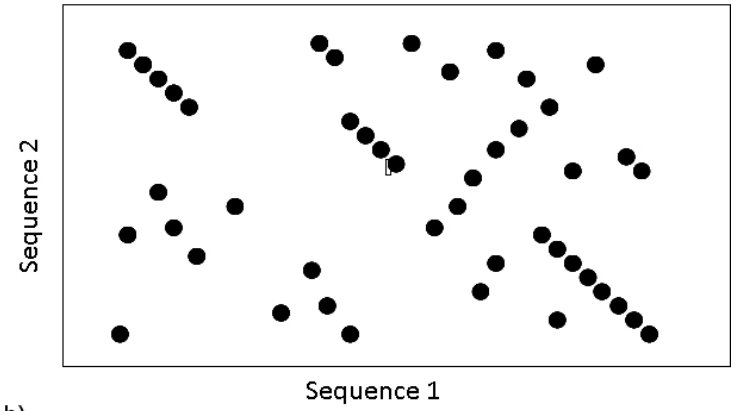
Exercício

- Alinhamentos globais são casos particulares de alinhamentos locais
  - Explique

BLAST é muito mais rápido do que programação dinâmica

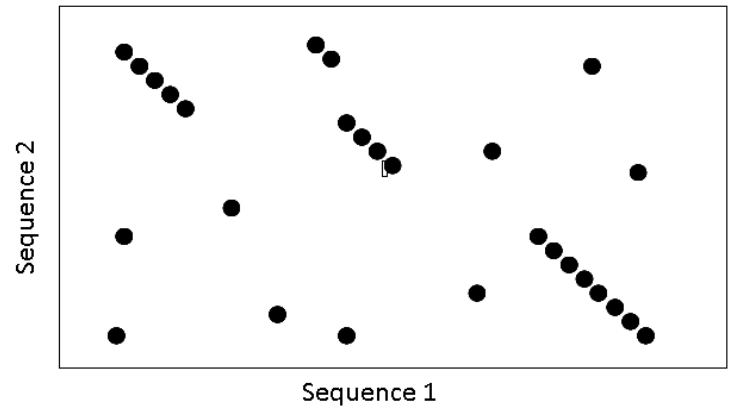
- “Truques de eficiência” usados por BLAST
- BLAST busca trechos parecidos (*palavras* ou *words*) entre as sequências = alinhamentos-semente
- Para nt, esses alinhamentos tem que ser exatos
- Para aa, esses alinhamentos tem que ter nota positiva
- Estende esses alinhamentos-semente até um ponto em que qualquer extensão adicional vai diminuir a nota

a)



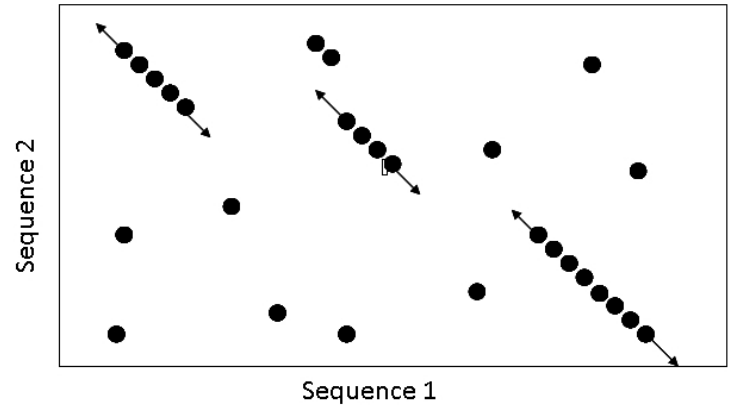
Todos os alinhamentos locais

b)



Remover alinhamentos com nota abaixo de um certo limiar

c)



Estender os alinhamentos locais de melhor nota

Crédito: I. Lobo

Exemplo de alinhamento ótimo que BLAST não encontraria

- Suponha nt e tamanho mínimo de palavra = 4

```
GTG-TGGCCTA-GAAGCT
|||  |  |  |||  ||  |  |
GTGGTCG-CTACGACG-T
```

16 posições, 12 identidades (75%)

Muito da eficiência de BLAST vem de pré-processamento

- As sequências do banco foram pré-processadas para tamanhos pré-definidos de palavras
- Ou seja, para tamanho de palavra = 11nt, já sabemos de antemão quais sequências do banco contém quais palavras

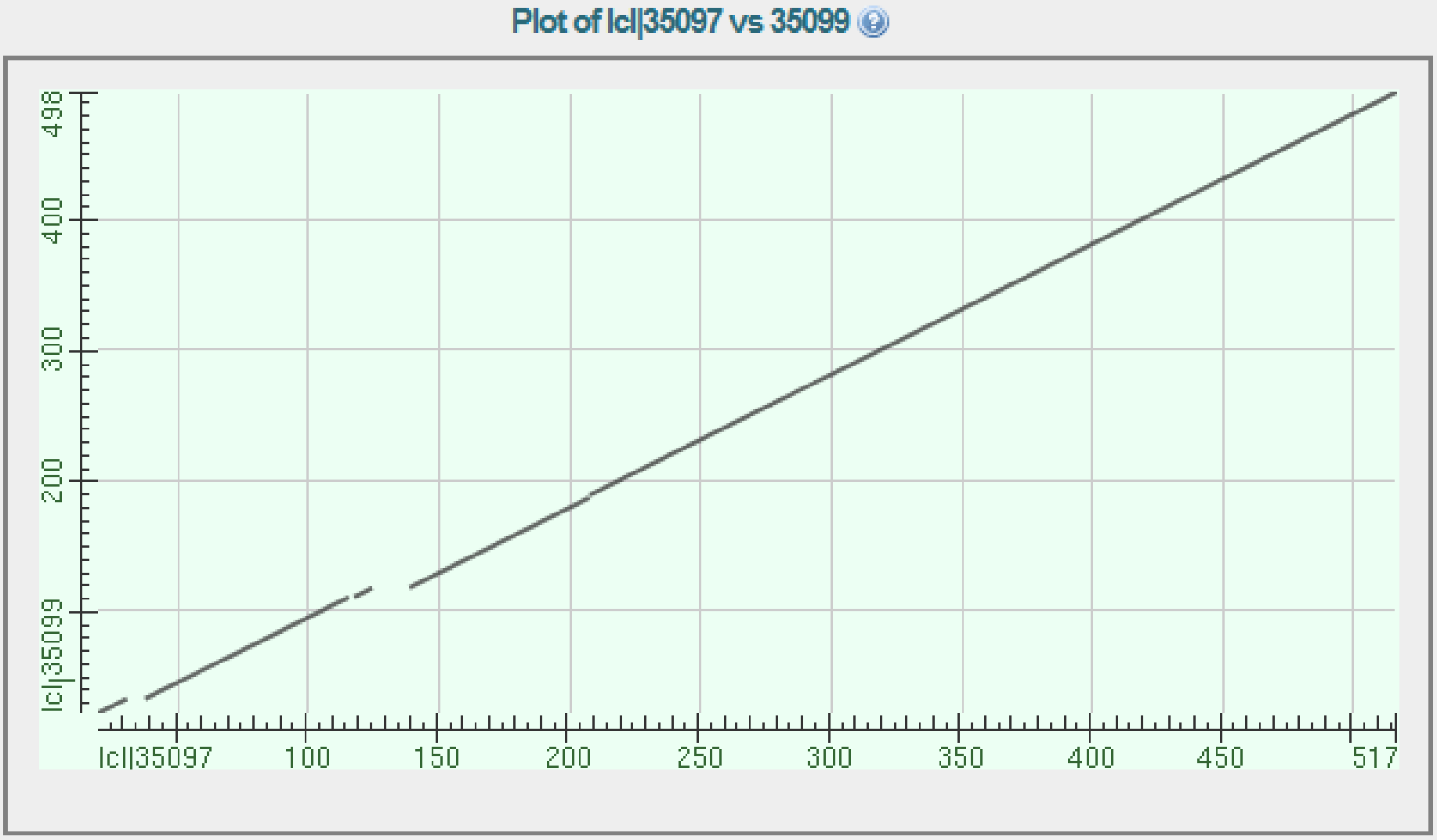
Características de BLAST

- Tamanho default das palavras
  - DNA (blastn): 11 nt
  - Proteínas: 6 aa
- Reporta bit score, raw score, e-value, identidades, positivos, buracos

Um alinhamento encontrado por BLAST

```
>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-------LFKNVSSKLEDQVGSDVYASWFQRLKFRSVSHNIVYLSVPTN  75

Query  81   FLKSWINNRYMDLITSLVQSEDPDVLKVEILVRSASRPVRPAQTEERAQPVQEVGAAPRN  140
            FLK+WI NRY+D IT L Q     +  VEI+VRSA+  + 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  SSAIAHTTAKPPIINTGKISTIQGKQSINPVFGSPLDSKFVFSNFIEGPSNRVALAAAHT  179

Query  201  IAEAGAGA--VRFNPLFIHAGVGLGKTHLLQAIANAAIDSPRNPRVVYLTAEYFMWRFAT  258
            IAE  + +  VRFNPLFIHA VGLGKTHLLQAIANAAI    N RVVYLTAEYFMWRFAT
Sbjct  180  IAEENSSSCTVRFNPLFIHASVGLGKTHLLQAIANAAIKKQNNLRVVYLTAEYFMWRFAT  239

Query  259  AIRDNDALTLKDTLRNIDLLVIDDMQFLQGKMIQHEFCHLLNMLLDSAKQVVVAADRAPW  318
            AIRDN AL  KD LRNIDLL+IDDMQFLQGK+IQHEFCHLLN LLDSAKQ+V AADR P
Sbjct  240  AIRDNYALNFKDCLRNIDLLLIDDMQFLQGKLIQHEFCHLLNSLLDSAKQIVAAADRPPS  299

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

Query  379  VTASGRELEGAFNQLMFRRSFEPNLSVDRVDELLSHLVGSGEAKRVRIEDIQRIVARHYN  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  EKDTVIKKEVELLKRLISE  498
```

Este é o dotplot do alinhamento do slide anterior. Observe que o início do alinhamento está no canto esquerdo inferior e o final está no canto direito superior, ao contrário dos dotplots anteriores que vimos, em que o início do alinhamento estava no canto esquerdo superior e o final no canto direito inferior. Esta diferença é apenas uma questão de convenção.



Resultado de uma busca com BLAST no GenBank

t (499 letters)

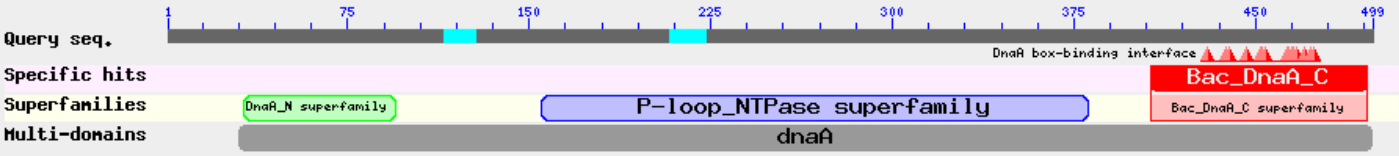
Query ID	lcl|78035	**Database Name**	nr
Description	t	**Description**	All non-redundant GenBank CDS translations+PDB+SwissProt+ environmental samples from WGS projects
Molecule type	amino acid	**Program**	BLASTP 2.2.26+ ▷Citation
Query Length	499		

Other reports: ▷Search Summary [Taxonomy reports] [Distance tree of results] [Multiple alignment]

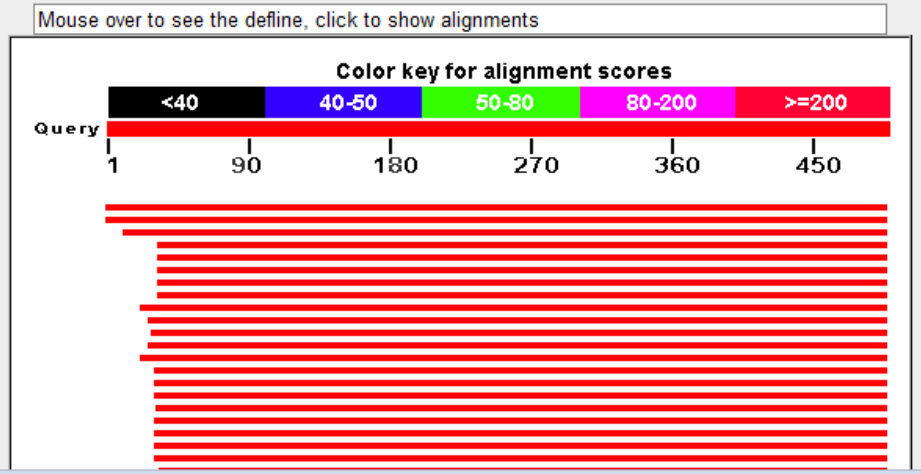
Graphic Summary

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



Lista de hits

Descriptions

Legend for links to other resources: U UniGene E GEO G Gene S Structure M Map Viewer 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|ACE8	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|ABC891	611	611	93%	0.0	63%	G
YP_002279530.1	dnaA gene product [Rhizobium leguminosarum bv. trifolii WSM2304] >gb|ACI5	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 CCNWG	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| chro	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 aquibiodomu	569	569	97%	0.0	56%	

Sabores de BLAST

sequências do banco ⇨ Query 👇	nucleotídeos	aminoácidos
nucleotídeos	BLASTN TBLASTX	BLASTX
aminoácidos	TBLASTN	BLASTP

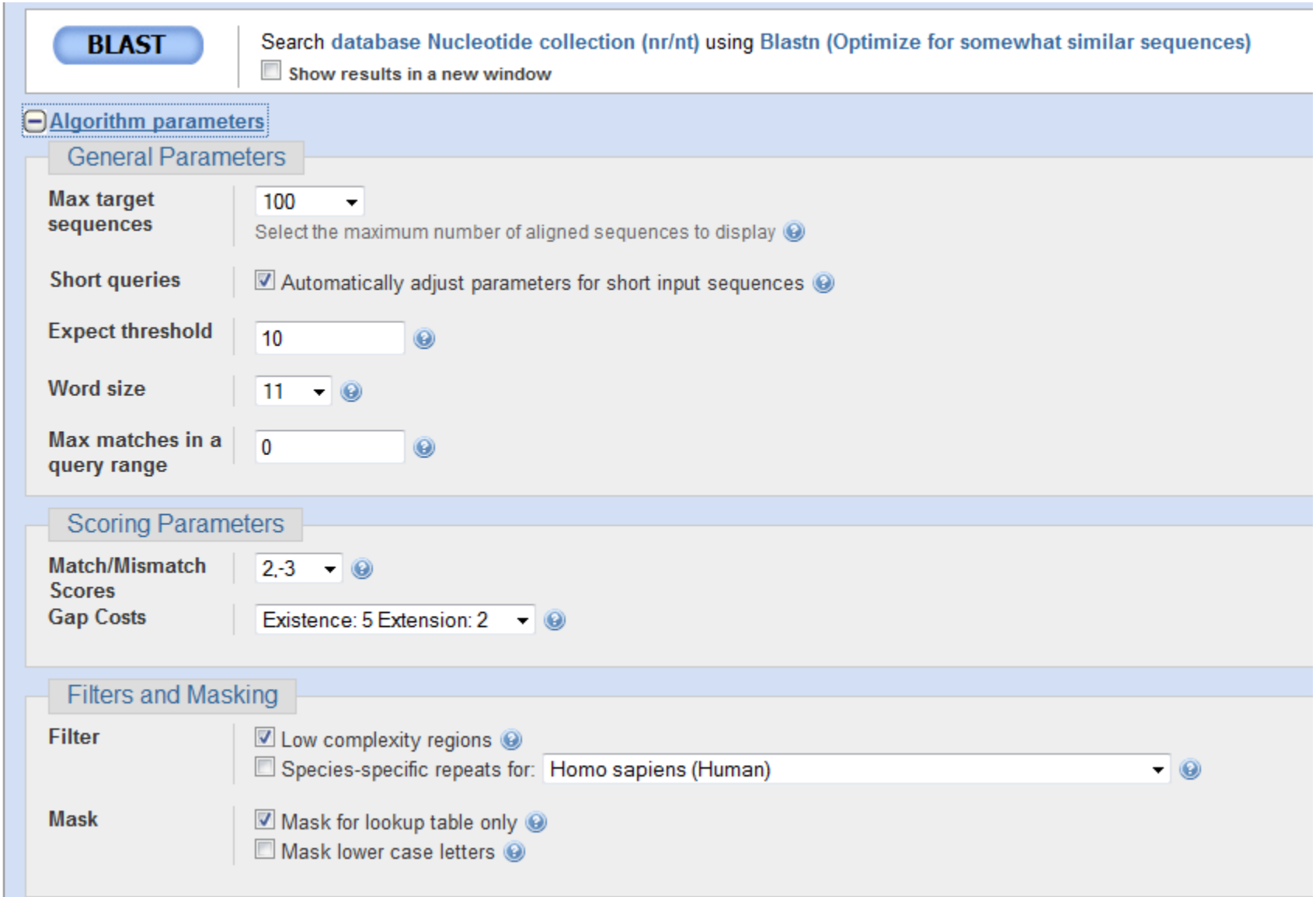
Tradução automática

- BlastX, tblastN, tblastX fazem tradução automática das sequências em todos os quadros de leitura
- Exercício: explique como é essa tradução em cada caso

Outras versões de BLAST

- megablast: apenas nt x nt, ainda mais eficiente do que blastn (portanto pode perder em especificidade – ou seja, deixa de encontrar alinhamentos relevantes)
- psi-blast
- phi-blast
- delta-blast
- smartBLAST
- todas essas são versões mantidas pelo NCBI
- existe também BLAT, que é uma versão de BLAST específica para genoma humano, mantida pela Universidade da California – Santa Cruz
- https://genome.ucsc.edu/cgi-bin/hgBlat

Parâmetros de BLASTn



Regiões de baixa complexidade

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

Nucleic Acids Research, 2019 1
doi: 10.1093/nar/gkz730

Low complexity regions in the proteins of prokaryotes perform important functional roles and are highly conserved

Chrysa Ntountoumi[1], Panayotis Vlastaridis[1], Dimitris Mossialos[2], Constantinos Stathopoulos[3], Ioannis Iliopoulos[4], Vasilios Promonas[5], Stephen G. Oliver[6,*] and Grigoris D. Amoutzias[1,*]

[1]Bioinformatics Laboratory, Department of Biochemistry and Biotechnology, University of Thessaly, 41500, Greece, [2]Microbial Biotechnology-Molecular Bacteriology-Virology Laboratory, Department of Biochemistry and Biotechnology, University of Thessaly, 41500, Greece, [3]Department of Biochemistry, School of Medicine, University of Patras, 26504, Greece, [4]Department of Medicine, University of Crete, Heraklion 71003, Greece, [5]Bioinformatics Research Laboratory, Department of Biological Sciences, New Campus, University of Cyprus, PO Box 20537, CY-1678 Nicosia, Cyprus and [6]Cambridge Systems Biology Centre & Department of Biochemistry, University of Cambridge, CB2 1GA, UK

Received May 15, 2019; Revised July 16, 2019; Editorial Decision August 12, 2019; Accepted August 15, 2019

ABSTRACT

We provide the first high-throughput analysis of the properties and functional role of Low Complexity Regions (LCRs) in more than 1500 prokaryotic and phage proteomes. We observe that, contrary to a widespread belief based on older and sparse data, LCRs actually have a significant, persistent and highly conserved presence and role in many and diverse prokaryotes. Their specific amino acid content is linked to proteins with certain molecular functions, such as the binding of RNA, DNA, metal-ions and polysaccharides. In addition, LCRs have been repeatedly identified in very ancient, and usually highly expressed proteins of the translation machinery. At last, based on the amino acid content enriched in certain categories, we have developed a neural network web server to identify LCRs and accurately predict whether they can bind nucleic acids, metal-ions or are involved in chaperone functions. An evaluation of the tool showed that it is highly accurate for eukaryotic proteins as well.

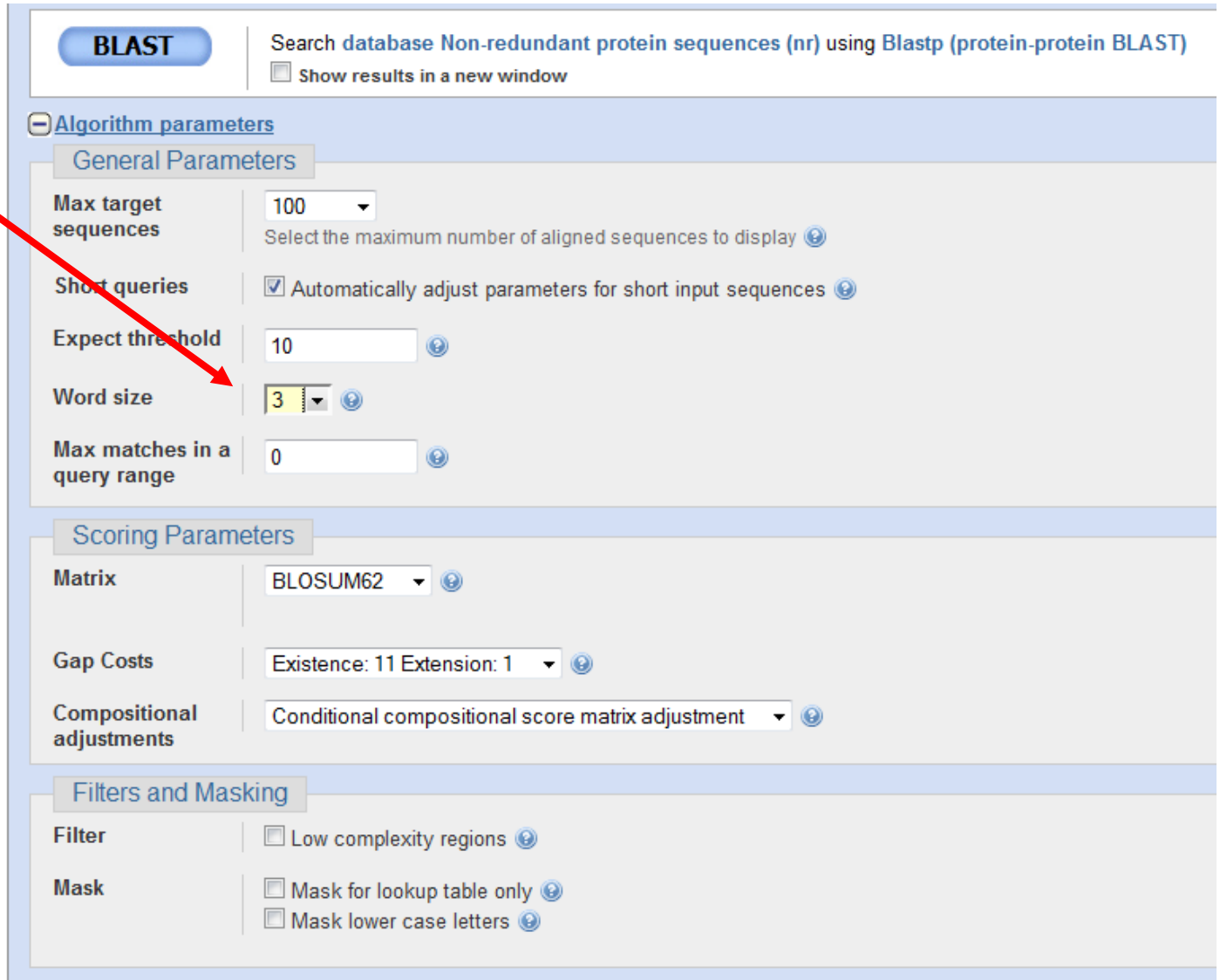
searching for homologs (5). However, emerging experimental evidence increasingly indicated that LCRs may play important adaptive and conserved roles that are highly relevant to biotechnology, heterologous protein expression, medicine, as well as to our understanding of protein evolution (6–9).

One of the established methodologies to identify LCRs is by measuring their Shannon entropy (1,10). The lower the value of the calculated entropy, the more homogeneous the region is in terms of amino acid content. Several computational tools have been developed to detect LCRs (11–17) (see especially (17) for a very recent and extended review on this topic). A subset of LCRs are single amino acid repeats (SARs) or tandem or interspersed repeats of short period (2–5 amino acids). Furthermore, repetition of large amino acid segments of >10 amino acids, or even motif or domain repeats are also categorized as protein repeats (9,18,19), but are not the subject of the current study.

The LCRs of eukaryotic proteins have been the focus of many past and recent studies, due to their involvement in human diseases (20), especially neurodegenerative ones. For example, hydrophobic LCRs tend to form amyloids in humans and other eukaryotes (21). From a mechanistic point of view, LCRs were originally proposed to be unstructured

Parâmetros de BLASTp

agora é 6



Notícia recente!

New BLAST Default Parameters and Search Limits.

Thu, 10 Sep 2020 12:00:00 EST

To provide a more useful BLAST output, maximize performance, and to make search time more consistent, webBLAST is updating some of the default parameters and search limits.

As previously announced, we have just made the following changes:

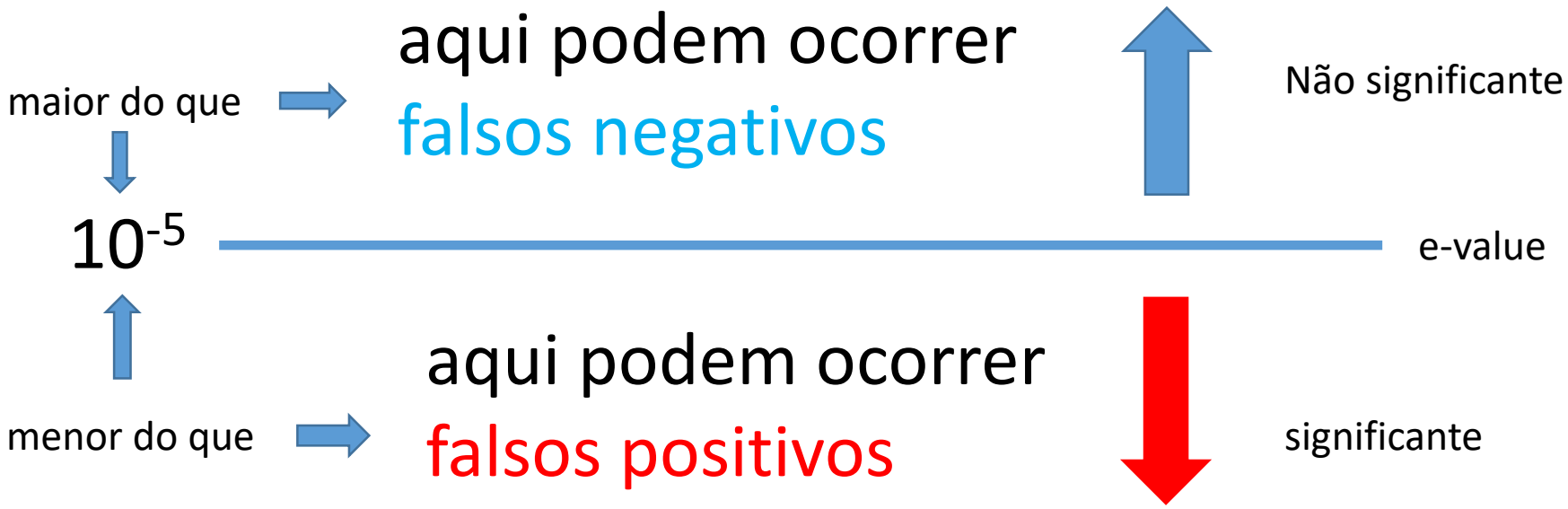
1. Expect Value default will be 0.05
2. Max target sequences limit will be no more than 5,000
3. Max query sequence size for BLASTn, blastx and tblastx and will be 1,000,000
4. Max query sequence size for BLASTp and tblastn will be 100,000
5. Max query/subject sequence size for blast2Sequences mode will be 10,000,000

If you have any questions or concerns, please email us at blast-help@ncbi.nlm.nih.gov

Exercício

- Para ganhar prática com BLAST, faça buscas usando blastp usando nomes de pessoas de diversos tamanhos
- Exemplos
  - ELVIS
  - ALICE
  - REGINA
  - ARISTIDES
  - PEDRALVARES
- Suas palavras tem que usar apenas as letras dos 20 aminoácidos
  - não use O, U, B, J, X, Z
- Altere parâmetros de busca para ver como mudam os resultados (e-value, penalizações de buracos)

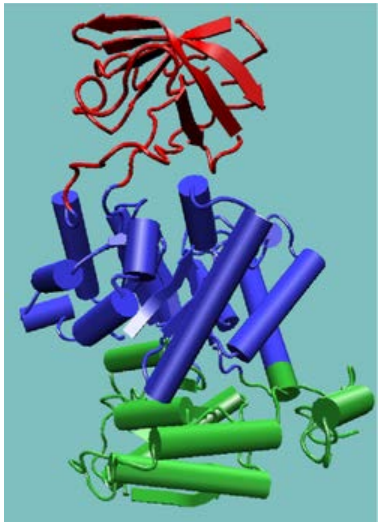
Posso acreditar nos resultados do BLAST?



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)

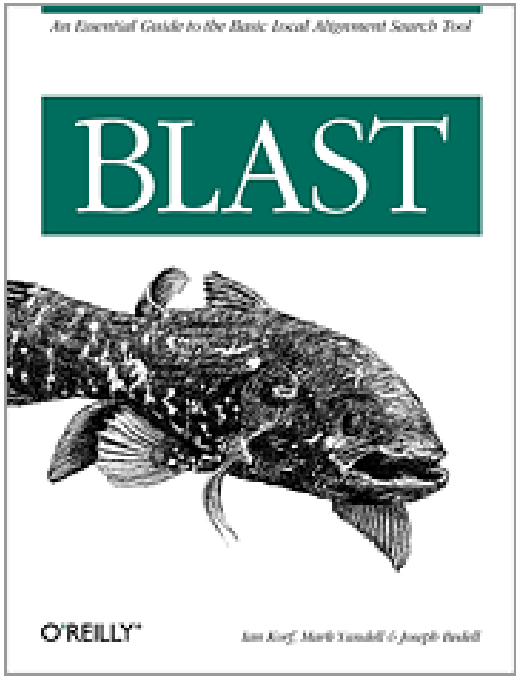
- Duas proteínas podem compartilhar um domínio e não serem relacionadas
  - falsos positivos de BLAST
- Acontece quando as proteínas tem múltiplos domínios



Pyruvate kinase

http://en.wikipedia.org/wiki/File:1pkn.png

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

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

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!