



Universidade de São Paulo
Instituto de Química

Bioinformática

Prof. João Carlos Setubal

Alinhamento de sequências de DNA

Dadas duas sequências de DNA (suas representações como cadeias de letras), queremos **compará-las**. Isto se faz **alinhando** uma com a outra, conforme ilustrado abaixo

```
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-CCTTCGAAGGGT

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

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
 - nome genérico para inserções/remoções: **indels**

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

Cepas de SARS-CoV-2

cepa Wuhan-Hu-1

28851 GTTCAAGAAATTCAACTCCAGGCAGCAGTAGGGGAACTTCTCCTGCTAGA 28900
vv iiv

28851 GTTCAAGAAATTCAACTCCAGGCAGCTCTAAACGAACTTCTCCTGCTAGA 28900

cepa P1 (ou gama)

Comparação de sequências de aminoácidos

- Cadeias de DNA podem codificar proteínas
- Proteínas são sequências de aminoácidos
- Existem 20 aminoácidos que são usados em proteínas de seres vivos
- são abreviados por 20 letras do alfabeto ou por abreviações de 3 letras
- Vamos considerar agora a 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

O alinhamento acima tem 50 posições. Em 28 posições o aminoácido de cima é igual ao aminoácido de baixo. Portanto **% de identidade** = $28/50 = 0,56 = 56\%$

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

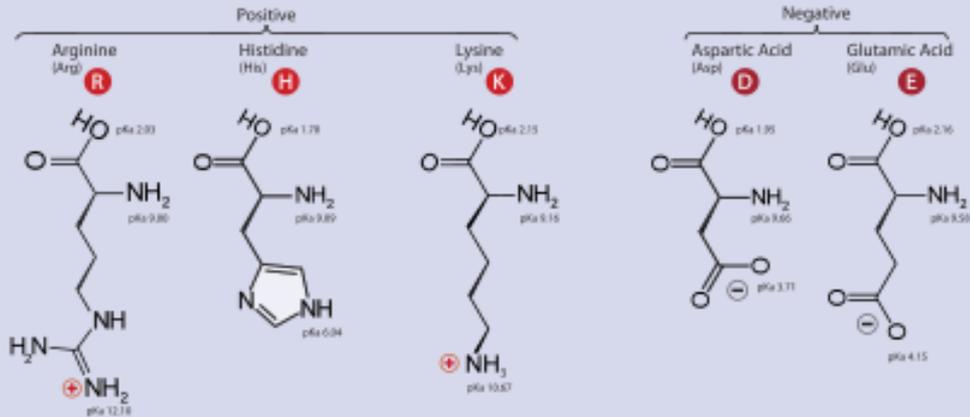
Twenty-One Amino Acids

⊕ Positive

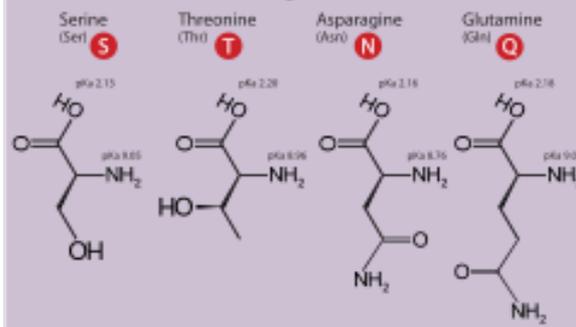
⊖ Negative

* Side chain charge at physiological pH 7.4

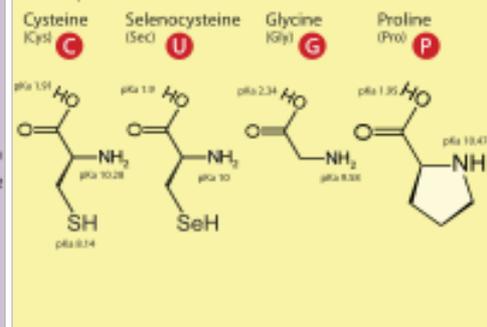
A. Amino Acids with Electrically Charged Side Chains



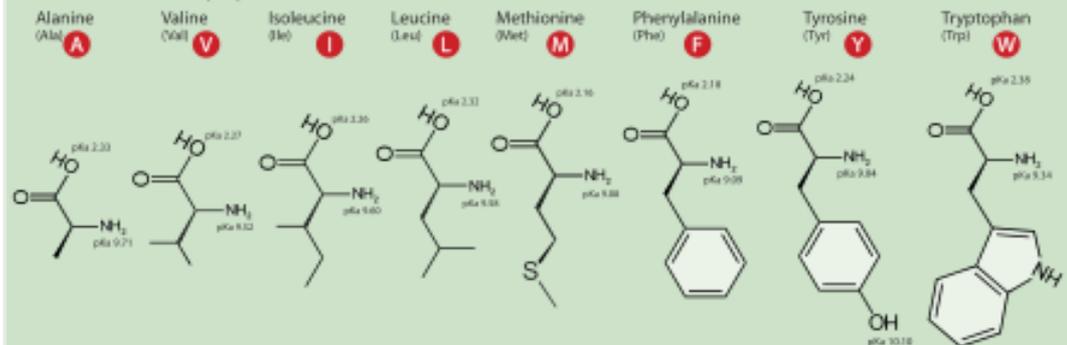
B. Amino Acids with Polar Uncharged Side Chains



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.ii
# * 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: GDVEKGKKIFIMKCSQCHTVEKGGKHKKTGPNLHGLFGRKTGQAPGYSYTAANKNKGI IWG
GD EKGKK++ +C QCH V+      KTGP LHG+ GR +G    G+ Y+AANKNKG++W
C: GDYKGGKKVYKQRCLQCHVVDSTAT-KTGPTLHGVI GRTSGTVSGFDYSAANKNKGVVWT
```

Exercício

- Usando a matriz BLOSUM62 de slide anterior, diga qual é a pontuação dos seguintes pareamentos:
 - E com D
 - W com P
 - G com S
 - L com L
 - W com W

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](#)

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...](#)

Comparação de sequências

- Similaridade “suficiente”
- O que é similaridade?
- O que é “suficiente”?
- Google das sequências: BLAST
 - Basic Local Alignment Search Tool
 - Altschul et al., 1990, 1997

blastn

blastp

blastx

tblastn

tblastx

Enter Query Sequence

BLASTP programs search protein subjects using a protein query. [more..](#)

Enter accession number(s), gi(s), or FASTA sequence(s) ?

```
>s
MQLNLAMGAVADGDRAPKACDAACSEAAGDKSAMMHDALFERFSARLKAQVGPVYASWFA
RLKHLTVSKSVVRFTVPTIFLKSWINNRYMDLITSLVQSEDDPDLKVEILVRSASRPVSPA
QTEERAQPVQEVGAAPRNKSFIPQSATAPAAQPMAAQATLRQGGSGPLFGSPLDTRFTFD
TFVEGSSNRVALAAAKTIAEAGAGAVRFNPLFIHAGVGLKTHLLQAIANAALDSRRNPRV
VYLTAEFMWRFATAIRDNDALTLKDTLRNIDLLVIDDMQFLQGMIOHEFCHLLNMLLDS
```

Clear

Query subrange ?

From

To

Or, upload file

Browse...

Job Title

s

Enter a descriptive title for your BLAST search ?

 Align two or more sequences ?

Enter Subject Sequence

Enter accession number, gi, or FASTA sequence ?

Clear

Subject subrange ?

```
>t
MBSRGISACIQENNYETPETNADARCLETTICEELFKNVSSKLEDQVGSVDVYASWFQRLKFR
SVSHNIVYLSVPTNFKAWIKNRYIDTITKLFQESISSIQEVEIIVRSAALMPSETSSSSA
IAHTTAKPRIINTGKISTIQGKOSINRVFGSPIDSKFVFSNFIQSPNSVALAAHTIAEE
NSSSCTVRFNPLFIHASVGLKTHLLQAIANAALKKONLRVVYLTAEYFMWRFATAIRDN
YALNFKDCLRNIIDLLLIDDMQFLOCKLIQHEFCHLLNSLLDSAKQIVAAADRPPSELESLD
```

From

To

Or, upload file

Browse...

Program Selection

Algorithm

 blastp (protein-protein BLAST)

Choose a BLAST algorithm ?

>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 | DAACSEAAGDKSAMHDALFERFSARLKAQVGPEVYASWFARLKLHTVSKSVVRFTVPT | 80 |
| | | DA C E ++ LF+ S++L+ QVG +VYASWF RLK +VS ++V +VPT | |
| Sbjct | 23 | DARCLETTCEE-----LFKNVSSKLEDQVGSVDVYASWFQRLKFRSVSHNIVYLSVPTN | 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 | SSAIAHTTAKPPIINTGKISTIQGKQSINPVFGSPLDSKFVFSNFIEGPSNRVALAAAHT | 179 |
| Query | 201 | IAEAGAGA--VRFNPLFIHAGVGLGKTHLLQAIANAIDS PRNPRVVYLTAEYFMWRFAT | 258 |
| | | IAE + + VRFNPLFIHA VGLGKTHLLQAIANAAI N RVVYLTAEYFMWRFAT | |
| Sbjct | 180 | IAEENSSSCTVRFNPLFIHASVGLGKTHLLQAIANAIAIKKQNNLRVVYLTAEYFMWRFAT | 239 |
| Query | 259 | AIRDNDALTLKDTLRNIDLLVIDDMQFLQGKMIQHEFCHLLNMLLDSAKQVVVAADRAPW | 318 |
| | | AIRDN AL KD LRNIDLL+IDDMQFLQGK+IQHEFCHLLN LLSAKQ+V AADR P | |
| Sbjct | 240 | AIRDNYALNFKDCLRNIDLLLIDDMQFLQGKLIQHEFCHLLNSLLDSAKQIVAAADRPPS | 299 |
| Query | 319 | ELESLDPRVRSRLQGGMAIEIEGPDYDMRYEMLNRRRMSARQDDPSFEISDEILTHVAKS | 378 |
| | | ELESLD R+RSRLQGG+A+ + D +MR +L R+ A++D+P IS+EIL VA++ | |
| Sbjct | 300 | ELESLDSRIRSRLQGGVAVPLGAHDIEMRLTILKNRLKMAKKDNPPLYISEEILQRVAQT | 359 |
| Query | 379 | VIASGRELEGAFNQLMFRRSFEPNLSVDRVDELLSHLVGSGEAKRVRIEDIQRIVARHYN | 438 |
| | | VI SGREL+GAFNQL+FR SFEP L++ VDELLSHLV +GE K++RIEDIQR+V++HYN | |
| Sbjct | 360 | VITSGRELDGAFNQLVFRNSFEPVLTIKMVEDELLSHLVSAGETKKIRIEDIQRMVSKHYN | 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 | |

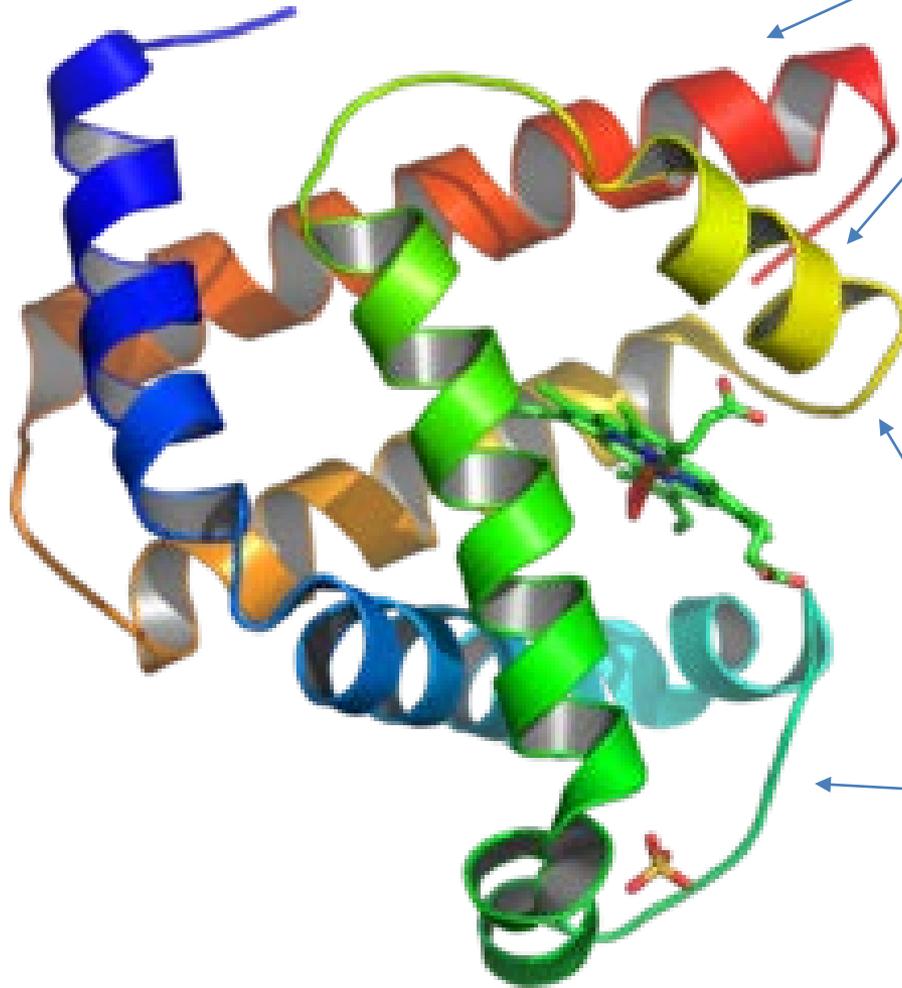
Respostas das perguntas no caso de BLAST

- O que é similaridade?
 - É medida de 3 formas
 - nota (score)
 - e-value
 - % de identidade
- O que é “suficiente”?
 - Depende do objetivo da comparação
 - No caso do e-value, um valor **máximo** em geral adotado é 10^{-5}
- Consulte a descrição do projeto para mais detalhes

No alinhamento apresentado

- algumas regiões tem mais similaridade do que outras
- isto significa que nas proteínas correspondentes há trechos ou regiões que são mais *conservadas* (entre proteínas de diferentes organismos) do que outras

regiões conservadas

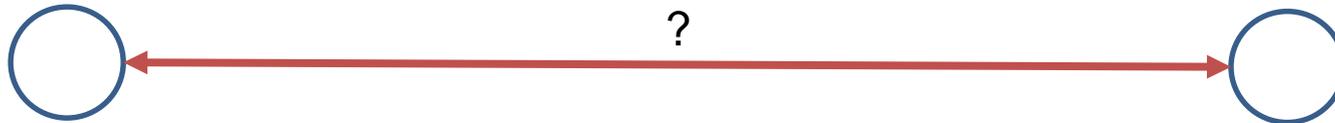


regiões variáveis

Esclarecimento de um **crime** por identificação de DNA

DNA da cena do crime

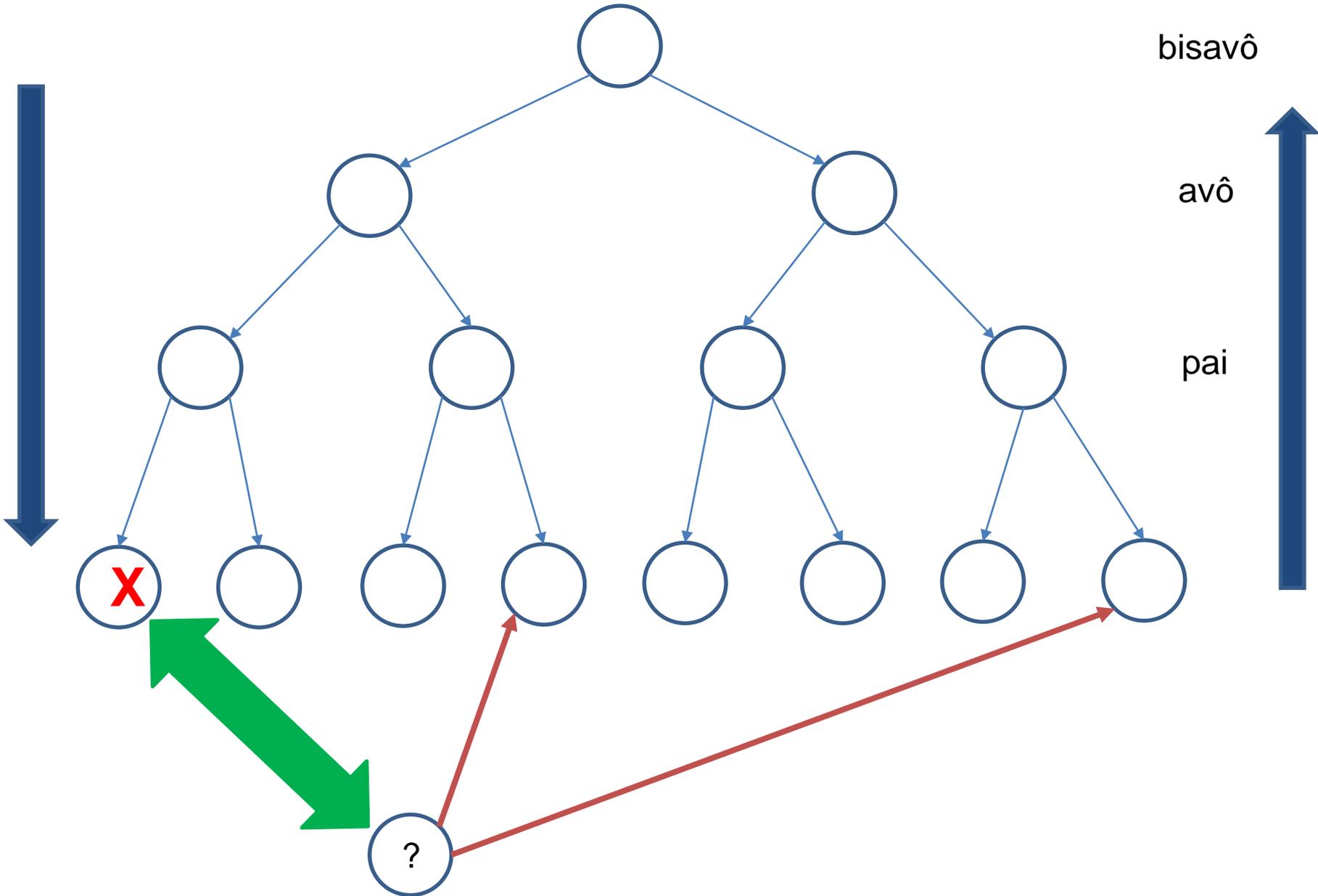
DNA do suspeito



E se não há suspeito?

O caso do assassino de Sacramento (*Golden State killer*)

- Um homem matou cerca de 12 pessoas nas décadas de 70 e 80 na região de Sacramento (California)
- Não havia nenhuma pista sobre o criminoso
- mas (na época) seu DNA foi coletado



bisavô

avô

pai

operações

- coletar amostra contendo DNA de dono desconhecido
- comparar esse DNA com DNAs de um banco de genomas
 - resultado: hits
- descobrir ancestral comum mais próximo (ACMP) dos hits [genealogia]
- descobrir descendentes do ACMP [genealogia]
- refinar a lista com base em metadados [“cadastros”]
 - resultado: alguns poucos candidatos
- pegar amostra “pública” do DNA dos candidatos
- comparar com amostra original

Qual é o perigo?

- Privacidade do nosso DNA e de nossos metadados

Exercício de bioinformática

- Disponível em

<http://www.iq.usp.br/setubal/qbq107/2021/>

- Recomendo que façam, mas não será corrigido