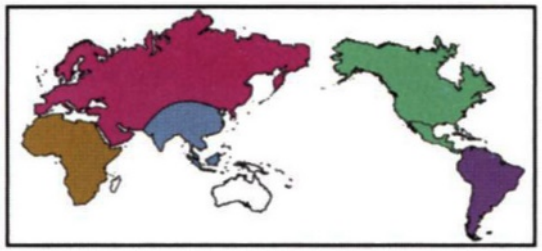
Bioinformática

Prof. João Carlos Setubal

Universidade de São Paulo
Instituto de Química

Evolução



Zoogeographical Regions

Blue: Oriental
Purple: Palearctic
Gold: Ethiopian (African)
Violet: Neotropical
Green: Nearctic

(6.2MY) 7
(6.7MY) 6
(7.2MY) 5*
(8.0MY) 4
(8.5MY) 3
(9.4MY) 2
(10.8MY) 1

Fca:	*Felis catus*	Domestic cat	Domestic Cat Lineage
Fsi	*F. silvestris*	European wild cat	
Fli:	*F. libyca*	African wild cat	
Fbi:	*F. bieti*	Chinese desert cat	
Fma:	*F. margarita*	Desert cat	
Fni:	*F. nigripes*	Black-footed cat	
Fch:	*F. chaus*	Jungle cat	
Oma:	*Otocolobus manul*	Pallas cat	Leopard Cat Lineage
Pru:	*Prionailurus rubiginosus*	Rusty spotted cat	
Pbe:	*P. bengalensis*	Asian leopard cat	
Pvi:	*P. viverrinus*	Fishing cat	
Ppl:	*P. planiceps*	Flat-headed cat	
Pco:	*Puma concolor*	Puma	Puma Lineage
Pya:	*P. yagouaroundi*	Jaguarundi	
Aju:	*Acinonyx jubatus*	Cheetah	
Lyp:	*Lynx pardinus*	Iberian lynx	Lynx Lineage
Lly:	*L. lynx*	Eurasian lynx	
Lca:	*L. canadensis*	Canada lynx	
Lru:	*L. rufus*	Bobcat	
Lpa:	*Leopardus pardalis*	Ocelot	Ocelot Lineage
Lwi:	*L. wiedii*	Margay	
Lja:	*L. jacobita*	Andean mt. cat	
Lco:	*L. colocolo*	Pampas cat	
Lge:	*L. geoffroyi*	Geoffroy's cat	
Lgu:	*L. guigna*	Kodkod	
Lti:	*L. tigrinus*	Tigrina	
Cca:	*Caracal caracal*	Caracal	Caracal Lineage
Cau:	*C. aurata*	African golden cat	
Cse:	*C. serval*	Serval	
Pba:	*Pardofelis badia*	Bay cat	Bay Cat Lineage
Pte:	*P. temminckii*	Asian golden cat	
Pma:	*P. marmorata*	Marbled cat	
Ple:	*Panthera leo*	Lion	Panthera Lineage
Pon:	*P. onca*	Jaguar	
Ppa:	*P. pardus*	Leopard	
Pti:	*P. tigris*	Tiger	
Pun:	*P. uncia*	Snow leopard	
Nne:	*Neofelis nebulosa*	Clouded leopard	
	Prionodon linsang	Banded linsang	
	Crocuta crocuta	Spotted hyaena	
	Helogale parvula	Dwarf mongoose	
	Suricata suricatta	Suricate	
	Cryptoprocta ferox	Fossa	
	Paradoxurus hermaphroditus	Palm Civet	
	Gennetta genetta	Genet	

0.005 substitutions/site

Na base da evolução dos organismos está a evolução dos genes

genes também tem “parentesco”

- podemos quantificar o parentesco entre genes fazendo alinhamento entre suas sequências

- mas o alinhamento deve ser feito sempre entre genes homólogos

Alinhamento de DNA

GTGGTGGCCTACGAAGGT
GTAGTGCCTTCGAAGGGT

Como avaliar um alinhamento?

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

Pontuação do alinhamento

GTGGTGGCCTACGAAGGT

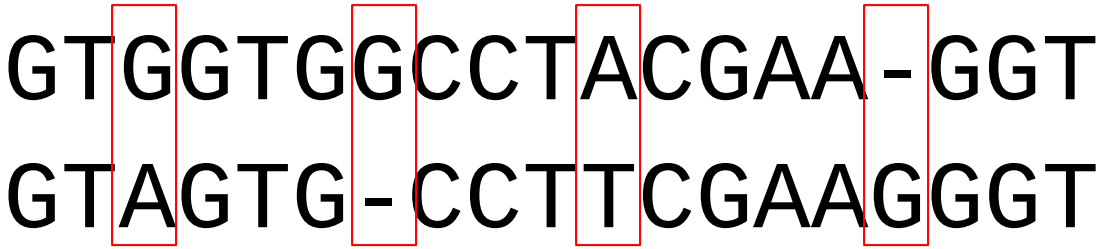
GTAGTGCCTTCGAAGGGT

+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



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

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:     IIWGEDTLMEYLENPKKYIPGTKMIFVGIKKKEERADLIAYLKKATNE
C:     VVWTKETLFEYLLNPKKYIPGTKMVFAGLKKADERADLIKYIEVESAKSL
         *   ** *** ********** * * **  ****** *
```

% 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

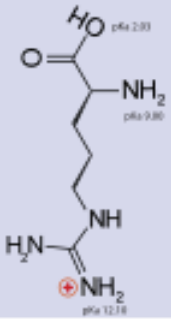
Positive Negative

• Side chain charge at physiological pH 7.4

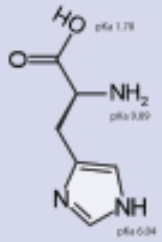
A. Amino Acids with Electrically Charged Side Chains

Positive

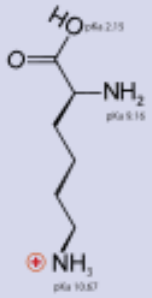
Arginine (Arg) R



Histidine (His) H

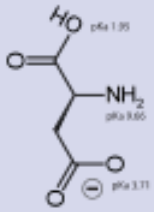


Lysine (Lys) K

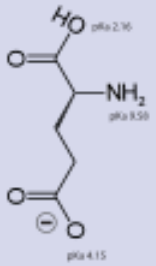


Negative

Aspartic Acid (Asp) D

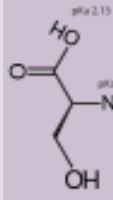


Glutamic Acid (Glu) E

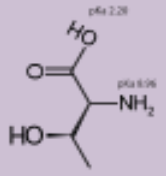


B. Amino Acids with Polar Uncharged Side Chains

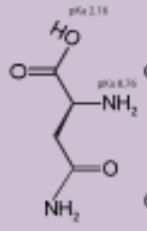
Serine (Ser) S



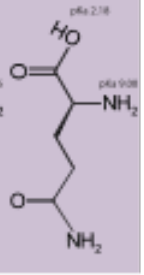
Threonine (Thr) T



Asparagine (Asn) N

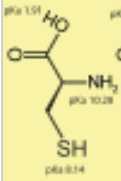


Glutamine (Gln) Q

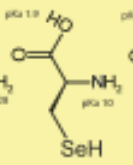


C. Special Cases

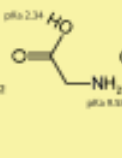
Cysteine (Cys) C



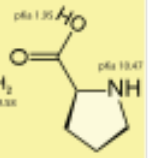
Selenocysteine (Sec) U



Glycine (Gly) G

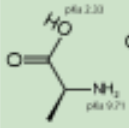


Proline (Pro) P

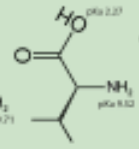


D. Amino Acids with Hydrophobic Side Chain

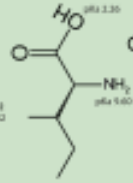
Alanine (Ala) A



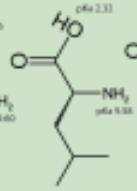
Valine (Val) V



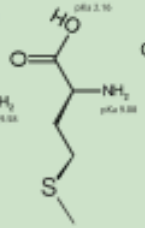
Isoleucine (Ile) I



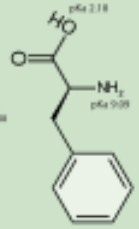
Leucine (Leu) L



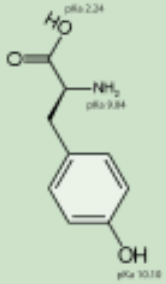
Methionine (Met) M



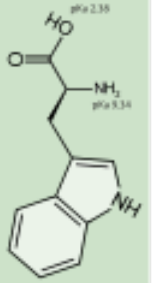
Phenylalanine (Phe) F



Tyrosine (Tyr) Y



Tryptophan (Trp) W



pKa Data: **CRC Handbook of Chemistry**, v. 2010

Dan Cojocari, Department of Medical Biophysics, **University of Toronto** 2011

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.iij
#  * 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 -4  1
```

Fonte: NCBI

Pontuação leva em conta a matriz

- Match: blosum62(i,i) sempre positivo
- Mismatch: blosum62(i,j) positivo, nulo, negativo
- Espaço: –2

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

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

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...

Comparação de sequencias

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

BLAST®

Basic Local Alignment Search Tool

Home | Recent Results | Saved Strategies | Help

NCBI/ BLAST/ blastp suite

Align Sequences Protein BLAST

blastn | **blastp** | blastx | tblastn | tblastx

BLASTP programs search protein subjects using a protein query. more..

Enter Query Sequence

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

```
>s
MQLNLAMGAVADGDRAPKACDAACSEAAGDKSAMMHDALFERFSARLKAQVGPEVYASWFA
RLKLHTVSKSVVRFTVPTTFLKSWINNRYMDLITSLVQSEDPDVLKVEILVRSASRPVRPA
QTEERAQPVQEVGAAPRNKSFIPSQSATAPAAQPMAAQATLRQGGSGPLFGSPLDTRFTFD
TFVEGSSNRVALAAAKTIAEAGAGAVRFNPLFIHAGVGLGKTHLLQAIANAAIDSPRNPRV
VYLTAEYFMWRFATAIRDNDALTLKDTLRNIDLLVIDDMQFLQGKMIQHEFCHLLNMLLDS
```

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

```
>t
MRSRGISACIQENNYETPETNADARCLETTCEELFKNVSSKLEDQVGSDVYASWFQRLKFR
SVSHNIVYLSVPTNFLKAWIKNRYIDTITKLFQESISSIQGVEIIVRSAAIMPSETSSSSA
IAHTTAKPRIINTGKISTIQGKQSINPVFGSPLDSKFVFSNFIEGPSNRVALAAAAHTIAEE
NSSSCTVRFNPLFIHASVGLGKTHLLQAIANAAIKKQNNLRVVYLTAEYFMWRFATAIRDN
YALNFKDCLRNIDLLLLIDDMQFLQGKLIQHEFCHLLNSLLDSAKQIVAAADRPPSELESLD
```

Subject subrange

From

To

Or, upload file Browse...

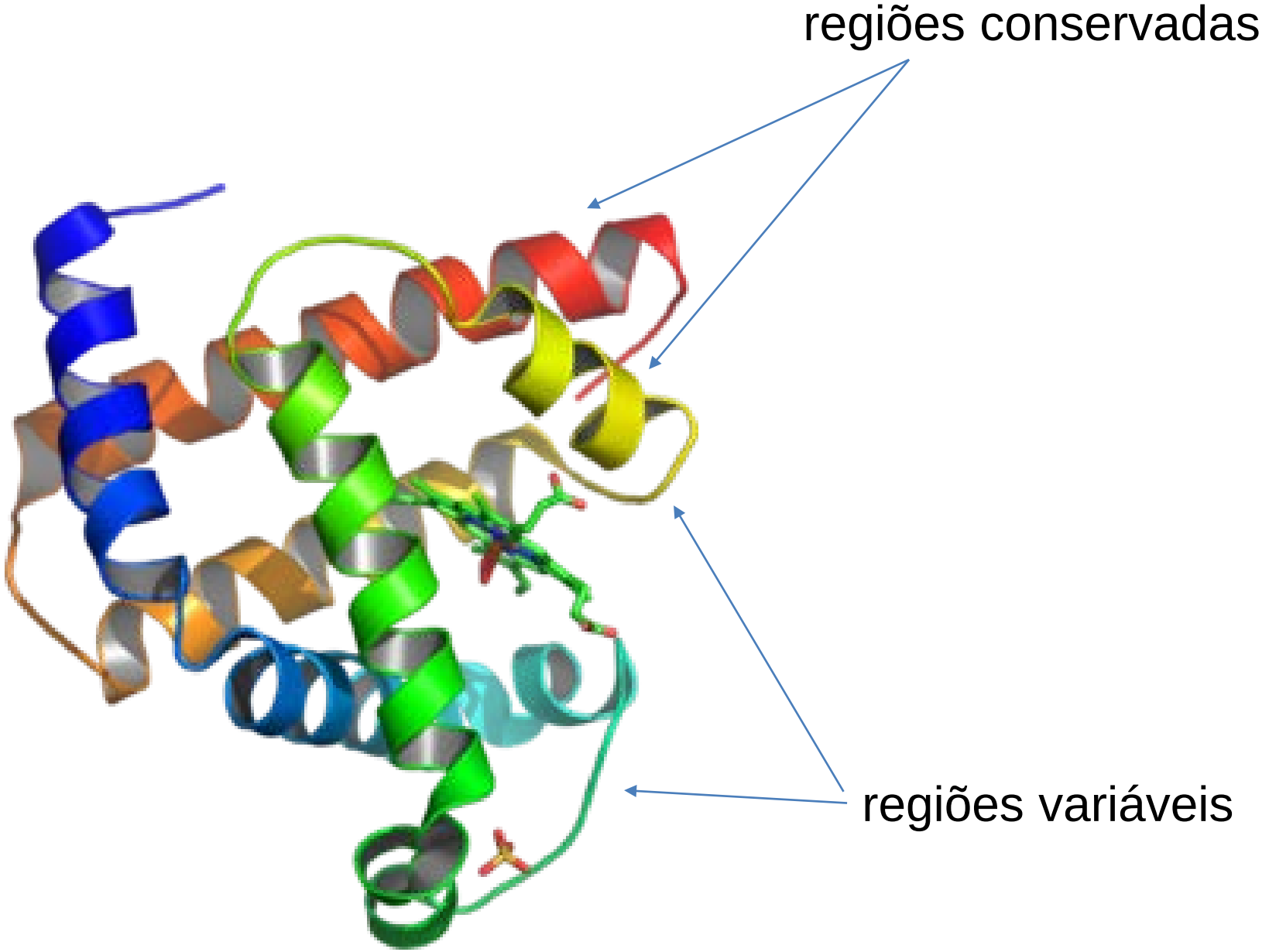
Program Selection

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   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
```


regiões conservadas
regiões variáveis