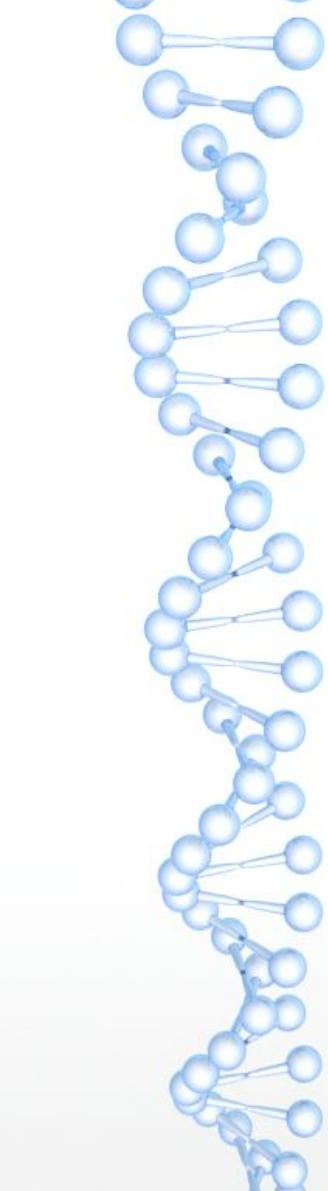
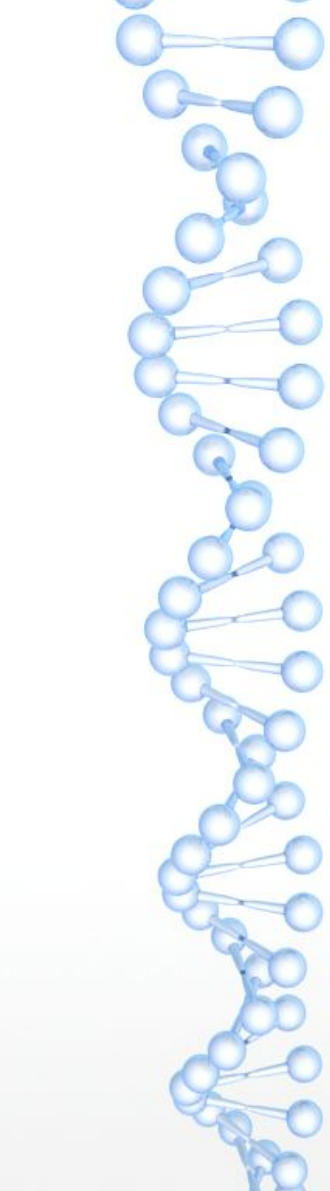


Biological Sequence Analysis

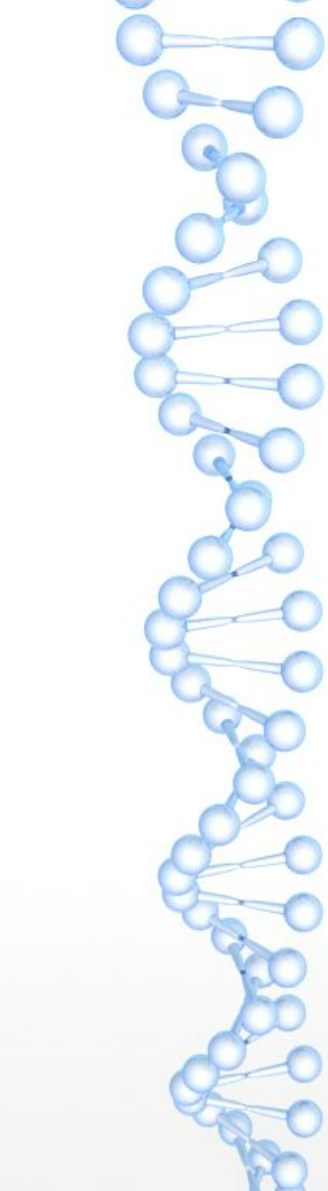
Alberto Pallavicini
Applied genomics

Sequence Alignments: Determining Similarity and Deducing Homology



Why construct sequence alignments?

- Provide a measure of relatedness between nucleotide or amino acid sequences
- Determining relatedness allows one to draw biological inferences regarding
  - structural relationships
  - functional relationships
  - evolutionary relationships
- Important to use correct terminology when describing phylogenetic relationships



Defining the Terms

- The quantitative measure: ***Similarity***
  - Always based on an observable
  - Usually expressed as percent identity
  - Quantify changes that occur as two sequences diverge (substitutions, insertions, or deletions)
  - Identify residues crucial for maintaining a protein's structure or function
- High degrees of sequence similarity *might* imply
  - a common evolutionary history
  - possible commonality in biological function

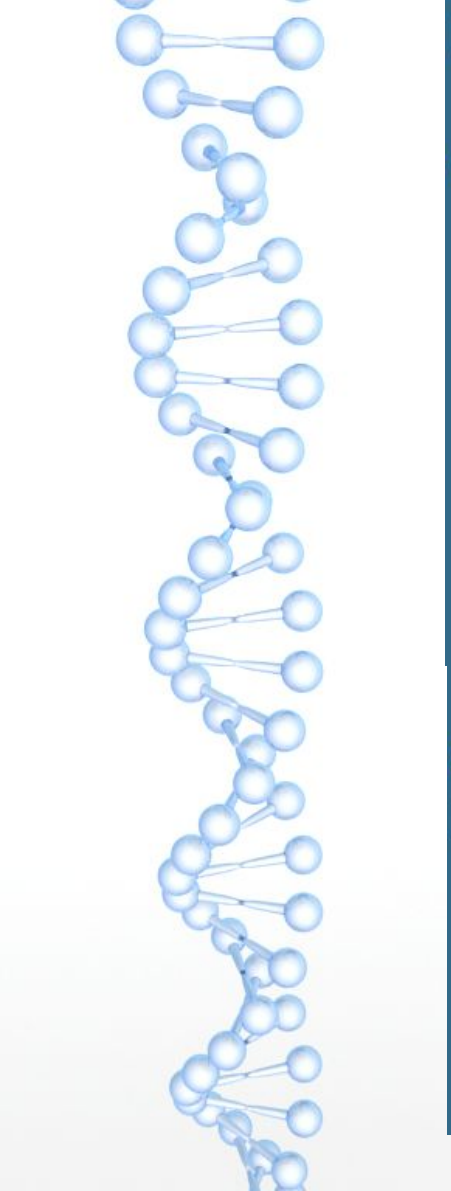
Defining the Terms

The conclusion: ***Homology***

- ***Homology***: Implies an evolutionary relationship
- ***Homologs***: Genes that have arisen from a common ancestor
- Genes either *are* or *are not* homologous
  (not measured in degrees)

> It is worth repeating here that homology, like pregnancy, is indivisible[8]. You either are homologous (pregnant) or you are not. Thus, if what one means to assert is that 80% of the character states are identical one should speak of 80% identity, and not 80% homology.

Fitch, Trends Genet. 16: 227-231, 2000



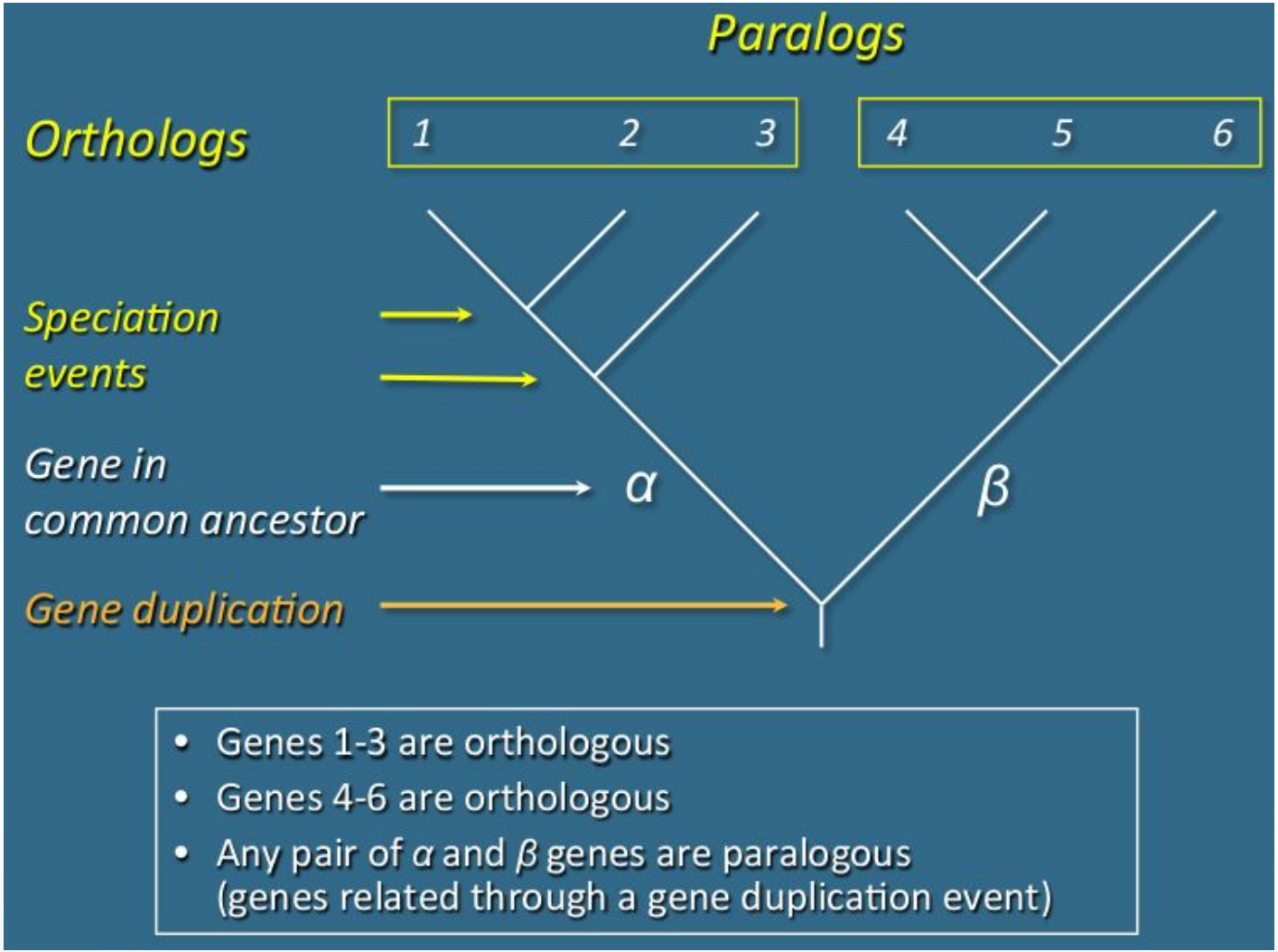
Defining the Terms

Orthologs: Genes that diverged as a result of a speciation event

- Sequences are direct descendants of a sequence in a common ancestor (share a common origin)
- Most likely have similar domain and three-dimensional structure
- Usually retain same biological function over evolutionary time
- Can be used to predict gene function in novel genomes

Paralogs: Genes that arose by the duplication of a single gene in a particular lineage

- Perhaps less likely to perform similar functions
- Can take on new functions over evolutionary time
- Provides insight into 'evolutionary innovation'

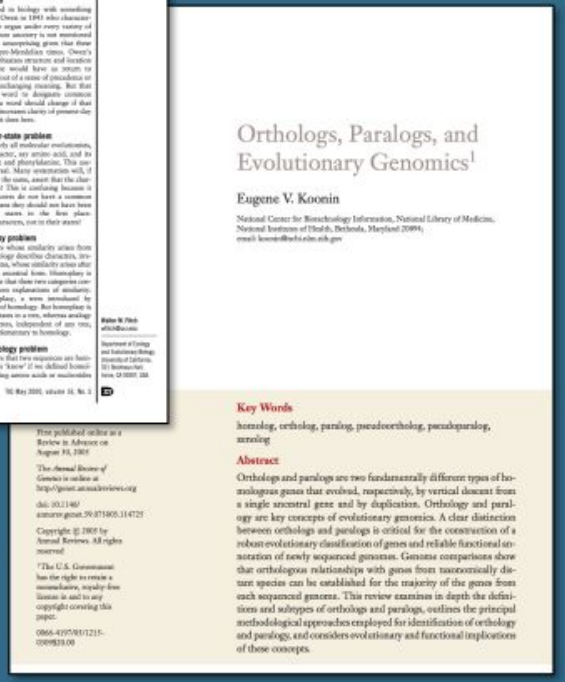

Paralogs
Orthologs
1
2
3
4
5
6
Speciation events
Gene in common ancestor
Gene duplication
α
β
• Genes 1-3 are orthologous
• Genes 4-6 are orthologous
• Any pair of α and β genes are paralogous (genes related through a gene duplication event)

Orthology and Paralogy: Further Reading

Homology

a personal view on some of the problems

Walter Fitch
Trends Genet.
16: 227-231, 2000

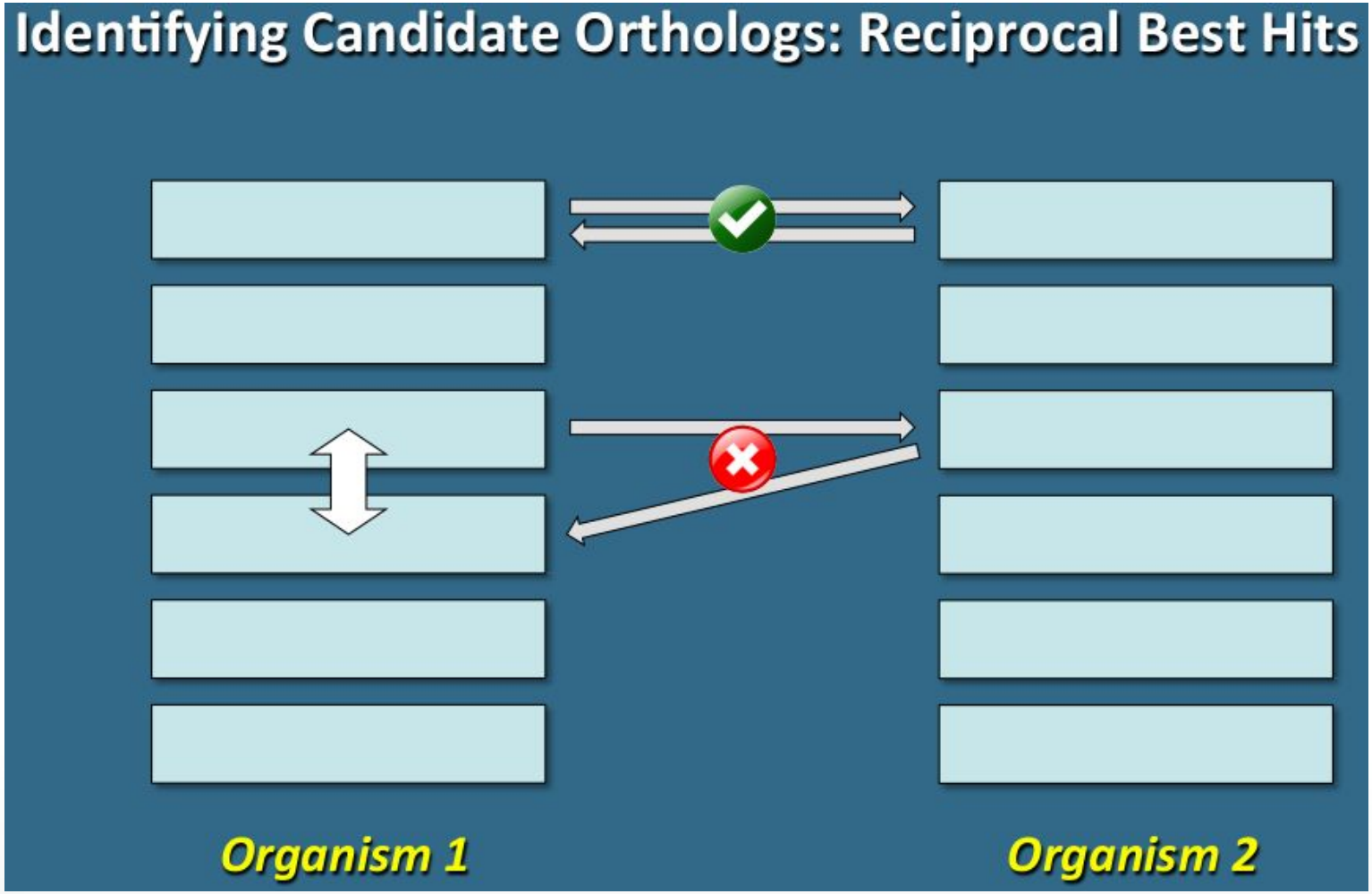


Orthologs, Paralogs, and Evolutionary Genomics[1]

Eugene V. Koonin

Eugene Koonin
Annu. Rev. Genet.
39: 309-338, 2005

Identifying Candidate Orthologs: Reciprocal Best Hits



Global Sequence Alignments

- Sequence comparison along the entire length of the two sequences being aligned
- Best for highly-similar sequences of similar length
- As the degree of sequence similarity declines, global alignment methods tend to miss important biological relationships

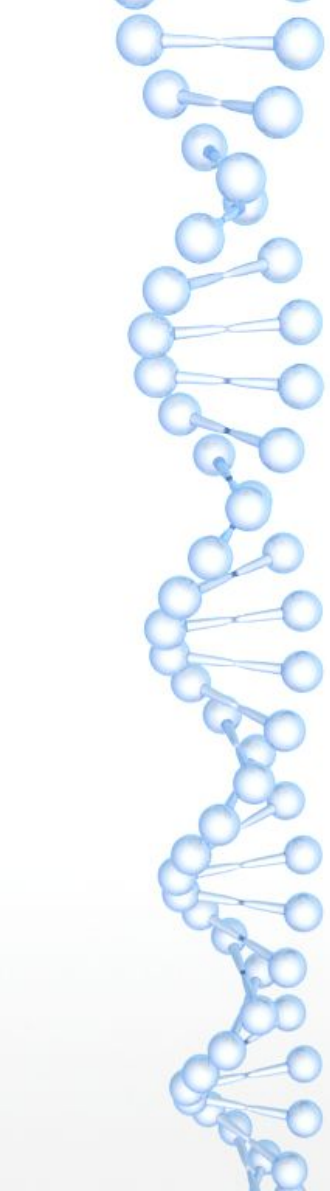
Local Sequence Alignments

- Sequence comparison intended to find the most similar regions in the two sequences being aligned ('paired subsequences')
- Regions outside the area of local alignment are excluded
- More than one local alignment could be generated for any two sequences being compared
- Best for sequences that share some similarity, or for sequences of different lengths

Scoring Matrices: Construction and Proper Selection

Scoring Matrices

- Empirical weighting scheme representing physicochemical and biological characteristics of nucleotides and amino acids
  - Side chain structure and chemistry
  - Side chain function
- Amino acid-based examples of considerations:
  - Cys/Pro are important for structure and function
  - Trp has a bulky side chain
  - Lys/Arg have positively charged side chains



Scoring Matrices

- ***Conservation:*** What residues can substitute for another residue and not adversely affect the function of the protein?
  - `Ile/Val` - both small and hydrophobic
  - `Ser/Thr` - both polar
  - *Conserve charge, size, hydrophobicity, additional physicochemical factors*
- ***Frequency:*** How often does a particular residue occur amongst the entire constellation of proteins?

Why is understanding scoring matrices important?

- Appear in all analyses involving sequence comparison
- Implicitly represent particular evolutionary patterns
- Choice of matrix can strongly influence outcomes of analyses

Matrix Structure: Nucleotides

- Simple match/mismatch scoring scheme:

  Match +2
  Mismatch −3

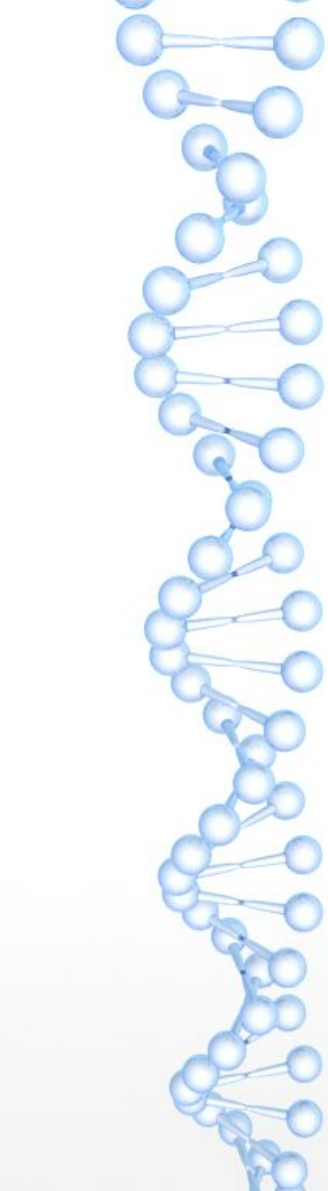
- Assumes each nucleotide occurs 25% of the time

	A	T	G	C
A	2	−3	−3	−3
T	−3	2	−3	−3
G	−3	−3	2	−3
C	−3	−3	−3	2

Matrix Structure: Proteins

	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

BLOSUM62

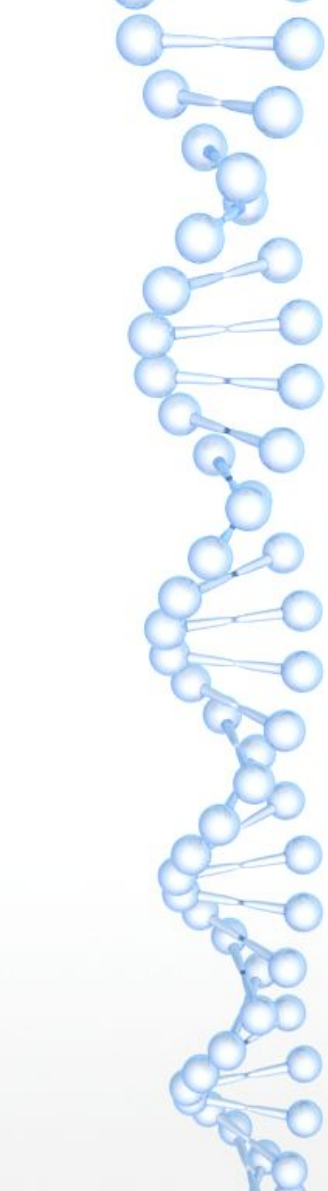


BLOSUM Matrices

- Look only for differences in conserved, ungapped regions of a protein family ('blocks')
- Directly calculated based on local alignments
  - Substitution probabilities (*conservation*)
  - Overall *frequency* of amino acids
- Sensitive to detecting structural or functional substitutions
- Generally perform better than PAM matrices for local similarity searches *(Henikoff and Henikoff, 1993)*
- BLOSUM series can be used to identify both closely and distantly related sequences

BLOSUM *n*

- Built using sequences sharing no more than *n*% identity
- Contribution of sequences > *n*% identical clustered and replaced by a sequence that represents the cluster



- Clustering reduces contribution of closely related sequences (less bias towards substitutions that occur in the most closely related members of a family)
- Reducing *n* yields more distantly related sequences
- Increasing *n* yields more closely related sequences

Which one to choose?

BLOSUM		% Similarity
90	Short alignments, highly similar	70-90
80	Best for detecting known members of a protein family	50-60
62	Most effective in finding all potential similarities	30-40
30	Longer, weaker local alignments	< 30

The takeaway…

No single matrix is the complete answer for all sequence comparisons

David Wheeler
Curr. Protoc. Bioinformatics
3.5.1 – 3.5.6, 2003

Selecting the Right Protein-Scoring Matrix

UNIT 3.5

OVERVIEW

Every program for searching protein sequences against a database includes a choice of a "protein-scoring matrix," also called a "weight matrix." Weight matrices add sensitivity to the search, while statistical significance adds selectivity (see UNIT 4.4). Virtually every user chooses the default, typically PAM 250 or BLOSUM62. Despite the fact that the choice of matrix can strongly influence the outcome of the analysis, most users do not know why a particular matrix should be used. In general, scoring matrices implicitly represent a particular theory of protein sequence evolution. This unit provides guidance in the choice of a scoring matrix, as understanding the assumptions underlying the PAM and BLOSUM scoring matrices can aid in making the proper choice. The selection of PAM matrices is covered first, after which the selection of BLOSUM matrices is discussed, and finally a brief overview of the wide variety of specialized scoring matrices is provided.

PAM MATRICES

PAM, a rearranged acronym derived from Accepted Point Mutation (Dayhoff, 1978) is a probabilistic model for amino acid replacement derived by comparing the frequencies of replacement in closely related sequences to the frequency expected from the completely random replacement of amino acids. The basis of this scoring system is the observation that the evolution of protein sequences is a nonrandom process—i.e., some amino acid replacements occur much more frequently than others, especially in related sequences. Amino acid substitutions tend to conserve charge, size, and hydrophobicity among other characteristics. One would expect that the substitution of glycine for alanine (CH_3 versus H) would have less of an effect on a protein's structure and function than the substitution of alanine for threonine (CH_3 versus substituted indole ring). The inference is that if two aligned sequences manifest a higher than expected prevalence of these characteristic replacements, the sequences are related. An excellent discussion of the derivation and use of the PAM matrices is given in George et al. (1990).

PAM matrices are the result of computing the probability of one substitution per 100 amino acids, called the PAM 1 matrix. Higher PAM matrices are derived by multiplying the PAM 1 matrix by itself a defined number of times. Thus, a PAM 160 matrix is the result of performing 160 matrix multiplications of the PAM 1 matrix against itself. Similarly, the PAM 250 matrix is derived by multiplying the PAM 1 matrix against itself 250 times.

Biologically, the PAM 50 matrix means that in 100 amino acids there have been 50 substitutions, while the PAM 250 matrix means there have been 2.5 amino acid replacements at each site (see UNIT 3.1 regarding insertions and deletions). This sounds unusual, but remember that over evolutionary time, it is possible that an alanine was changed to a glycine, then to a valine, and then back to an alanine. These silent substitutions are derived from observed amino acid frequency data in protein families and superfamilies.

Choosing a PAM Matrix

It is extremely important to note that PAM matrixes are derived from protein sequence data available in the late 1960s and early 1970s. Most proteins known at that time were small, globular, and hydrophilic. If the researcher believes their protein contains substantial hydrophobic regions, such as membrane-spanning helices or sheets, the PAM matrices are less useful than others described in this unit. Dayhoff et al. (1978) were the first to define the terms protein family and superfamily. A protein *family* is defined as sequences 85% identical or greater to each other. A protein *superfamily* is defined as sequences related from 30% identical or greater to each other. A protein superfamily may contain many protein families. The user should be aware that while the terms "family" and "superfamily" are widely used in biology, most of the time the original definition of Dayhoff and collaborators is not being used (see below).

Locating all potential similarities: PAM 250

The most widely used PAM matrix is PAM 250 (Fig. 3.5.1). It has been chosen because it is capable of accurately detecting similarities in the 30% range (i.e., superfamilies), that is, when the two proteins are up to 70% different from each other (George et al., 1990). Another way to think about this is that the PAM 250

Finding Similarities and Inferring Homologies

3.5.1

Contributed by David Wheeler
Current Protocols in Bioinformatics (2003) 3.5.1-3.5.6

Gaps

- Used to improve alignments between two sequences
  - Compensate for insertions and deletions
  - As such, *gaps represent biological events*
- Gaps must be kept to a reasonable number, to not reflect a biologically implausible scenario. About one gap per 20 residues is a good rule-of-thumb.
- Cannot be scored simply as a ‘match’ or a ‘mismatch’

Affine Gap Penalty

Fixed deduction for introducing a gap *plus*
an additional deduction proportional to the length of the gap

Deduction for a gap = $G + Ln$

				nucleotide	protein
where	G	=	gap-opening penalty	5	11
	L	=	gap-extension penalty	2	1
	n	=	length of the gap		
and	G	>	L		

BLAST:
The Basic Local Alignment Search Tool

- Seeks high-scoring segment pairs (HSPs)
  - Pair of sequences that can be aligned with one another
  - When aligned, have maximal aggregate score (score cannot be improved by extension or trimming)
  - Score must be above score threshold (*S*)
  - Gapped or ungapped
- Results not limited to the 'best' high-scoring segment pair for the two sequences being aligned

J Mol Biol. 1990 Oct 5;215(3):403-10.

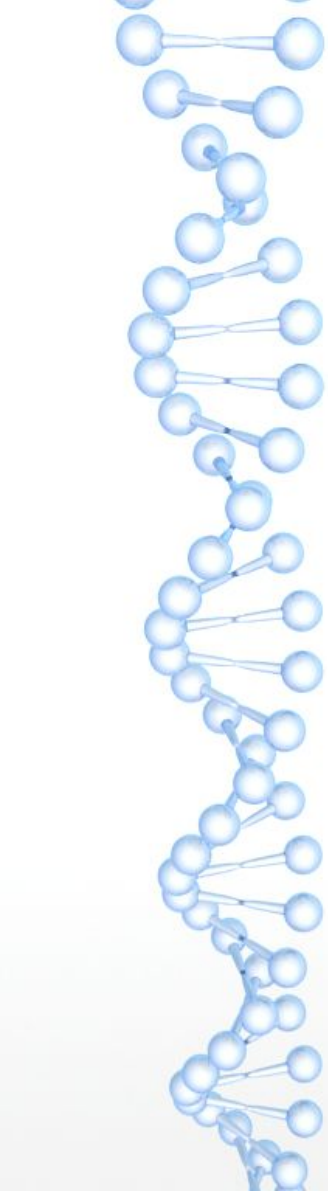
Basic local alignment search tool.

Altschul SF[1], Gish W, Miller W, Myers EW, Lipman DJ.

Author information

Abstract

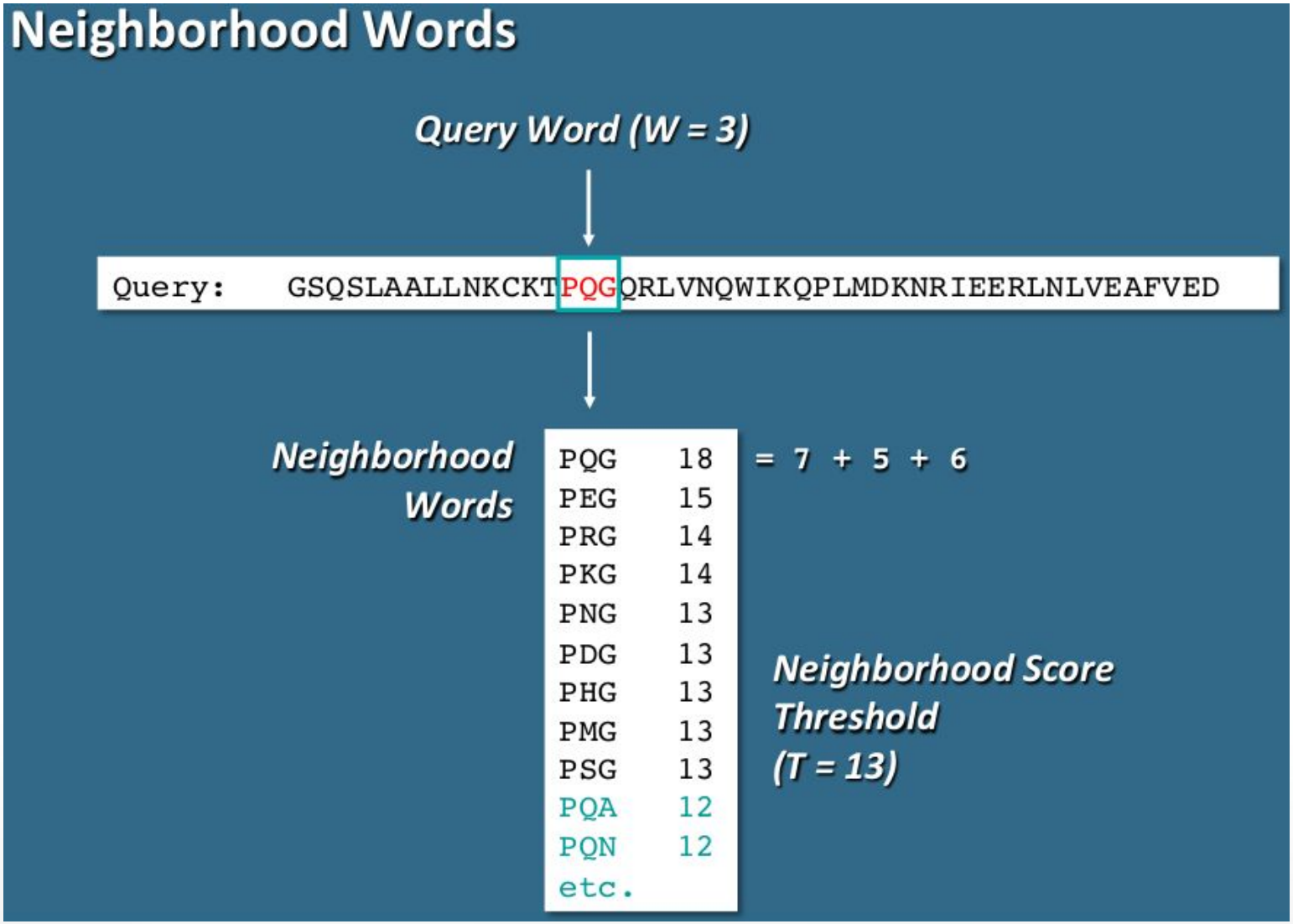
A new approach to rapid sequence comparison, basic local alignment search tool (BLAST), directly approximates alignments that optimize a measure of local similarity, the maximal segment pair (MSP) score. Recent mathematical results on the stochastic properties of MSP scores allow an analysis of the performance of this method as well as the statistical significance of alignments it generates. The basic algorithm is simple and robust; it can be implemented in a number of ways and applied in a variety of contexts including straightforward DNA and protein sequence database searches, motif searches, gene identification searches, and in the analysis of multiple regions of similarity in long DNA sequences. In addition to its flexibility and tractability to mathematical analysis, BLAST is an order of magnitude faster than existing sequence comparison tools of comparable sensitivity.



BLAST Algorithms

Program	*Query Sequence*	*Target Sequence*
BLASTN	Nucleotide	Nucleotide
BLASTP	Protein	Protein
BLASTX	Nucleotide, six-frame translation	Protein
TBLASTN	Protein	Nucleotide, six-frame translation
TBLASTX	Nucleotide, six-frame translation	Nucleotide, six-frame translation

Neighborhood Words

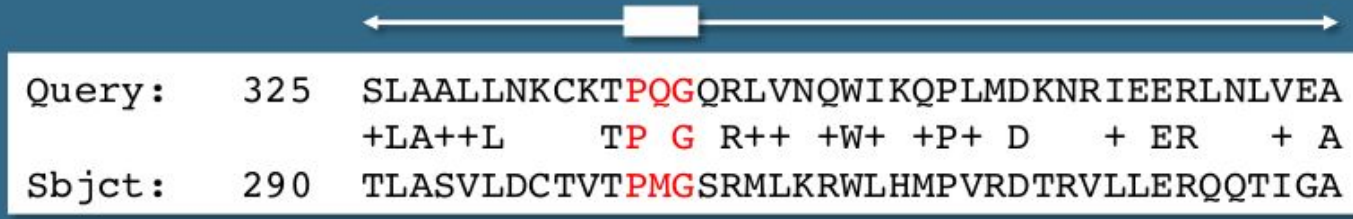


High-Scoring Segment Pairs

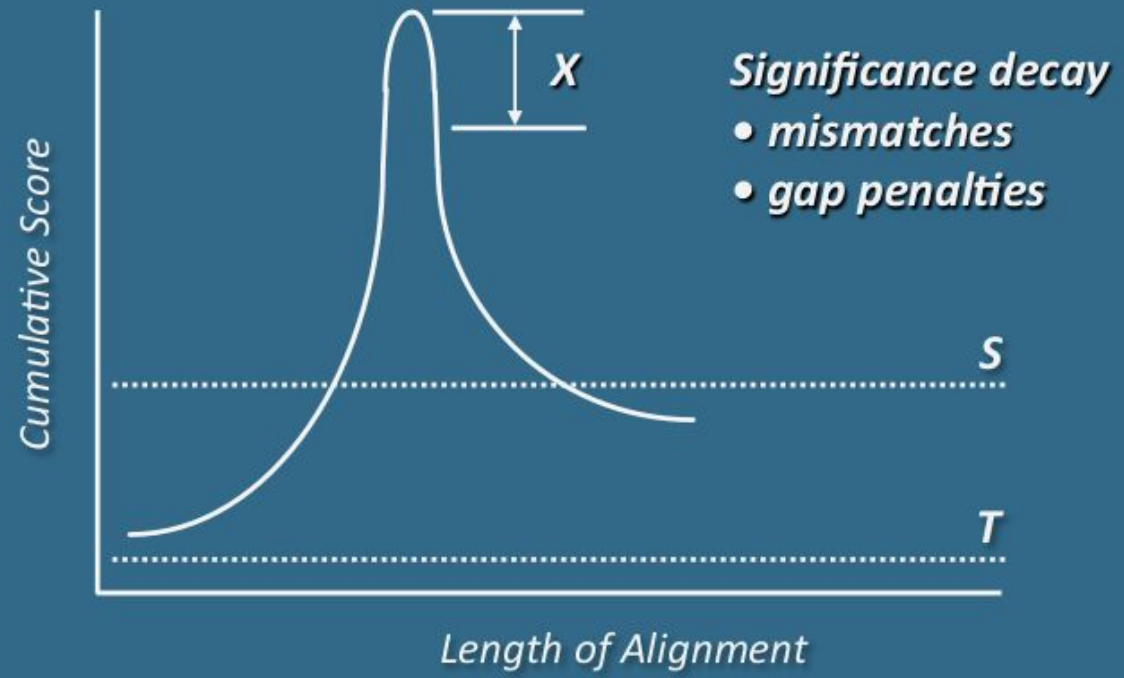
PQG	18
PEG	15
PRG	14
PKG	14
PNG	13
PDG	13
PHG	13
PMG	13
PSG	13
PQA	12
PQN	12
etc.	

```
Query:   325  SLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEA  365
              +LA++L    TP G R++ +W+ +P+ D   + ER   + A
Sbjct:   290  TLASVLDCTVTPMGSRMLKRWLHMPVRDTRVLLERQQTIGA  330
```

Extension



```
Query:   325  SLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEA  365
              +LA++L    TP G R++ +W+ +P+ D   + ER   + A
Sbjct:   290  TLASVLDCTVTPMGSRMLKRWLHMPVRDTRVLLERQQTIGA  330
```



Significance decay

- ***mismatches***
- ***gap penalties***

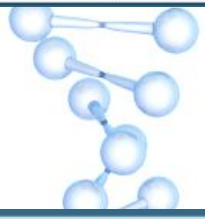
Scores and Alignment Length Don't Tell the Whole Story

```
Query: 1 SGLKSLVGKTALLSGTSSKL 20
         SGLKSLVGKTALLSGTSSKL
Sbjct: 1 SGLKSLVGKTALLSGTSSKL 20
```

Score = 91

```
Query: 1 CQHMWYQWMIQCIWMYHCMQ 20
         CQHMWYQWMIQCIWMYHCMQ
Sbjct: 1 CQHMWYQWMIQCIWMYHCMQ 20
```

Score = 138

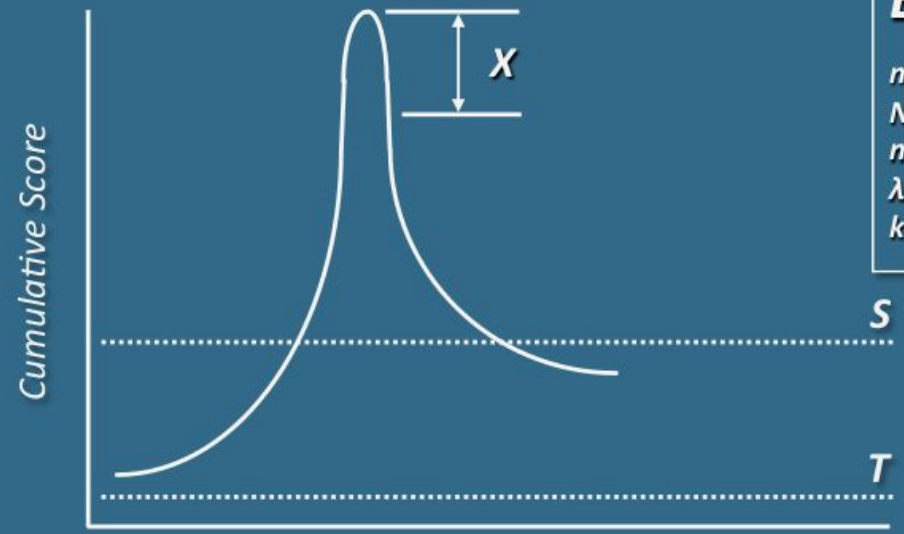


$E \leq 10^{-6}$ for nucleotides

$E \leq 10^{-3}$ for proteins

Scores and Probabilities

```
Query:   325  SLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEA  365
              +LA++L    TP G R++ +W+ +P+ D   + ER   + A
Sbjct:   290  TLASVLDCTVTPMGSRMLKRWLHMPVRDTRVLLERQQTIGA  330
```



$E = kmNe^{-\lambda S}$

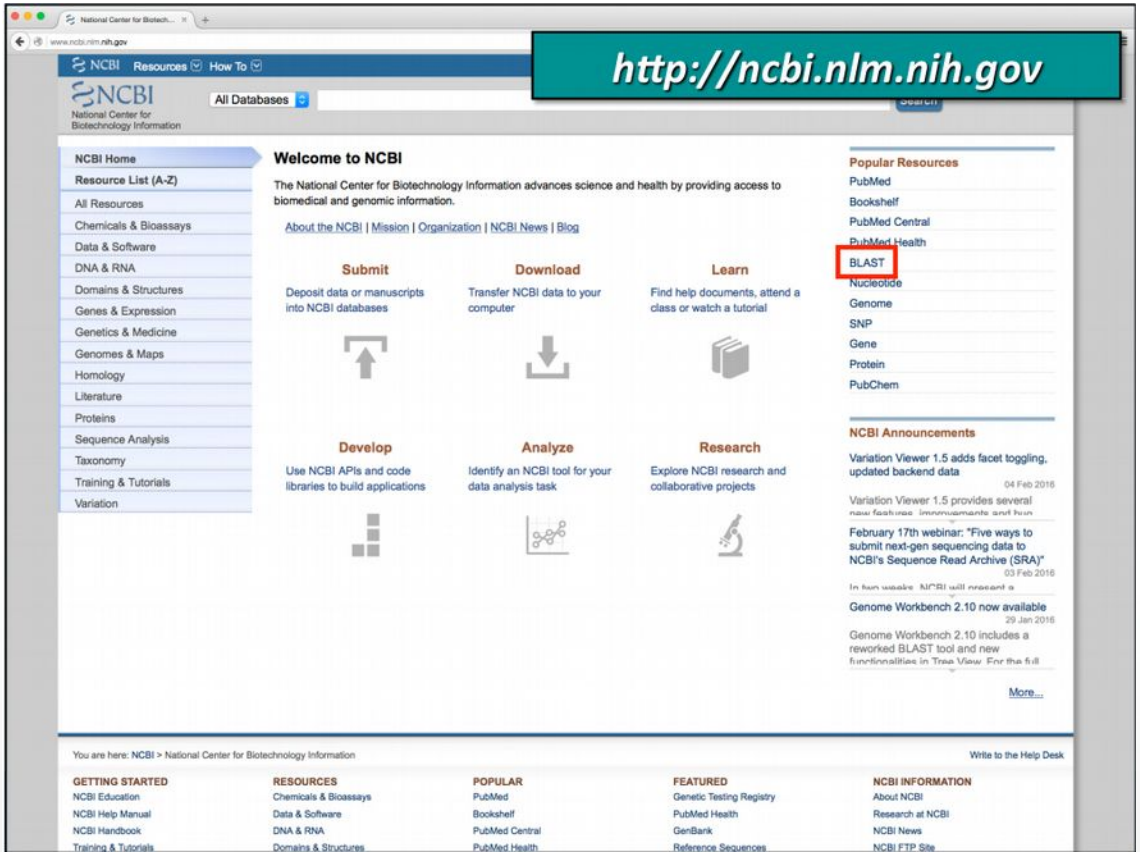
m	# letters in query
N	# letters in database
mN	size of search space
λS	normalized score
k	minor constant

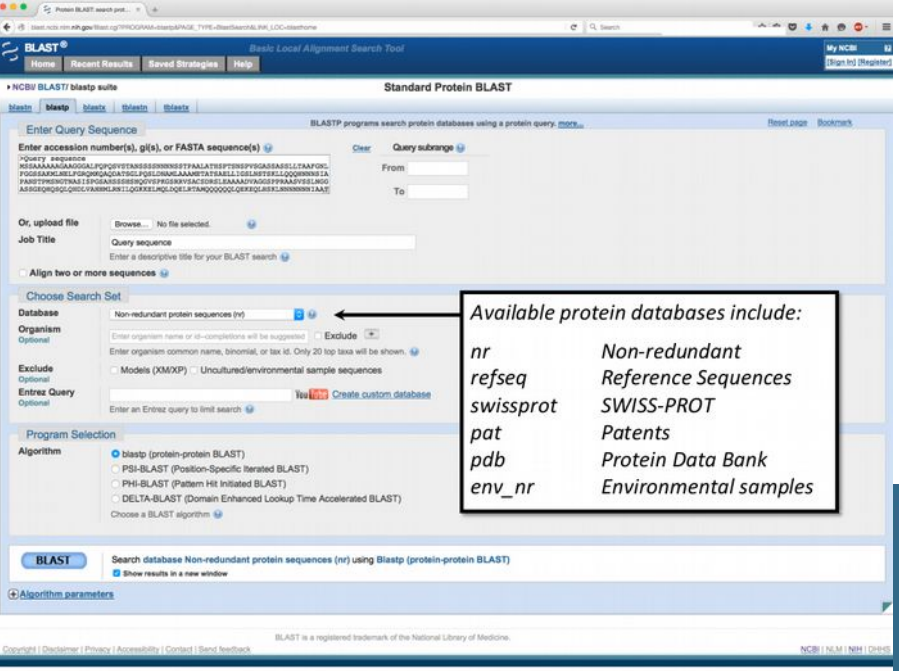
Number of HSPs found purely by chance

Lower values signify higher similarity

Using Blast for protein similarity searches

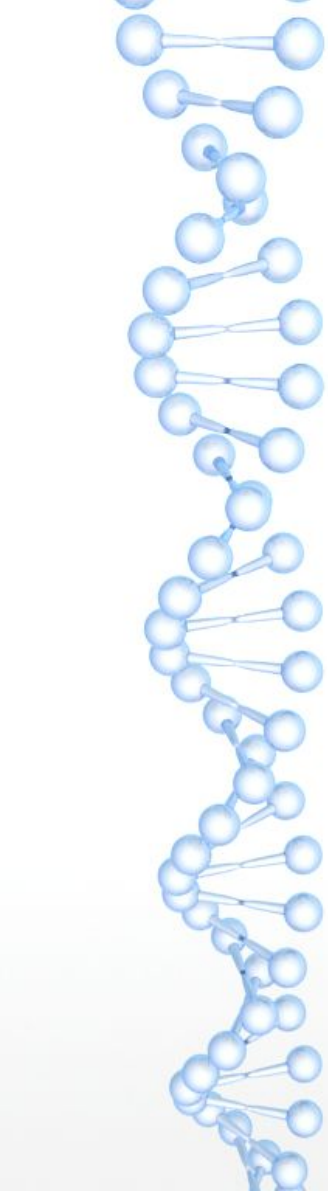
- https://mega.nz/#!Uh0DCIwS!KezJonumqHAcL4XTFdFRJdG8j-qxWk25WuWj9puh42E





NCBI RefSeq Database

- *Goal:* Provide a single reference sequence for each molecule of the central dogma (DNA, mRNA, and protein)
- Distinguishing features
  - Non-redundancy
  - Updates to reflect the current knowledge of sequence data and biology
  - Includes biological attributes of the gene, gene transcript, or protein
  - Encompasses a wide taxonomic range, with primary focus on mammalian and human species
  - Ongoing updates and curation (both automated and manual review), with review status indicated on each record



RefSeq Accession Number Prefixes

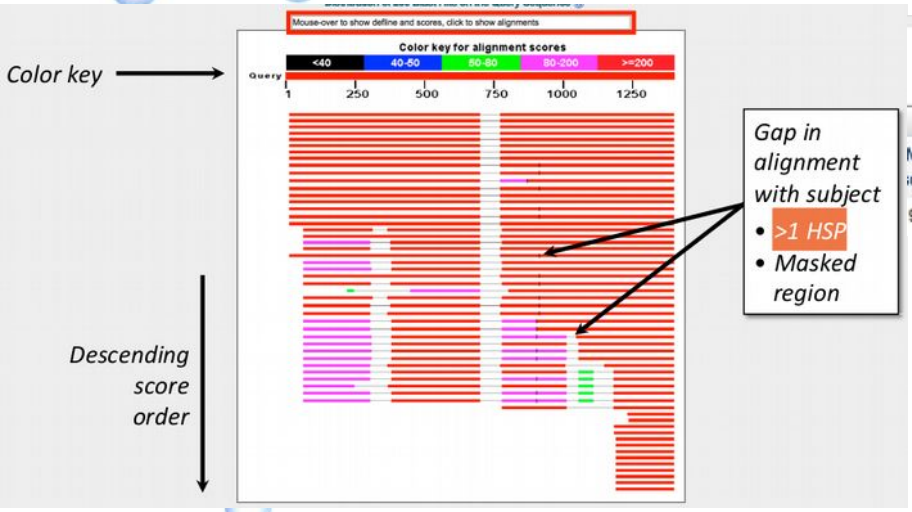
From curation of GenBank entries:

NT_	Genomic contigs
NM_	mRNAs
NP_	Proteins
NR_	Non-coding transcripts

From genome annotation:

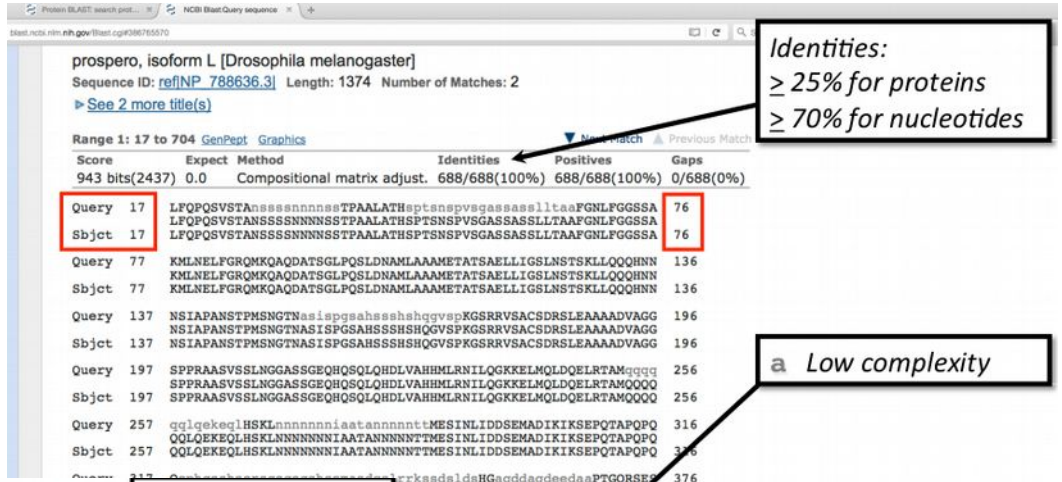
XM_	Model mRNA
XP_	Model proteins

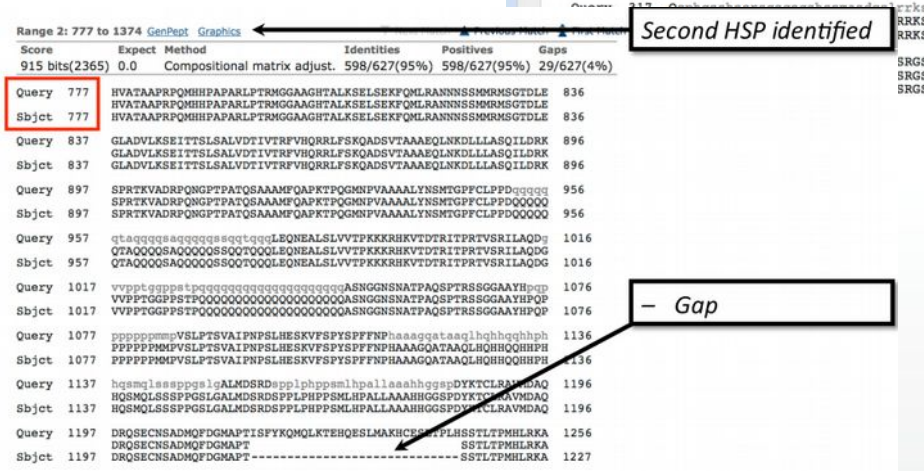
Complete list of molecule types in Chapter 18 of the NCBI Handbook
http://ncbi.nlm.nih.gov/books/NBK21091



0.0 means ≤ 10^{-1000}

Max score	Total score	Query cover	E value	Ident	Accession
994	1938	93%	0.0	100%	NP_001247046.1





Score	Expect	Method	Identities	Positives	Gaps
943 bits(2437)	0.0	Compositional matrix adjust.	688/688(100%)	688/688(100%)	0/688(0%)

Score	Expect	Method	Identities	Positives	Gaps
915 bits(2365)	0.0	Compositional matrix adjust.	598/627(95%)	598/627(95%)	29/627(4%)

HSP 1

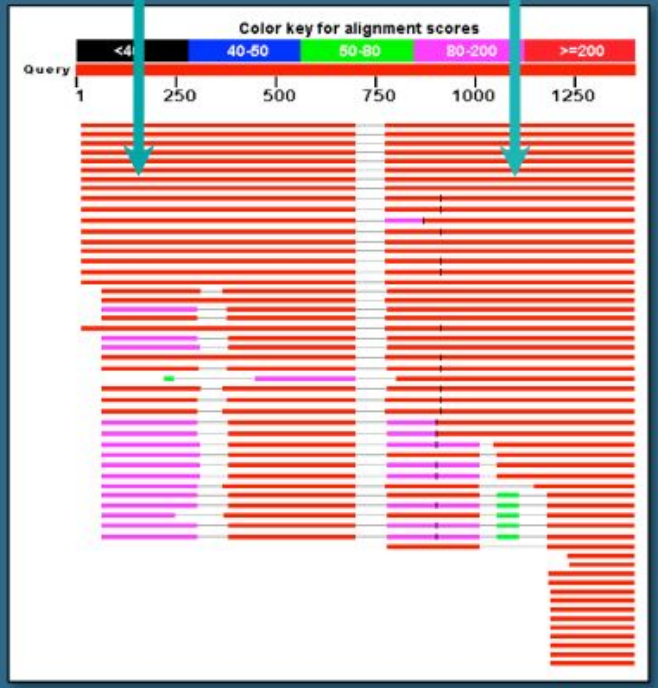
Q: 17– 704

S: 17– 704

HSP 2

Q: 777–1403

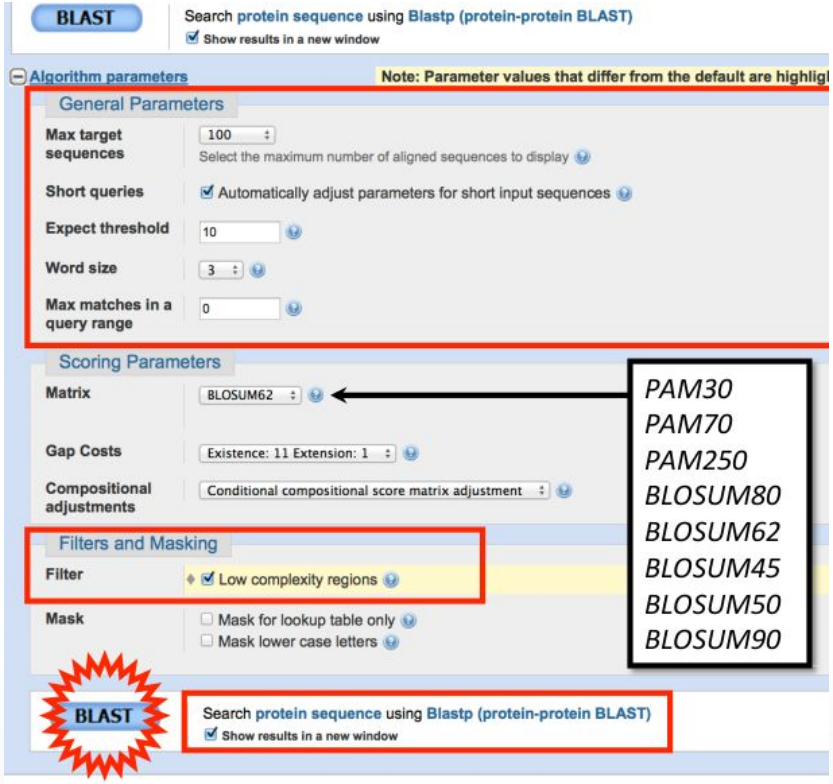
S: 777–1374



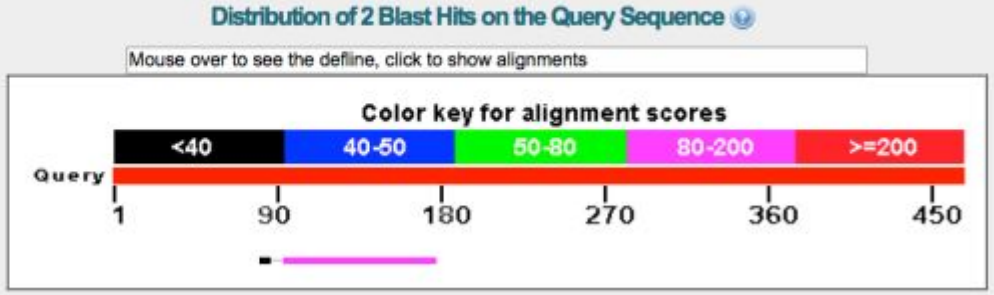
BLAST 2 Sequences

- Finds local alignments between two protein or nucleotide sequences of interest
- All BLAST programs available
- Select BLOSUM and PAM matrices available for protein comparisons
- Same affine gap costs (adjustable)
- Input sequences can be masked


BLAST®
Basic Local Alignment Searc
Home
Recent Results
Saved Strategies
Help
NCBI/ BLAST/ blastp suite
Align Seque
blastn
blastp
blastx
tblastn
tblastx
BLASTP programs search prot
Enter Query Sequence
Enter accession number(s), gi(s), or FASTA sequence(s)
Clear
>NP_008872.1 SOX-10 [Homo sapiens]
MAEEQDLSEVELSPVGSEEPRCLSPGSAPSLGPDGGGGGSGLRASPGPGELGKVKKEQQDGEADDDKFPV
CIREAVSQVLSGYDWTLVPMPVRVNGASKSKPHVKRPMNAFMVWAQAARRKLADQYPHLHNAELSKTLGK
LWRLLNESDKRPFIEEAERLRMQHKKDHPDYKYQPRRRKNGKAAQGEAECPGGEAEQGGTAAIQAHYKSA
HLDHRHPGEGSPMSDGNPEHPSGQSHGPPTPPTTPKTELQSGKADPKRDGRSMGEGGKPHIDFGNVDIGE
Or, upload file
Browse...
No file selected.
Job Title
NP_008872.1 SOX-10 [Homo sapiens]
Enter a descriptive title for your BLAST search
Align two or more sequences
Enter Subject Sequence
Enter accession number(s), gi(s), or FASTA sequence(s)
Clear
>NP_003131.1 sex determining region Y [Homo sapiens]
MQSYASAMLSVFNSDDYSPAVQENIPALRRSSSFLCTESCNSKYQCETGENSKGNVQDRVKRPMNAFIVW
SRDQRRKMALENPRMRNSEISKQLGYQWKMLTEAEKWPFFQEAQKLQAMHREKYPNYKYRPRRKAKMLPK
NCSLLPADPASVLCSEVQLDNRLYRDDCTKATHSRMEHQLGHLPPINAASSPQQRDRYSHWTKL
Or, upload file
Browse...
No file selected.
Program Selection
Algorithm
blastp (protein-protein BLAST)
Choose a BLAST algorithm
BLAST
Search protein sequence using Blastp (protein-protein BLAST)
Show results in a new window
Algorithm parameters


BLAST
Search protein sequence using Blastp (protein-protein BLAST)
Show results in a new window
Algorithm parameters
Note: Parameter values that differ from the default are highlig
General Parameters
Max target sequences
100
Select the maximum number of aligned sequences to display
Short queries
Automatically adjust parameters for short input sequences
Expect threshold
10
Word size
3
Max matches in a query range
0
Scoring Parameters
Matrix
BLOSUM62
Gap Costs
Existence: 11 Extension: 1
Compositional adjustments
Conditional compositional score matrix adjustment
Filters and Masking
Filter
Low complexity regions
Mask
Mask for lookup table only
Mask lower case letters
PAM30
PAM70
PAM250
BLOSUM80
BLOSUM62
BLOSUM45
BLOSUM50
BLOSUM90
BLAST
Search protein sequence using Blastp (protein-protein BLAST)
Show results in a new window

Graphic Summary



NP_003131.1 sex determining region Y [Homo sapiens]

Sequence ID: lcl|Query_213411 Length: 204 Number of Matches: 2

Range 1: 51 to 134 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
94.0 bits(232)	1e-26	Compositional matrix adjust.	39/84(46%)	62/84(73%)	0/84(0%)

```
Query  95   NGASKSKPHVKRPMNAFMVWAQAARRKLADQYPHLHNAELSKTLGKLWRLLNESDKRPFI  154
            N     +  VKRPMNAF+VW++  RRK+A + P + N+E+SK LG  W++L E++K PF
Sbjct  51   NSKGNVQDRVKRPMNAFIVWSRDQRRKMALENPRMRNSEISKQLGYQWKMLTEAEKWPFF  110

Query  155  EEAERLRMQHKKDHPDYKYQPRRR  178
            +EA++L+  H++ +P+YKY+PRR+
Sbjct  111  QEAQKLQAMHREKYPNYKYRPRRK  134
```

Range 2: 95 to 101 Graphics ▼ Next Match ▲ Previous Match ▲ First Match

Score	Expect	Method	Identities	Positives	Gaps
15.4 bits(28)	1.9	Compositional matrix adjust.	3/7(43%)	5/7(71%)	0/7(0%)

```
Query  82   GYDWTLV  88
            GY W ++
Sbjct  95   GYQWKML  101
```

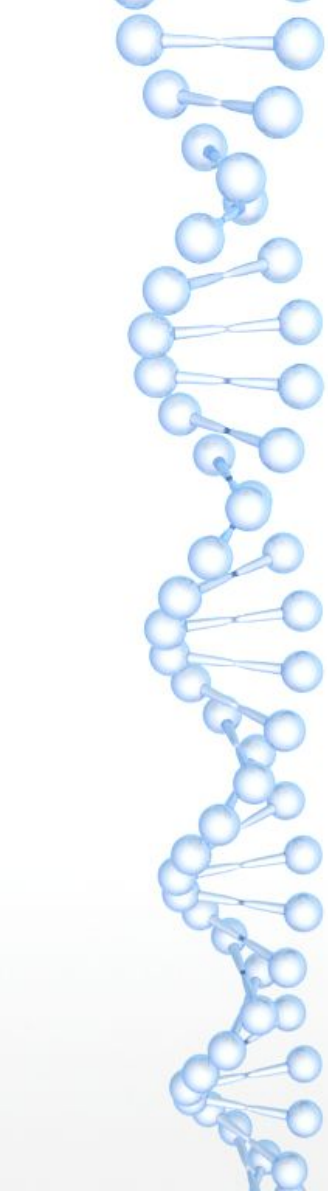
Global alignment

- https://www.ebi.ac.uk/Tools/psa/emboss_needle/

```
EMBOSS_001         1 MAEEQDLSEVELSPVGSEEPRCLSPGSAPSLGPDGGGGGGSGLRASPGPGE     50

EMBOSS_001         1 --------------------------------------------------       0

EMBOSS_001        51 LGKVKKEQQDGEA------DDDKFPV------CIREAVSQVLSGYDWTLV     88
                        .|....|      .||..|.      .:|.:.|.:.:....:..
EMBOSS_001         1 ------MQSYASAMLSVFNSDDYSPAVQENIPALRRSSSFLCTESCNSKY     44

EMBOSS_001        89 PMPVRVNGASKSKPHVKRPMNAFMVWAQAARRKLADQYPHLHNAELSKTL    138
                     ......|....:..||||||||:||::..|||:|.:.|.:.|:|:||.|
EMBOSS_001        45 QCETGENSKGNVQDRVKRPMNAFIVWSRDQRRKMALENPRMRNSEISKQL     94

EMBOSS_001       139 GKLWRLLNESDKRPFIEEAERLRMQHKKDHPDYKYQPRRRKNGKAAQGEA    188
                     |..|::|.|::|.||.:||::|:..|::.:|:|||:|||    ||.....
EMBOSS_001        95 GYQWKMLTEAEKWPFFQEAQKLQAMHREKYPNYKYRPRR----KAKMLPK    140

EMBOSS_001       189 ECPGGEAEQGGTAAIQAHYKSAHLDHR-HPGEGSPMSDGNPEHPSGQSHG    237
                     .|....|:.....     ....||:| :..:.:..:....||..|  |.
EMBOSS_001       141 NCSLLPADPASVLC-----SEVQLDNRLYRDDCTKATHSRMEHQLG--HL    183

EMBOSS_001       238 PP-TPPTTPK--------TELQSGKADPKRDGRSMGEGGKPHIDFGNVDI    278
                     || ...::|:         |:|
EMBOSS_001       184 PPINAASSPQQRDRYSHWTKL----------------------------    204

EMBOSS_001       279 GEISHEVMSNMETFDVAELDQYLPPNGHPGHVSSYSAAGYGLGSALAVAS    328

EMBOSS_001       205 --------------------------------------------------    204

EMBOSS_001       329 GHSAWISKPPGVALPTVSPPGVDAKAQVKTETAGPQGPPHYTDQPSTSQI    378

EMBOSS_001       205 --------------------------------------------------    204

EMBOSS_001       379 AYTSLSLPHYGSAFPSISRPQFDYSDHQPSGPYYGHSGQASGLYSAFSYM    428

EMBOSS_001       205 --------------------------------------------------    204

EMBOSS_001       429 GPSQRPLYTAISDPSPSGPQSHSPTHWEQPVYTTLSRP    466

EMBOSS_001       205 ------------------------------------    204
```

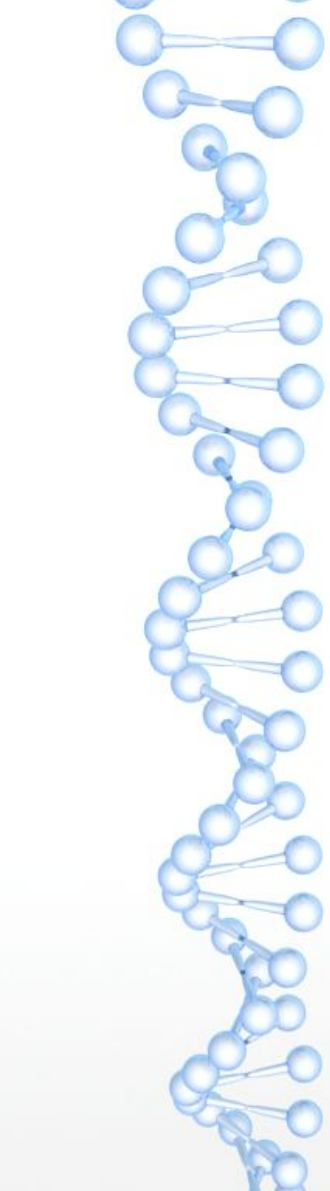
Nucleotide-Based BLAST Algorithms

	W	*+/–*	*Gaps*
Optimized for aligning very long and/or highly similar sequences (> 95%)			
MegaBLAST *(default)*	28	1, –2	Linear
Better for diverged sequences and/or cross-species comparisons (< 80%)			
Discontiguous MegaBLAST	11	2, –3	Affine
BLASTN	11	2, –3	Affine
Finding short, nearly exact matches (< 20 bases)			
BLASTN	7	2, –3	Affine



BLAT

- "BLAST-Like Alignment Tool"
- Designed to rapidly align longer nucleotide sequences ($L \geq 40$) having ≥ 95% sequence similarity
- Can find exact matches reliably down to $L = 33$
- Method of choice when looking for exact matches in nucleotide databases
- 500 times faster than BLAST for mRNA/DNA searches
- May miss divergent or shorter sequence alignments
- Can be used on protein sequences, but BLASTP is more efficient



BLAT

- "BLAT-Like Alignment Tool"
- Designed to rapidly align longer nucleotide sequences ($L \geq 40$) having ≥ 95% sequence similarity
- Can find exact matches reliably down to $L = 33$
- Method of choice when looking for exact matches in nucleotide databases
- 500 times faster than BLAST for mRNA/DNA searches
- May miss divergent or shorter sequence alignments
- Can be used on protein sequences, but BLASTP is more efficient

BLAT Search Genome

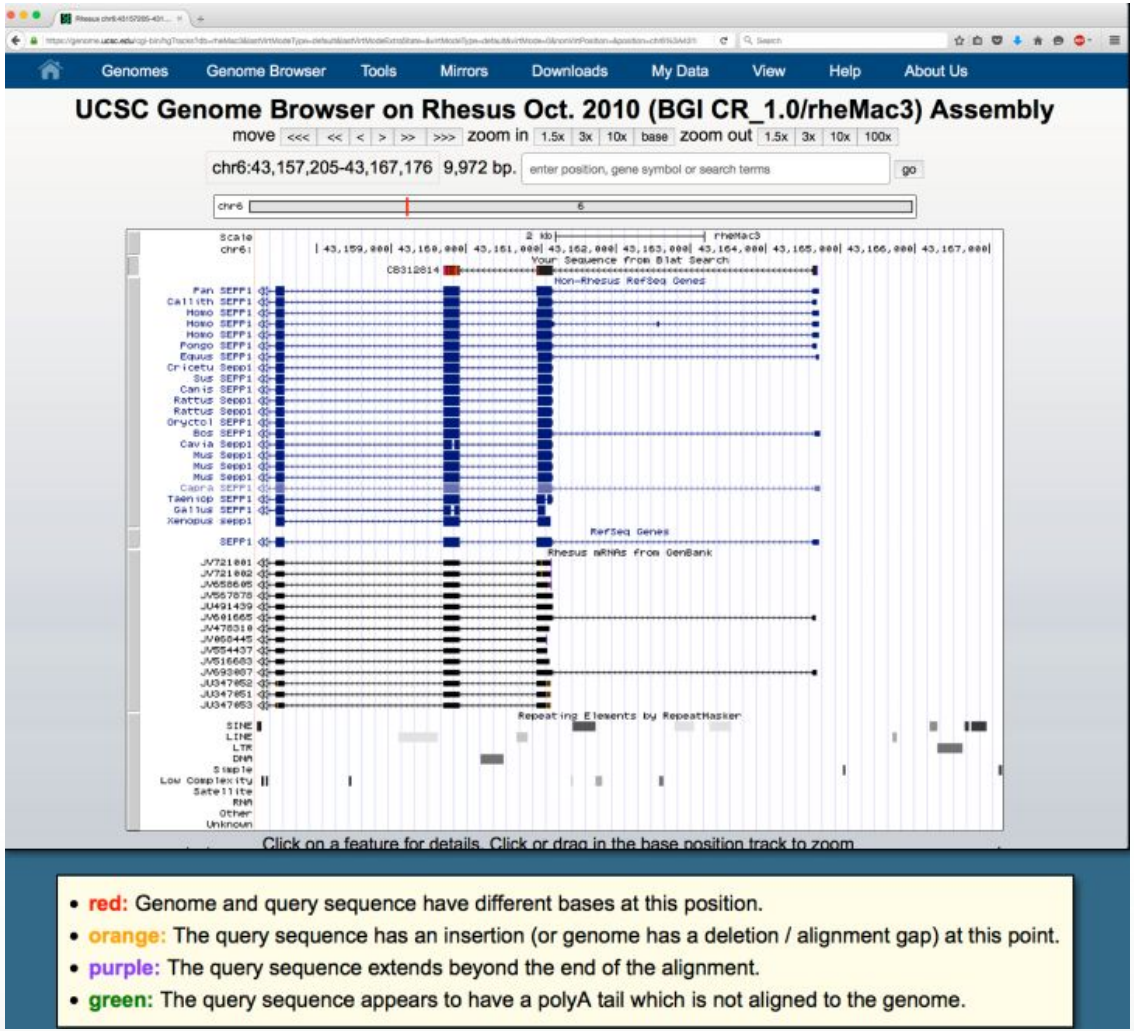
Genome: Rhesus | Assembly: Oct. 2010 (BGI CR_1.0/rheMac3) | Query type: DNA | Sort output: query,score | Output type: hyperlink

```
>CB312814 NICHD_Rh_Ov1 Macaca mulatta cDNA clone
GGGGGTGGAGCTGCCAGAGTAAAGCAAAGAGCAAGGAAGCAGGCTCGTTGGAAGGGGTTGTGACAGCCCC
AGCAATGTGGAGAAGTCTGGGGCTTGCCCTGGCTCTCTGTCTCCTTCCATCGGGAGGAACAGAGAGCCAG
GACCAAAGCTCCTTCTGTAAGCAACCCCCAGCCTGGAGCATAAGAGATCAAGATCCAATGCTAGACTCCA
ATGGTTCAGTGACTGTGGTCGCTCTTCTTCAAGCCAGCTGATACCTGTGCATACTGCANGCATCTAAATT
GGAAGAACTGCGAGTAAAACTGGAGAAAGAAGGATATTCTAAATATTCCTATATTGGTGGTAATCATCAA
GGGATCTCTTCTCGATTAAAATACACACATCTTTAGAAAAAAGGTTTCAGAGCATATTCCTGTATATTCA
CCAGAAGAAAACCCAACCGATGTCTGGACTCTTTTAATGGAAACCAAGAAGACCTCCTCATATATGACGG
ATGTGGCCTTCCTGGAAAACACCCTGGTTGGCCTTTTTCCTTCCCAACCTTGGCGAATGGTAAAAAAACC
CCTTTAAATGGTTTTCCGGGAAAAAAAAAGTGGGAAATTGGTCTCCTCCCAAATCTCAAAAAAGAAAAAA
TTTTTGTAAAAAGGGATCTTTTTGGGCACCGGGGGGAAAAAAAAATTTGAAAACTTCCCCCACCCCCCTT
TTTCCCTCTTTGGGGACTCCTTCCCAAATTCCGGGGACATCCCCCCT
```

submit | I'm feeling lucky | clear

Paste in a query sequence to find its location in the the genome. Multiple sequences may be searched if separated by lines starting with '>' followed by the sequence name.

I'm feeling lucky returns only the highest scoring alignment (direct path to genome browser)


Genomes
Genome Browser
Tools
Mirrors
Downloads
My Data
View
Help
About Us
UCSC Genome Browser on Rhesus Oct. 2010 (BGI CR_1.0/rheMac3) Assembly
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x
chr6:43,157,205-43,167,176 9,972 bp.
enter position, gene symbol or search terms
go
Click on a feature for details. Click or drag in the base position track to zoom
red: Genome and query sequence have different bases at this position.
orange: The query sequence has an insertion (or genome has a deletion / alignment gap) at this point.
purple: The query sequence extends beyond the end of the alignment.
green: The query sequence appears to have a polyA tail which is not aligned to the genome.