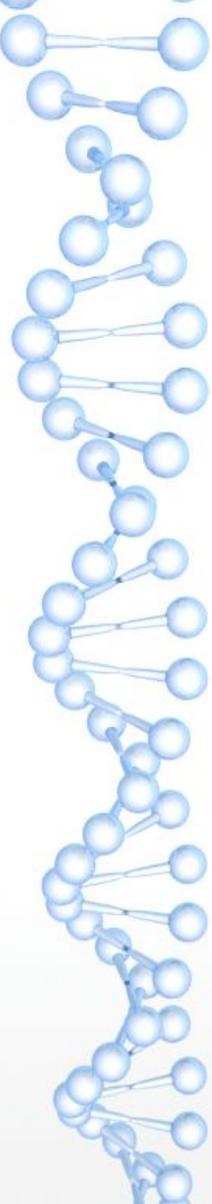


Biological Sequence Analysis

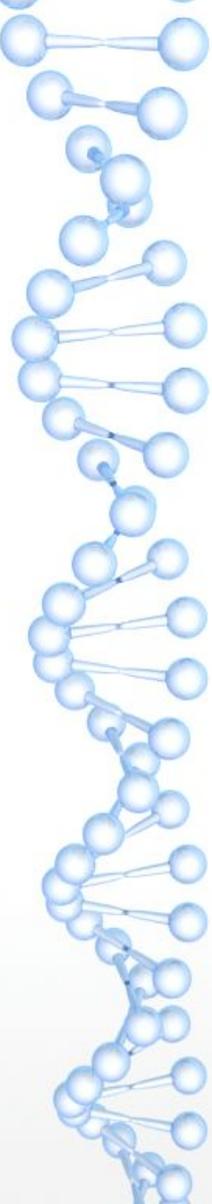
Alberto Pallavicini
Applied genomics



Sequence Alignments: Determining Similarity and Deducing Homology

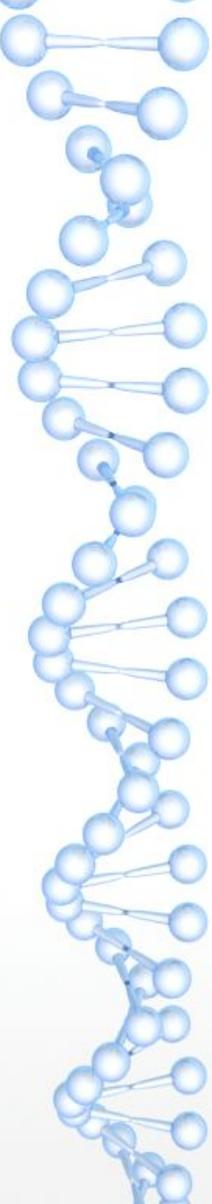
Why construct sequence alignments?

- 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



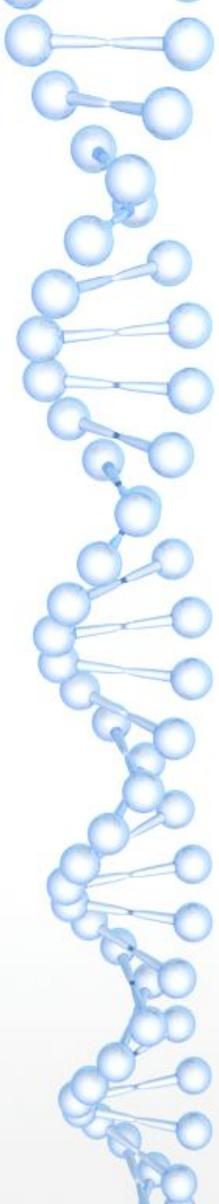
Defining the Terms

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 indivisible⁸. 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



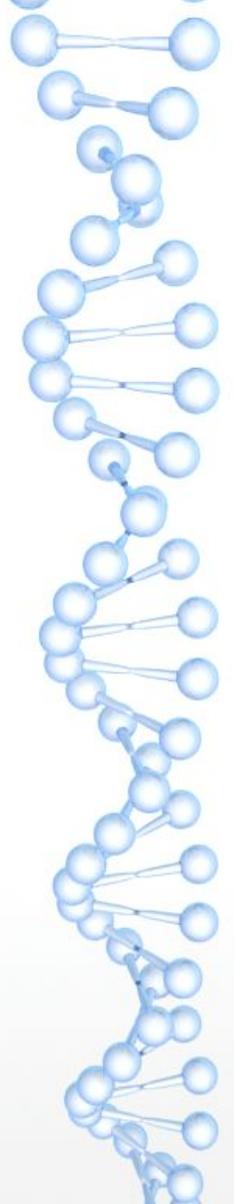
Speciation events



Gene in common ancestor

Gene duplication

- 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

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

Orthologs, Paralogs, and Evolutionary Genomics¹

Eugene V. Koonin

National Center for Biotechnology Information, National Library of Medicine,
National Institutes of Health, Bethesda, Maryland 20894;
e-mail: koonin@ncbi.nlm.nih.gov

Key Words

homology, ortholog, paralog, pseudortholog, pseudoparalog, xenolog

Abstract

Orthologs and paralogs are two fundamentally different types of homologous genes that evolved, respectively, by vertical descent from a single ancestral gene and by duplication. Orthology and paralogy are key concepts of evolutionary genomics. A clear distinction between orthologs and paralogues is critical for the construction of a robust evolutionary classification of genes and reliable functional annotation of newly sequenced genomes. Genome comparisons show that orthologous relationships with genes from taxonomically distant species can be established for the majority of the genes from each sequenced genome. This review examines in depth the definition and subtypes of orthologs and paralogs, outlines the principal methodological approaches employed for identification of orthology and paralogy, and considers evolutionary and functional implications of these concepts.

Homology

Reviews

Homology

a personal view on some of the problems

There are many problems relating to defining the terminology used to describe various biological relationships and getting agreement on which definitions are best. Here, I examine 15 homological problems, all of which are current, and all of which relate to the usage of homology and its associated terms. I suggest a set of definitions that are intended to be widely consistent among themselves and also as consistent as possible with most current usage.

I have frequently been asked about some controversial issues concerning the usage of homology and related terms. I maintain most of these terms as a set of 15 problems. This is an opinion on how best to maintain clarity on the use of these concepts with no link given to where the terms are possible. Part of this clarity lies in making sure the definitions are self-consistent. There are many alternative definitions for most of these terms and I might say that we don't need another paper discussing this. But there is so much squabbling about best usage, and so much misusage by non-investigators to the field, especially by molecular biologists, mathematicians and bioinformatics people, that, if this could help investigators express these ideas more clearly and get others to maintain their own definitions and keep these within some bounds, it will be worth the effort. I have avoided phrases like "I would suggest" and "we agree" to save space. Issues that I have listed wherever the tree seems too divergent. Although the examples are highly technical, the issues are to be as general as possible, and a glossary is found in Box 1. This article is an invited follow-up to the excellent paper here in 1997 (Box 1). Other good discussions of some of these topics can be found in Box 2 and 3. For a comparison between molecular and morphological, see Box 4.

Homology is the similarity of two characters that have descended, usually with divergence, from a common ancestral character. This is important because most of the terminological problems arise from different definitions of homology. Characters can be any traits, structural or behavioral, based on an organism. Analog is distinguished from homology in that its characters, although similar, have descended convergently from unrelated ancestral characters. The commonness in the most recent common ancestor of the tree being considered.

Other homologies problem
Organic chemists consider compounds such as methanol, ethanol and propanol as an homologous series because each differs from the next by a CH₂ group. This happens to be the same CH₂ group that occurs in Mathematics but are used for nothing for the same as well. There is no point in worrying about these differences, except to suggest that molecular biologists, mathematicians and bioinformaticians working in the field of biology have not done it the biological definition.

The relationship problem

Homology was first defined in biology with something like its present meaning by Owsen in 1861. The character and homology as the same origin under every variety of form and function¹. Common ancestry is not mentioned in that definition, which is somewhat given that these were pro-Farwellian and pro-Mendelian views. Owsen's definition of homology captures essence and function rather than ancestry. Some would like to return to Owsen's definition, perhaps out of a sense of nostalgia or some personal need for unchanging meaning. But that would mean inventing a word to designate common ancestry. The meaning of a word should change if that change is a reflection that increases clarity of purpose, insight and expression, as it does here.

The character-character-state problem

Many characters, and nearly all molecular mutations, distinguish between a character, say amino acid, and its character states, say glycine and phenylalanine. This method distinction is not concerned. Many scientists will, if two character states are not the same, assert that the characters are non-homologous. This is confusing because it implies that the two characters do not have a common ancestor, which, if true, means that should not have been comparing the character states in the first place. Homology results in the characters, not in their state!

The homology-homology problem

Analogy describes characters whose similar arise from convergent processes. Homology describes characters, irrespective of their character states, whose similarity arises from divergence from a common ancestral form. Homology is the complement of analogy in that there are no convergent states all known microevolutionary explanations of similarities could make homology, it was mentioned by Laskov², the complement of homology. But homology is a relationship character character states in a tree, whereas analogy is a relation of two characters, independent of any tree, making homology non-complementary to homology.

The recognition of homology problem

How can we know for sure that two sequences are homologous? One would always "know" if one defined homology objectively as those having common such as molecules.

John N. Huxley

Editorial Board

Member of Council

and Institute Board

of the Royal Society

© 1997, Royal Society

0013-0745

10.1093/ajph/87.10.1472

This published online as a Review in Advance on August 14, 2005

The Annual Review of Genomics is online at <http://genom.annualreviews.org>

doi: 10.1146/annurev.genom.06.080505.114722

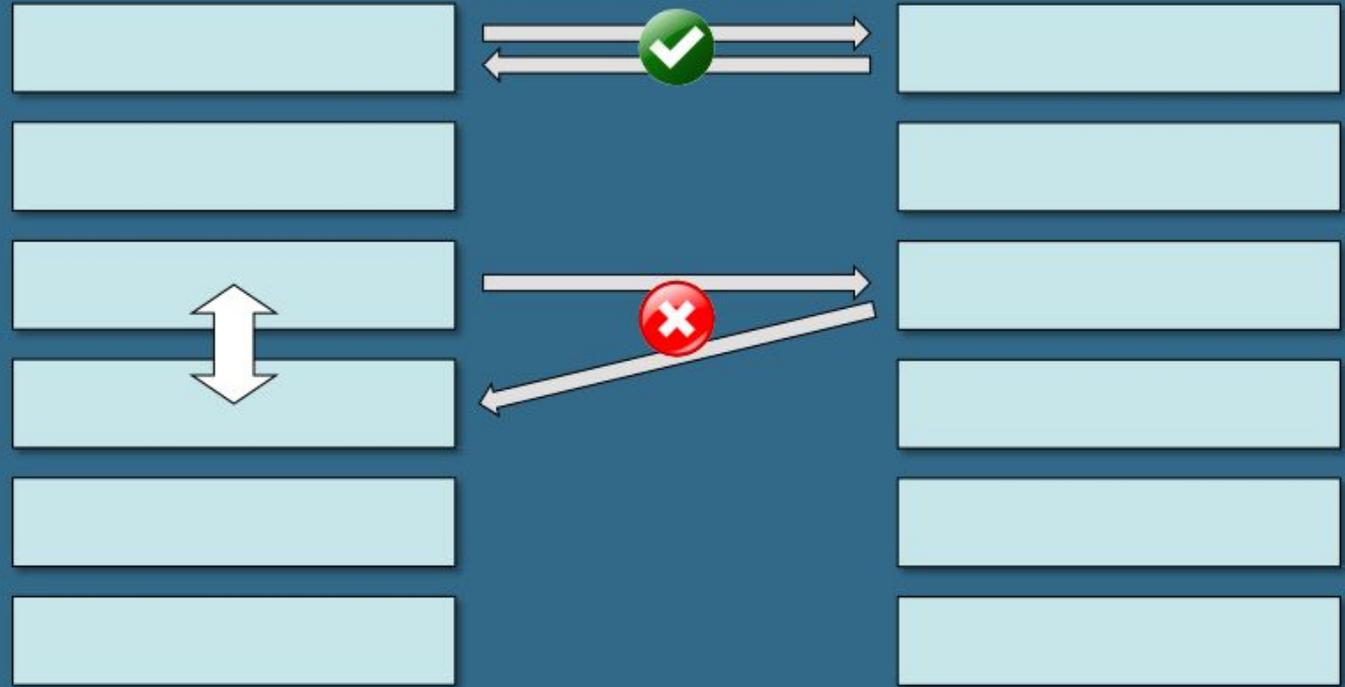
Copyright © 2005 by Annual Reviews. All rights reserved.

The U.S. Government has the right to reproduce, retransmit, copy, disseminate, or otherwise use or copyright copying this paper.

0864-4777/05/1215-1000\$3.00

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

Identifying Candidate Orthologs: Reciprocal Best Hits



Organism 1

Organism 2

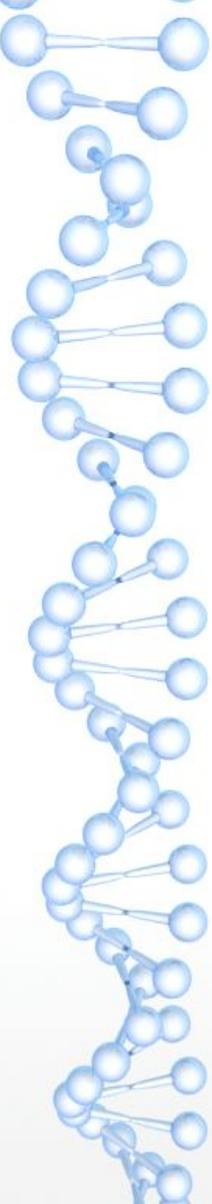
Global Sequence Alignments

- 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

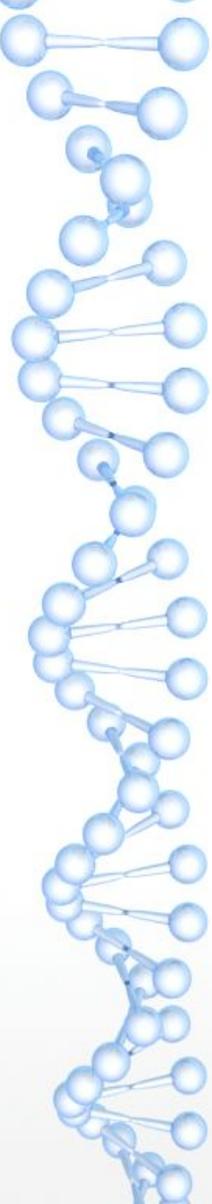
- Sequence comparison intended to find the most similar regions in the two sequences being aligned ('paired subsequences')
- Regions outside the area of local alignment are excluded
- More than one local alignment could be generated for any two sequences being compared
- Best for sequences that share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

	A	T	G	C
A	2	-3	-3	-3
T	-3	2	-3	-3
G	-3	-3	2	-3
C	-3	-3	-3	2

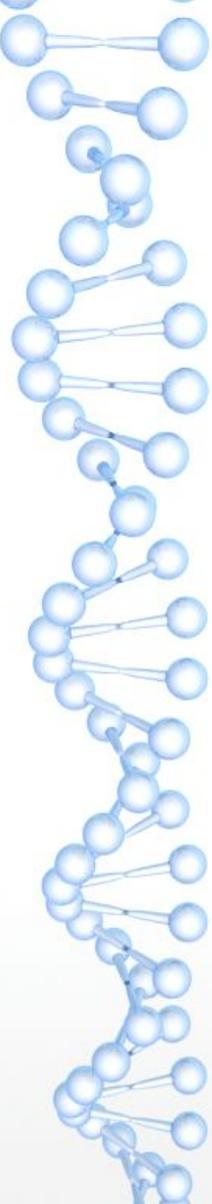
- Assumes each nucleotide occurs 25% of the time



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	-1	-2	-2	0	-1	-2	-2	-1	-2	-1	4	1	-3	-2	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	1	1	2	2	3	2	2	3	2	3	1	1	1	3	2	11	2	-3	-4	-3	-2	-4
Y	2	2	2	2	2	1	2	2	2	1	1	2	1	2	2	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

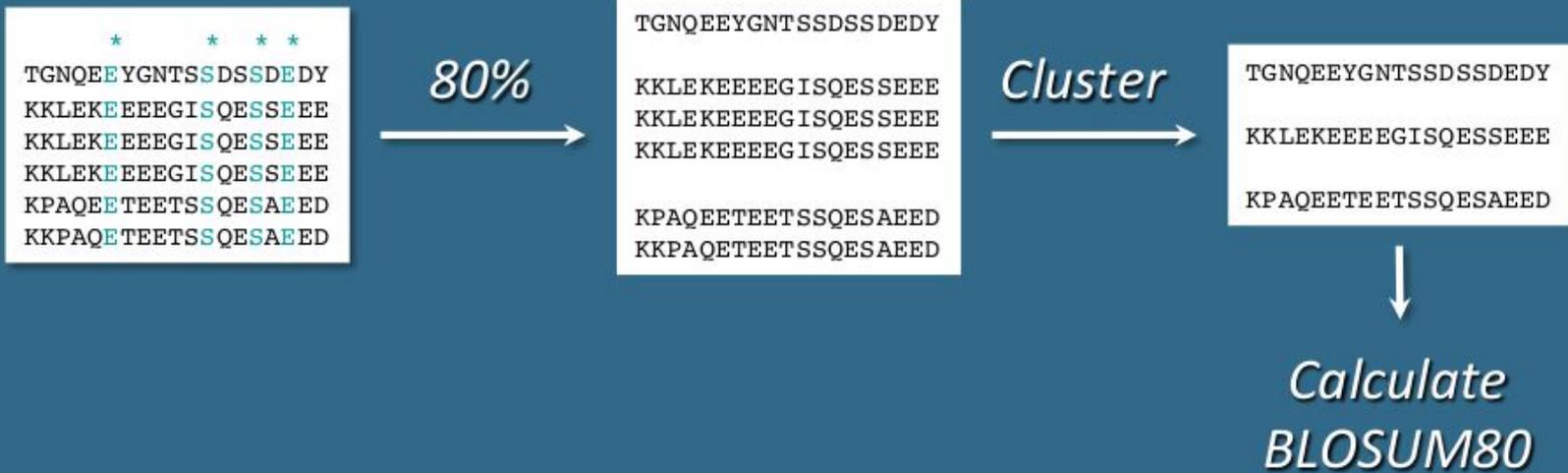


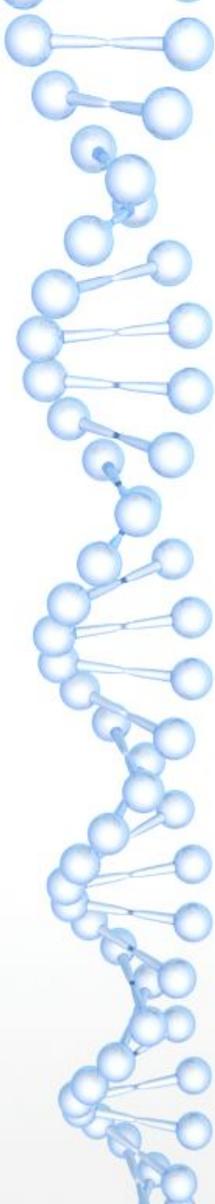
BLOSUM Matrices

- 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

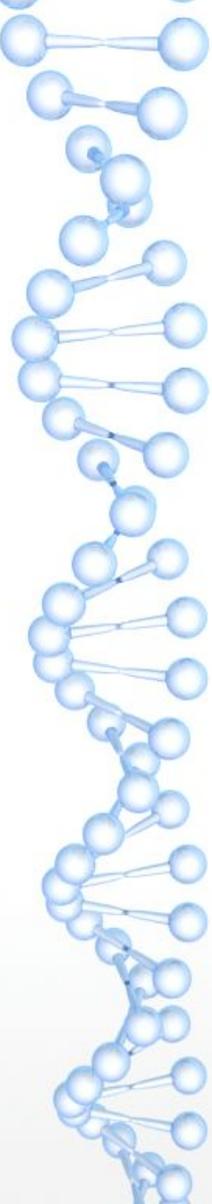




- Clustering reduces contribution of closely related sequences (less bias towards substitutions that occur in the most closely related members of a family)
- Reducing n yields more distantly related sequences
- Increasing n yields more closely related sequences

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 *over 2.1*). 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 (G1, versus H) would have less of an effect on a protein's structure and function than the substitution of alanine for threonine (A1) versus substituted indole (I). 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 *over 2.1* 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 matrices 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
Copyright © 2003 by John Wiley & Sons, Inc.

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'

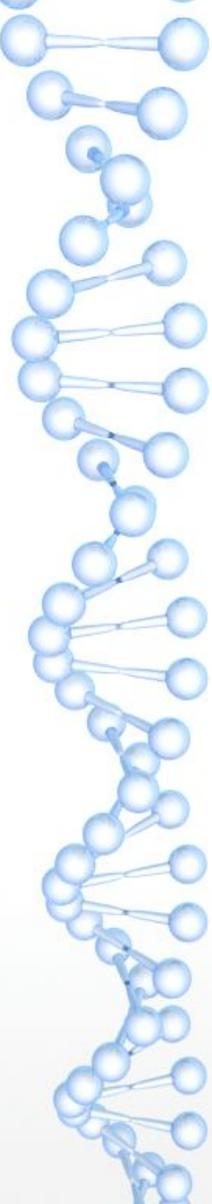


Affine Gap Penalty

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

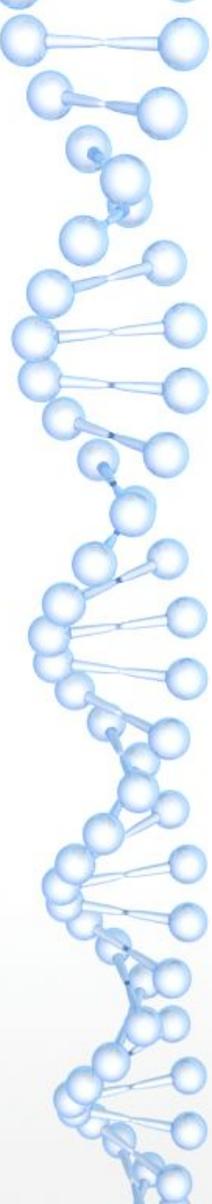
$$\text{Deduction for a gap} = G + Ln$$

		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.

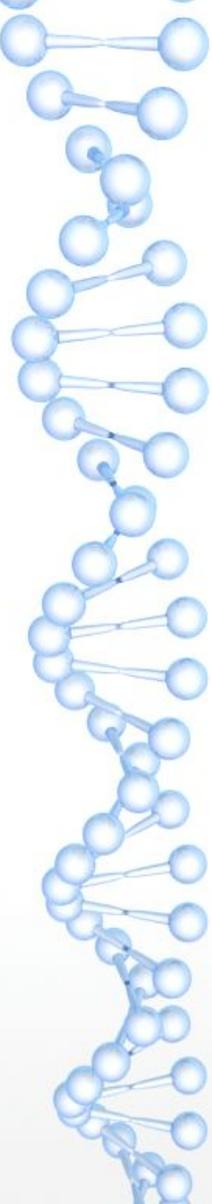
Basic local alignment search tool.

Altschul SF¹, 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

<i>Program</i>	<i>Query Sequence</i>	<i>Target Sequence</i>
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

Query Word ($W = 3$)

Query: GSQSLAALLNKCKT **PQG** QRLVNQWIKQPLMDKNRIEERLNLVEAFVED

*Neighborhood
Words*

PQG	18	= 7 + 5 + 6
PEG	15	
PRG	14	
PKG	14	
PNG	13	
PDG	13	
PHG	13	
PMG	13	
PSG	13	
PQA	12	
PQN	12	
etc.		

*Neighborhood Score
Threshold
($T = 13$)*

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	SLAALLNKCKT PQG QRLVNQWIKQPLMDKNRIEERLNLVEA	365
		+LA++L TP G R++ +W+ +P+ D + ER + A	
Sbjct:	290	TLASVLDCTVT PMG SRMLKRWLHMPVRDTRVLLERQQTIGA	330


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

Score = 91

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

Score = 138

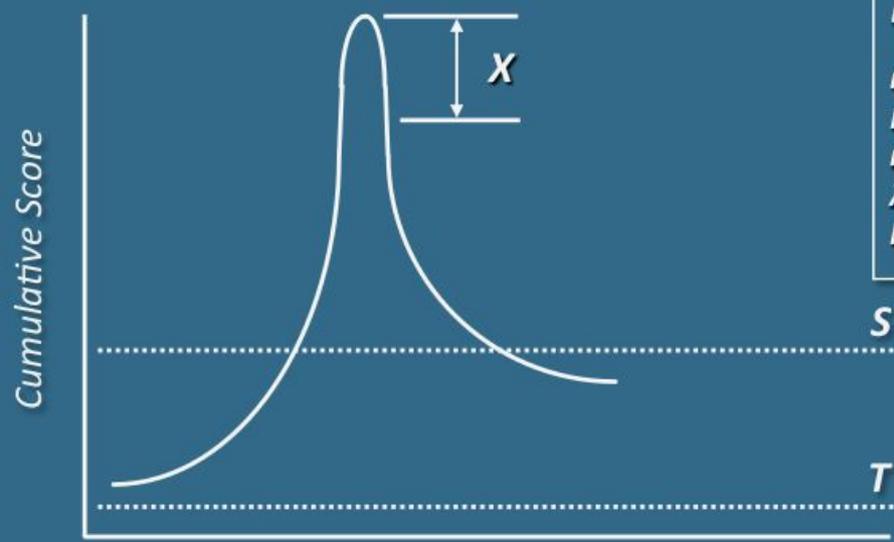
Scores and Probabilities

← [] →

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

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

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



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

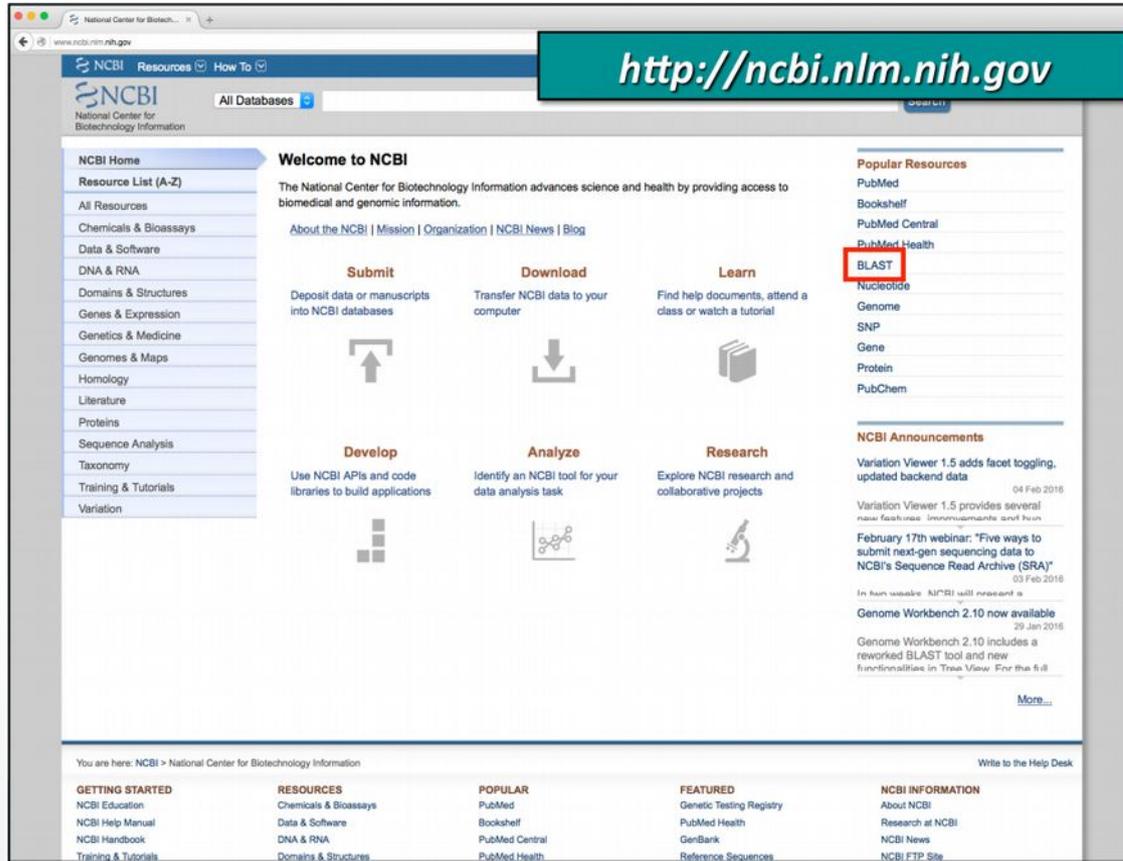
- m # letters in query
- N # letters in database
- mN size of search space
- λS normalized score
- k minor constant

S
Number of HSPs found purely by chance

T
Lower values signify higher similarity

Using Blast for protein similarity searches

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



The image shows a screenshot of the National Center for Biotechnology Information (NCBI) website. The URL `http://ncbi.nlm.nih.gov` is highlighted in a green box at the top. The page features a navigation menu on the left, a central content area with various service icons (Submit, Download, Learn, Develop, Analyze, Research), and a right-hand sidebar with 'Popular Resources' and 'NCBI Announcements'. The 'BLAST' link under 'Popular Resources' is highlighted with a red rectangular box.

Popular Resources

- PubMed
- Bookshelf
- PubMed Central
- PubMed Health
- BLAST**
- Nucleotide
- Genome
- SNP
- Gene
- Protein
- PubChem

NCBI Announcements

- Variation Viewer 1.5 adds facet toggling, updated backend data (04 Feb 2015)
- Variation Viewer 1.5 provides several new features: immunoassays and fun (04 Feb 2015)
- February 17th webinar: "Five ways to submit next-gen sequencing data to NCBI's Sequence Read Archive (SRA)" (03 Feb 2015)
- In two weeks, NCBI will present a (03 Feb 2015)
- Genome Workbench 2.10 now available (29 Jan 2015)
- Genome Workbench 2.10 includes a reworked BLAST tool and new functionalities in Tree View. For the full (29 Jan 2015)

[More...](#)

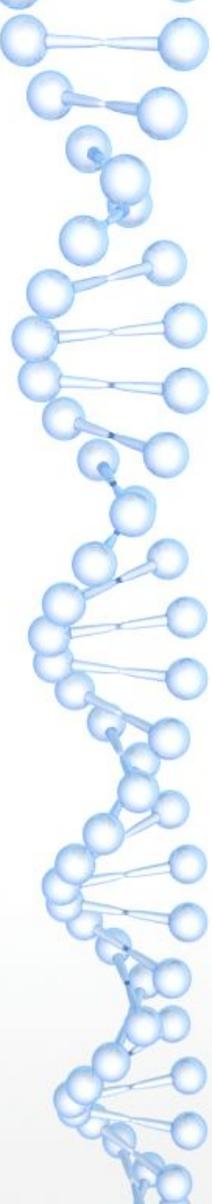
Available protein databases include:

<i>nr</i>	<i>Non-redundant</i>
<i>refseq</i>	<i>Reference Sequences</i>
<i>swissprot</i>	<i>SWISS-PROT</i>
<i>pat</i>	<i>Patents</i>
<i>pdb</i>	<i>Protein Data Bank</i>
<i>env_nr</i>	<i>Environmental samples</i>

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:

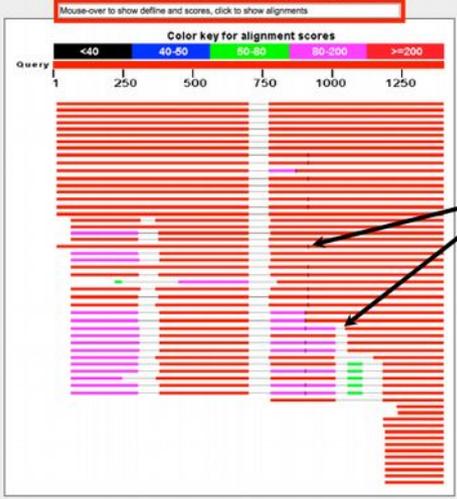
NT_	Genomic contigs
NM_	mRNAs
NP_	Proteins
NR_	Non-coding transcripts

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>

Color key



Gap in alignment with subject

- >1 HSP
- Masked region

Descending score order

0.0 means $\leq 10^{-1000}$

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

prosporo, isoform L [Drosophila melanogaster]
 Sequence ID: [ref|NP_788636.3|](#) Length: 1374 Number of Matches: 2
[See 2 more title\(s\)](#)

Range 1: 17 to 704 GenPept Graphics

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

Query 17 LFPQPSVSTAnssssnnnnssTPAALATHptnspsvsgssaslltaeFGNLFGGSSA 76
 LFPQPSVSTANSSNNNNSSTPAALATHSPNSPVGASASLLTAAGNLFGGSSA
 Sbjct 17 LFPQPSVSTANSSNNNNSSTPAALATHSPNSPVGASASLLTAAGNLFGGSSA 76

Query 77 KMLNELFGRQMKQADATSGLPQSLDNAMLAAMETATSSELLIGLSNSTKLLQQHNN 136
 KMLNELFGRQMKQADATSGLPQSLDNAMLAAMETATSSELLIGLSNSTKLLQQHNN
 Sbjct 77 KMLNELFGRQMKQADATSGLPQSLDNAMLAAMETATSSELLIGLSNSTKLLQQHNN 136

Query 137 NSIAPANSTPMSNGTNaaiapgsahssshhgqvspKGSRRVSACSDRSLEAAAADVAGG 196
 NSIAPANSTPMSNGTNASIFSGAHSSSHSHQGVSPKGSRRVSACSDRSLEAAAADVAGG
 Sbjct 137 NSIAPANSTPMSNGTNASIFSGAHSSSHSHQGVSPKGSRRVSACSDRSLEAAAADVAGG 196

Query 197 SPPRAASVSLNGGASGEGHQSLQHDLVAHHMLRNLQGGKELMQLDQLRTAMqqqq 256
 SPPRAASVSLNGGASGEGHQSLQHDLVAHHMLRNLQGGKELMQLDQLRTAMQ000
 Sbjct 197 SPPRAASVSLNGGASGEGHQSLQHDLVAHHMLRNLQGGKELMQLDQLRTAMQ000 256

Query 257 qqlqekeqlHSKLnannnnniiaatannnnntMESINLIDDEMAIDIKSEPTAPQPQ 316
 QQLQEKEQLHSKLNNNNNNI AATANNNNNTMESINLIDDEMAIDIKSEPTAPQPQ
 Sbjct 257 QQLQEKEQLHSKLNNNNNNI AATANNNNNTMESINLIDDEMAIDIKSEPTAPQPQ 316

Identities:
 ≥ 25% for proteins
 ≥ 70% for nucleotides

a Low complexity

Second HSP identified

Range 2: 777 to 1374 GenPept Graphics

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

Query 777 HVATAAPRQMHHPARLPTRMGGAAGHTALKSELSEKQMLRANNNSMMRMSGTDLE 836
 HVATAAPRQMHHPARLPTRMGGAAGHTALKSELSEKQMLRANNNSMMRMSGTDLE
 Sbjct 777 HVATAAPRQMHHPARLPTRMGGAAGHTALKSELSEKQMLRANNNSMMRMSGTDLE 836

Query 837 GLADVLSKSEITTSLSALVDTIVTRFVHQRRFLSKQADSVTAAEQNKDILLASQILDRK 896
 GLADVLSKSEITTSLSALVDTIVTRFVHQRRFLSKQADSVTAAEQNKDILLASQILDRK
 Sbjct 837 GLADVLSKSEITTSLSALVDTIVTRFVHQRRFLSKQADSVTAAEQNKDILLASQILDRK 896

Query 897 SPRTKADRPNQNGPTPATQSAAMFQAPKTPQGMNPFVAAAALYNSMTGPFCLPPDqqqqq 956
 SPRTKADRPNQNGPTPATQSAAMFQAPKTPQGMNPFVAAAALYNSMTGPFCLPPD00000
 Sbjct 897 SPRTKADRPNQNGPTPATQSAAMFQAPKTPQGMNPFVAAAALYNSMTGPFCLPPD00000 956

Query 957 qtaqqqqsaqqqqqsaqqqqLEQNEALSIVVTPKKRHKVTDTRIPTVSRILAQDG 1016
 QTAQQQSAQQQQSSAQTTQQLLEQNEALSIVVTPKKRHKVTDTRIPTVSRILAQDG
 Sbjct 957 QTAQQQSAQQQQSSAQTTQQLLEQNEALSIVVTPKKRHKVTDTRIPTVSRILAQDG 1016

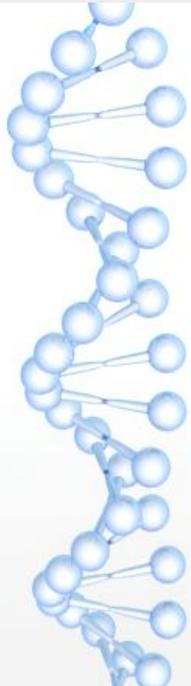
Query 1017 vvpptggpstpqqqqqqqqqqqqqqqqqqqqASNGGNSNATPAQSPTRSSGGAAYHpp 1076
 VVPPTGGPSTPQQQQQQQQQQQQQQQQQQQQASNGGNSNATPAQSPTRSSGGAAYHPQP
 Sbjct 1017 VVPPTGGPSTPQQQQQQQQQQQQQQQQQQQQASNGGNSNATPAQSPTRSSGGAAYHPQP 1076

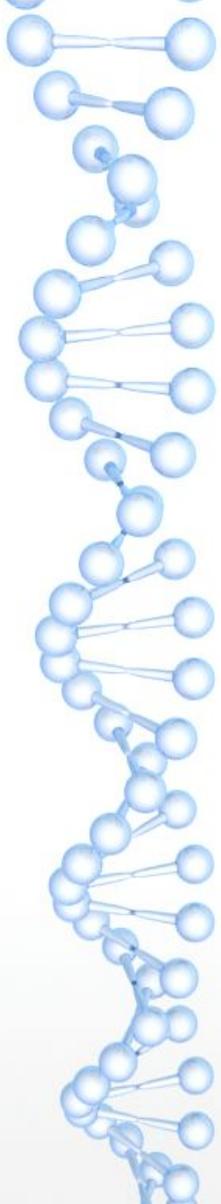
Query 1077 ppppppppVSLPTSAI PNPFLSHESKVFSPYSPFFPPhaaagqataaqlhghhqqhhph 1136
 PpppppppVSLPTSAI PNPFLSHESKVFSPYSPFFPPhaaagqataaqlhghhqqhhph
 Sbjct 1077 PpppppppVSLPTSAI PNPFLSHESKVFSPYSPFFPPhaaagqataaqlhghhqqhhph 1136

Query 1137 hqsmqlssppgsLgALMDSRDEpplphppsmLhpallaaahggspDYKTCRLRAVDAQ 1196
 HQSMQLSSPPGSLGALMDSRDEPPLPHPPSMLHPALLAAAHGGSPDYKTCRLRAVDAQ
 Sbjct 1137 HQSMQLSSPPGSLGALMDSRDEPPLPHPPSMLHPALLAAAHGGSPDYKTCRLRAVDAQ 1196

Query 1197 DRQSECSADMFQDGNAPTIFSYKQMLKTEHQESLMKHCESFPLHSSTLTPMHLRKA 1256
 DRQSECSADMFQDGNAPT-----SSTLTPMHLRKA
 Sbjct 1197 DRQSECSADMFQDGNAPT-----SSTLTPMHLRKA 1227

- Gap





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

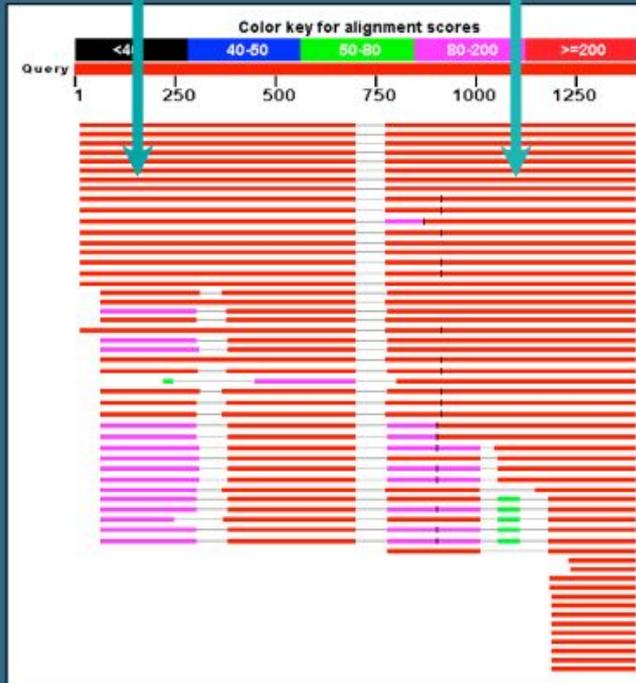
Q: 17- 704

S: 17- 704

HSP 2

Q: 777-1403

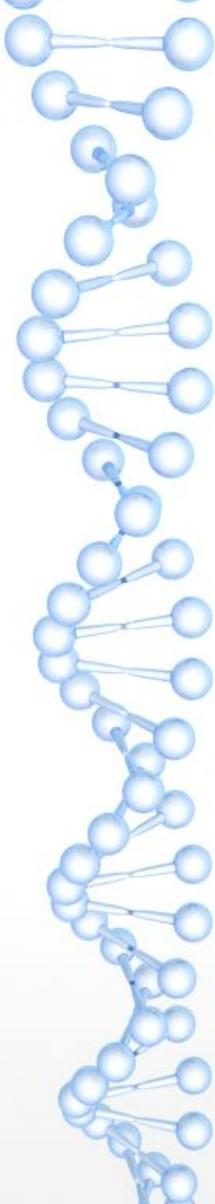
S: 777-1374





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



Protein BLAST: Align two or...
blast.ncbi.nlm.nih.gov/blast.cgi?PAGE=Proteins&PROGRAM=blastp&BLAST_PROGRAMS=blastp&PAGE_TYPE=blastSearch&BLAST_SPEC=

BLAST®

Basic Local Alignment Search

Home Recent Results Saved Strategies Help

NCBI/ BLAST/ blastp suite **Align Sequences**

blastn **blastp** blastx tblastn tblastx

Enter Query Sequence

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

```
>NP_008872.1 SOX-10 [Homo sapiens]  
MAEEQDLSEVELSPVGSSEPRCLSPGSAFSLGPDGGGGGSLRASPGPGLGKVKKEQQDGEADDDKFPV  
CI REAVSQVLSGYDWTLPMPVYVNGASKSKPHVVRPMNAPMVWAQAARRKLADQYPHLHNAELSKTLGK  
LWRLNESDKRPFIEEAERLRMQHKKDHPDYKQPRRRKNGAAQGEAECPGGEAEGGTAAIQAHYKSA  
HLDRHPGEGSFMSDGNPEHPGSGSHGPTPTPTPKTELQSGKADPKRGRSMGEGGKPHIDFGNVDIGP
```

Or, upload file No file selected.

Job Title
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]  
MQSYASAMLSVFNSSDDYSPAVQENIPALRRSSFLCTESCSNSKYQCETGENSKGNVQDRVSRPMNAFIVW  
SRDQRRKMALENPRMRNSEISKQLGYQWKMLTEAKWPFQEAQKLQAMHREKYPNYKYRPRRKAAMLPK  
NCSLLPADPASVLCSEVQLDNRLYRDDCTKATHSRMEHQLGHLPPINAASSPQQRDRYSHWTKL
```

Or, upload file 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 highlighted**

General Parameters

Max target sequences
Select the maximum number of aligned sequences to display

Short queries Automatically adjust parameters for short input sequences

Expect threshold

Word size

Max matches in a query range

Scoring Parameters

Matrix ← PAM30
PAM70
PAM250
BLOSUM80
BLOSUM62
BLOSUM45
BLOSUM50
BLOSUM90

Gap Costs Existence: 11 Extension: 1

Compositional adjustments Conditional compositional score matrix adjustment

Filters and Masking

Filter Low complexity regions

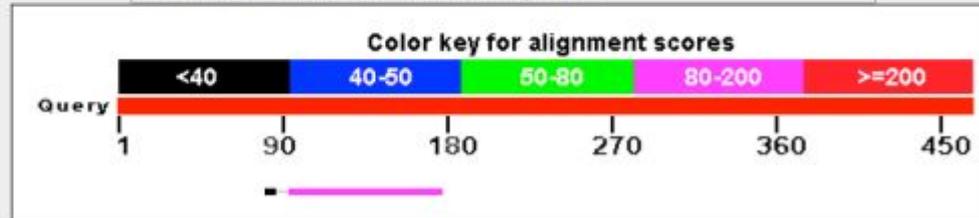
Mask Mask for lookup table only
 Mask lower case letters

BLAST Search protein sequence using Blastp (protein-protein BLAST)
 Show results in a new window

Graphic Summary

Distribution of 2 Blast Hits on the Query Sequence

Mouse over to see the define, click to show alignments



NP_003131.1 sex determining region Y [Homo sapiens]

Sequence ID: lc|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	NGASKSKPHVKRPMNAPMVWAQAARRKLADQYPHLHNAELSKTLGKLWRLLNESDKRPFI	154
		N + VKRPMNAP+VW++ RRK+A + P + N+E+SK LG W++L E++K PF	
Sbjct	51	NSKGNVQDRVKRPMNAFIVWSRDQRRKMALENPRMRNSEISKQLGYQWKMLTEAEKWPF	110
Query	155	EEAERLRMQHKKDHPDYKYQPRRR	178
		+EA++L+ H++ +P+YKY+PRR+	
Sbjct	111	QEAQKLQAMHREKYPNYKYPRRK	134

Range 2: 95 to 101 [Graphics](#)

[Next Match](#) [Previous Match](#) [First Match](#)

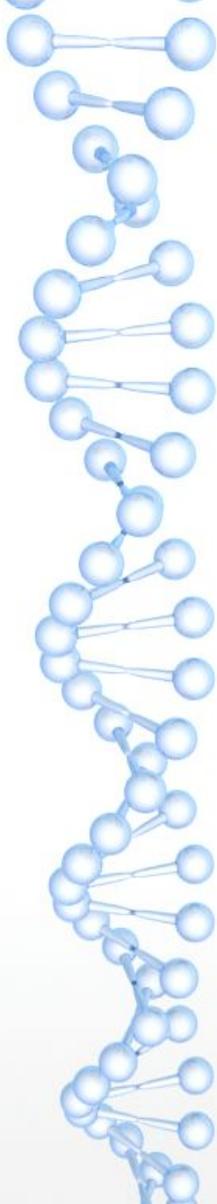
Score	Expect	Method	Identities	Positives	Gaps
15.4 bits(28)	1.9	Compositional matrix adjust.	3/7(43%)	5/7(71%)	0/7(0%)

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

Global alignment

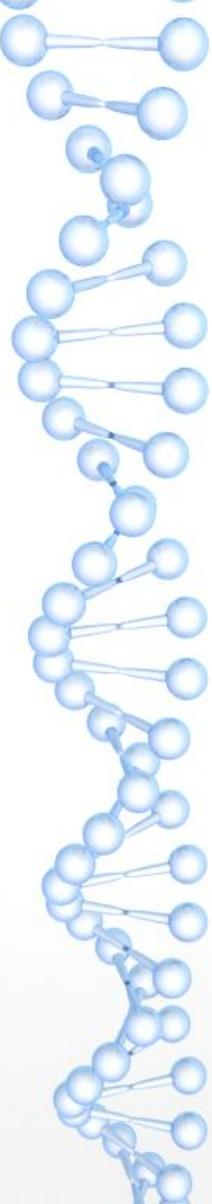
- https://www.ebi.ac.uk/Tools/psa/emboss_needle/

```
EMBOSS_001      1 MAEEQDLSEVELSPVGSEEPRLSPGSAPSLGPDGGGGGSLRASPGGE      50
EMBOSS_001      1 -----                                                0
EMBOSS_001     51 LGKVKKEQQDGEA-----DDDKFPV-----CIREAVSQVLSGYDWTLV      88
EMBOSS_001      1 -----MOSYASAMLSVFNSSDDYSPAVQENIPALRRSSSFLECTESCNSKY      44
EMBOSS_001     89 PMPVRVNGASKSKPHVKRPMNAFMVWAQAARRKLADQYPHLHNAELSKTL     138
EMBOSS_001     45 QCETGENSKGNVQDRVKRPMAFIVWSRDQRRKMALENPRMRNSEISKQL      94
EMBOSS_001    139 GKLWRLLNESDKRPFIEEAERLRMQHKKDHPDYKYQPRRRKNGKAAQGEA     188
EMBOSS_001     95 GYQWKMLTEAEKWPFQEAQKLQAMHREKYPNYKYRPRR----KAKMLPK     140
EMBOSS_001    189 ECPGGEAEQGGTAAIQAHYKSAHLDRH-HPGEGSPMSDGNPEHPSGQSHG     237
EMBOSS_001    141 NCSLLPADPASVLC-----SEVQLDNRLYRDDCTKATHSRMEHQLG--HL     183
EMBOSS_001    238 PP-TPPTTPK-----TELQSGKADPKRDGRSMGEGGKPHIDFGNVDI     278
EMBOSS_001    184 PPINAASSPQQRDRYSHWTKL-----                          204
EMBOSS_001    279 GEISHEVMSNMETFDDVAELDQYLPNGHPGHVSSYSAAGYGLGSALAVAS     328
EMBOSS_001    205 -----                                                204
EMBOSS_001    329 GHSAWISKPPGVALPTVSPPGVDAKAQVKTETAGPQGGPPHYTDQPSTSQI     378
EMBOSS_001    205 -----                                                204
EMBOSS_001    379 AYTSLSLPHYGSAFPISIRPQFDYSDHQPSGPPYGHSGQASGLYSAFSYM     428
EMBOSS_001    205 -----                                                204
EMBOSS_001    429 GPSQRPLYTAISDPSPSGQSHSPTHWEQPVYTTLSRP      466
EMBOSS_001    205 -----                                                204
```



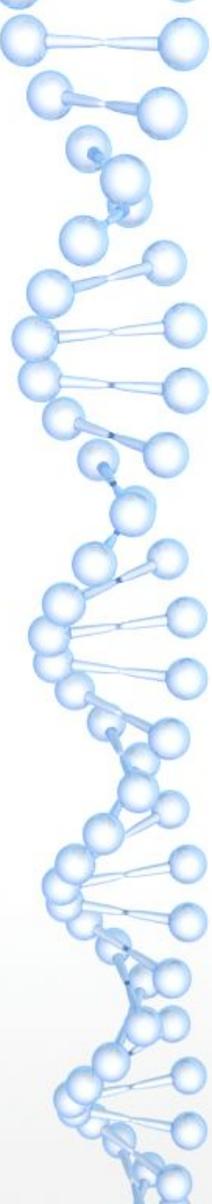
Nucleotide-Based BLAST Algorithms

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



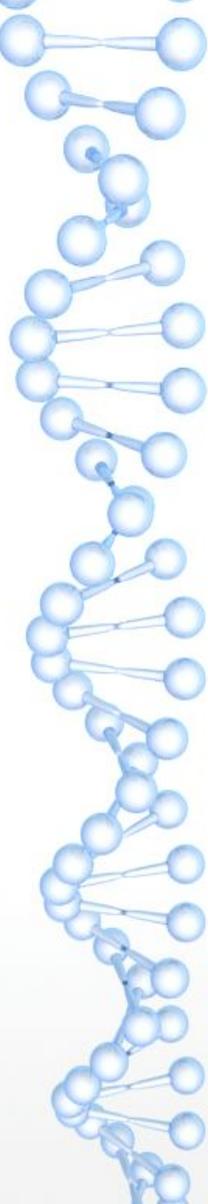
BLAT

- “BLAST-Like Alignment Tool”
- Designed to rapidly align longer nucleotide sequences ($L \geq 40$) having $\geq 95\%$ 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

- “BLAST-Like Alignment Tool”
- Designed to rapidly align longer nucleotide sequences ($L \geq 40$) having $\geq 95\%$ 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



UCSC Genome Browser HD...
genome.ucsc.edu

http://genome.ucsc.edu

UCSC Genome Bioinformatics

Genomes Genome Browser Tools Mirrors Downloads My Data Help About Us

Blat

About the UCSC Genome Bioinformatics Site

Welcome to the UCSC Genome Browser website. This site contains the reference sequence and working draft assemblies for a large collection of genomes. It also provides portals to [ENCODE](#) data at UCSC (2003 to 2012) and to the [Neanderthal](#) project. Download or purchase the Genome Browser source code, or the Genome Browser in a Box ([GBIB](#)) at our [online store](#).

BLAT Search Genome

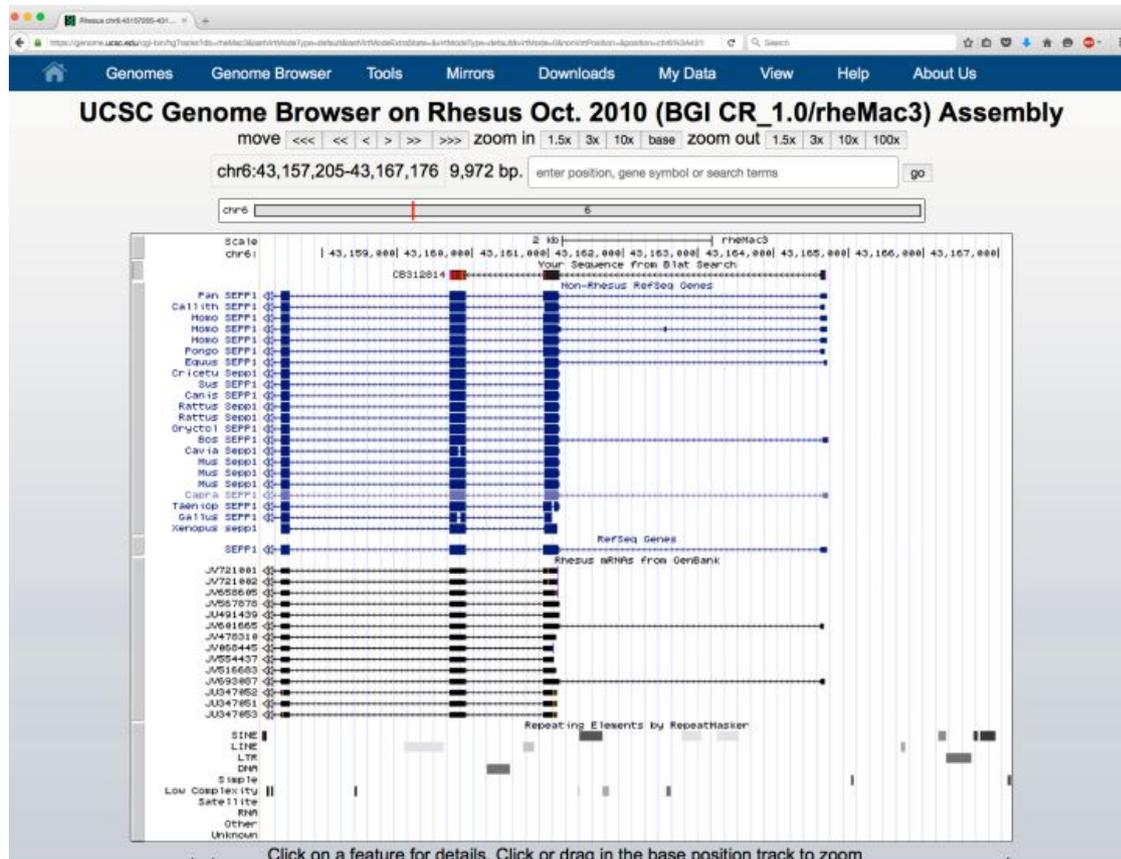
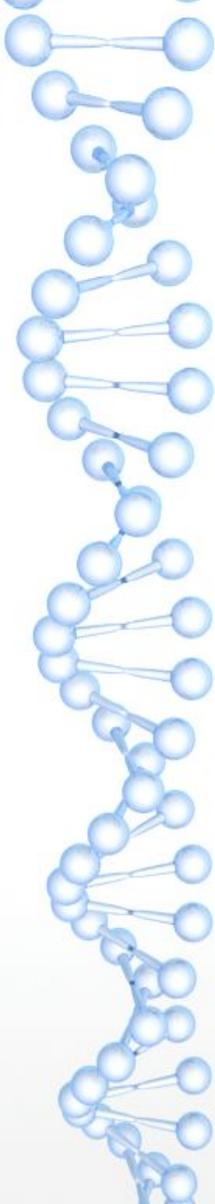
Genome: Rhesus Assembly: Oct. 2010 (BGI CR_1.0/rheMac3) Query type: DNA Sort output: query,score Output type: hyperlink

```
>CB312814 NICHD_Rh_Ovl_Macaca_mulatta_cDNA_clone
GGGGGTGGAGCTGCCAGAGTAAAGCAAAGAGCAAGGAAGCAGGCTCGTTGGAAGGGTGTGACAGCCCC
AGCAATGTGGAGAAGTCTGGGGCTTGCCTGGCTCTGTCTCTCCATCGGGAGGAACAGAGGCCAG
GACCAAAGCTCCTTCTGTAAGCAACCCAGCCTGGAGCATAAGAGATCAAGATCCAATGCTAGACTCCA
ATGGTTCAGTGACTGTGGTCTGCTTCTTCAAGCCAGCTGATACCTGTGCATACTGCANGCATCTAAATT
GGAAGAAGCTGCGAGTAAACTGGAGAAAGAAGGATATCTAAATATCCATATATGGTGGTAATCATCAA
GGGATCTCTTCTCGATTAAATACACACATCTTTAGAAAAAAGGTTTCAGAGCATATTCCTGTATATTCA
CCAGAAGAAAACCCACCGATGTCTGGACTCTTTAATGGAAACCAAGAAGACCTCCTCATATATGACGG
ATGTGGCCTTCTGGAAAACCCCTGGTGGCCTTTTCCTTCCCAACCTGGCGAATGGTAAAAAAACC
CCTTTAAATGGTTCGGGAAAAAAAAGTGGGAAATTTGGTCTCTCCCAAATCTCAAAAAAGAAAAA
TTTTGTAAAAAGGGATCTTTTGGGCACCGGGGAAAAAAAATTTGAAAACTTCCCCACCCCTT
TTCCCTCTTGGGGACTCCTTCCCAAATTCGGGGACATCCCCCT
```

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)



- **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.

