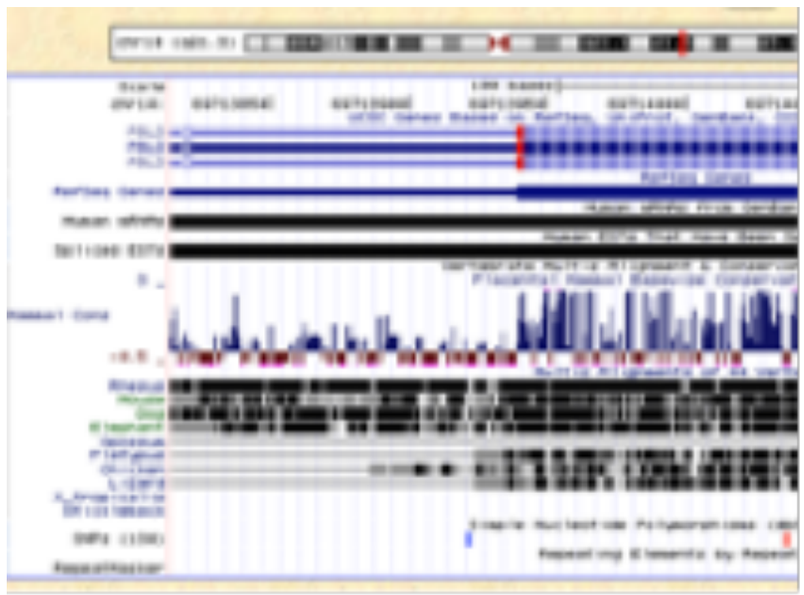
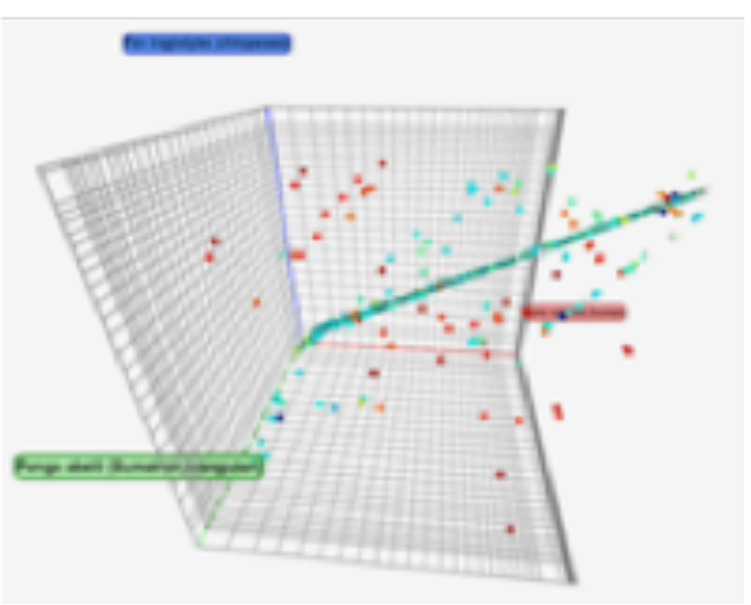
Computational Genomics

Introduction
to
Biological Sequence Analysis





NATIONAL HUMAN GENOME RESEARCH INSTITUTE Division of Intramural Research

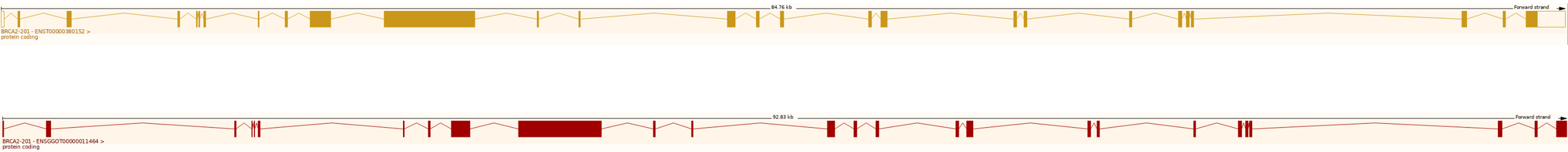
Current Topics in Genome Analysis 2016

Week 1: Biological Sequence Analysis I

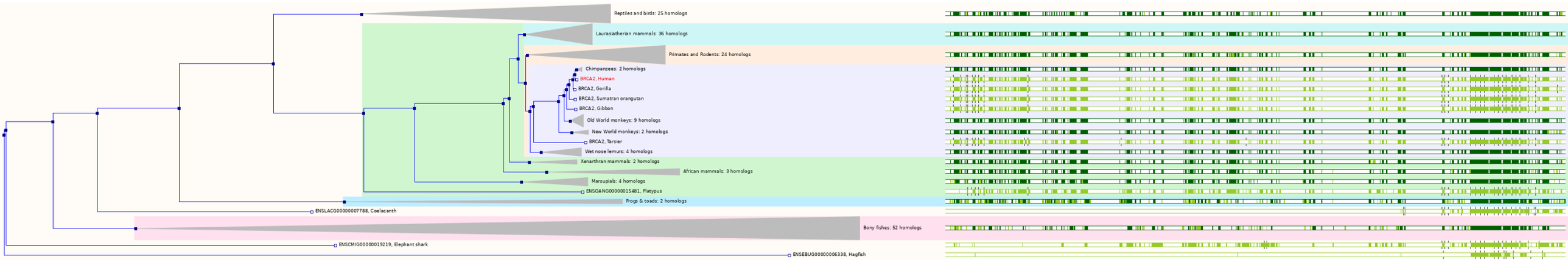
Andy Baxevanis, Ph.D.

Why Construct Sequence Alignments?

Why Construct Sequence Alignments?



To Determine Similarity and to Deduce Homology



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

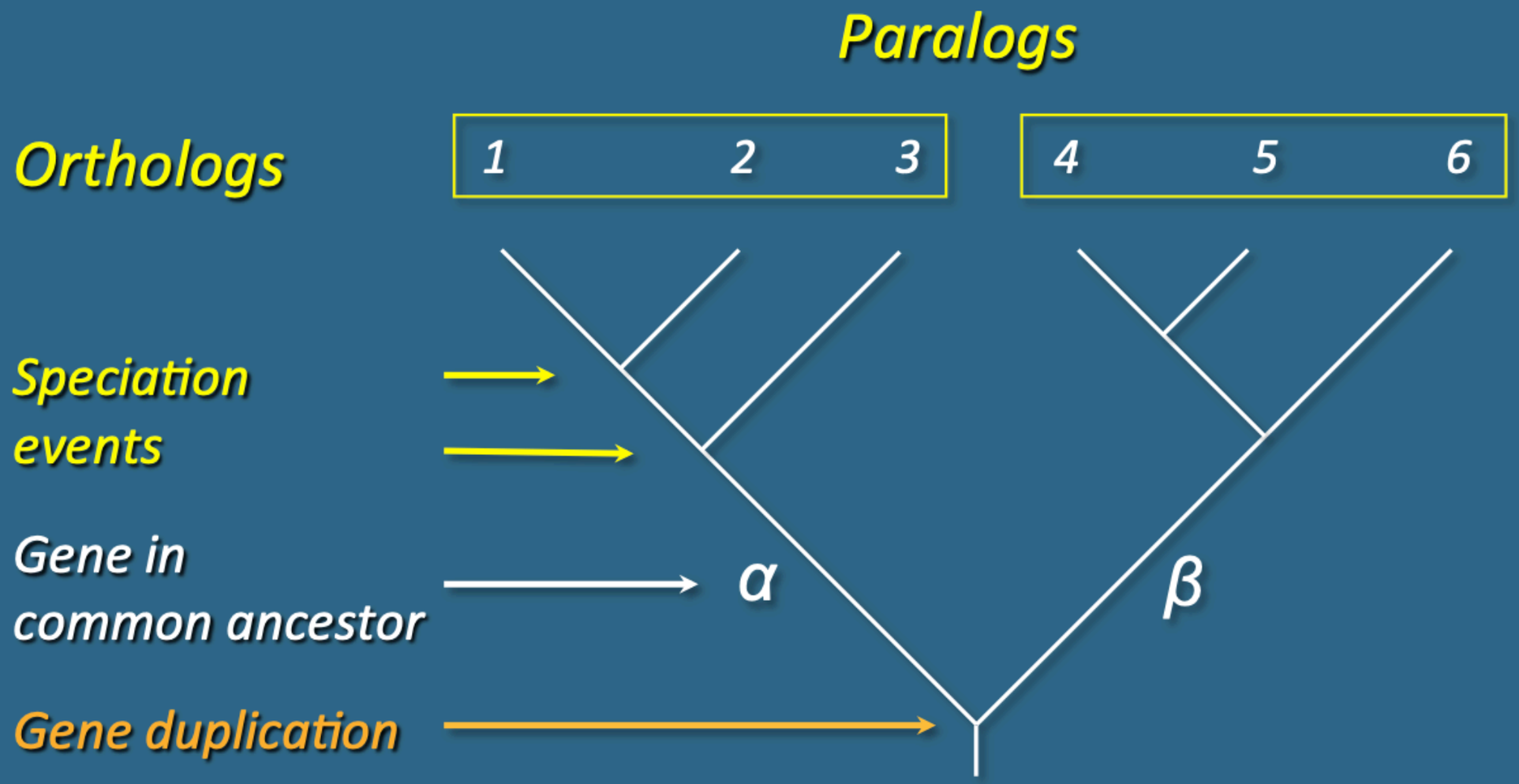
- Sequences are direct descendants of a sequence in a common ancestor (share a common origin)
- Most likely have similar domain and three-dimensional structure
- Usually retain same biological function over evolutionary time
- Can be used to predict gene function in novel genomes

Defining the Terms

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'

Defining the Terms



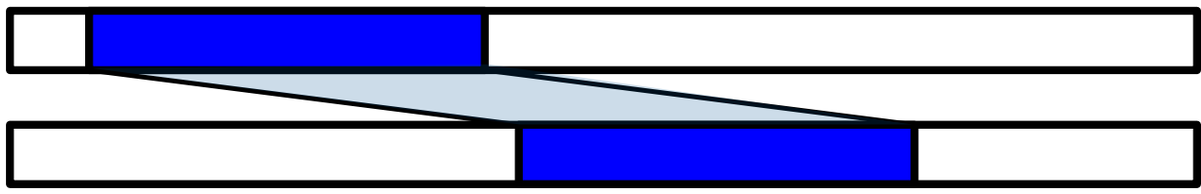
- Genes 1-3 are orthologous
- Genes 4-6 are orthologous
- Any pair of *α* and *β* genes are paralogous (genes related through a gene duplication event)

Kinds of Alignments

Global

```
5’–AGCTTTGACGAGCTCCTTACGTGGTACGTGCTAA–3
   ||  |||||||  |||||||||||| ||||||||
5’–AGAATTGACGATTTCCTTACGTGGTCCGTGCTAA–3’
```

Local



```
5’–AGCTTTGACGAGCTCCTTACGTGGTACGTGCTAA–3’
        |||||||  |||||||||
      5’–CTCCTTAGATGGTACGTG–3’
```

Kinds of Alignments

Global

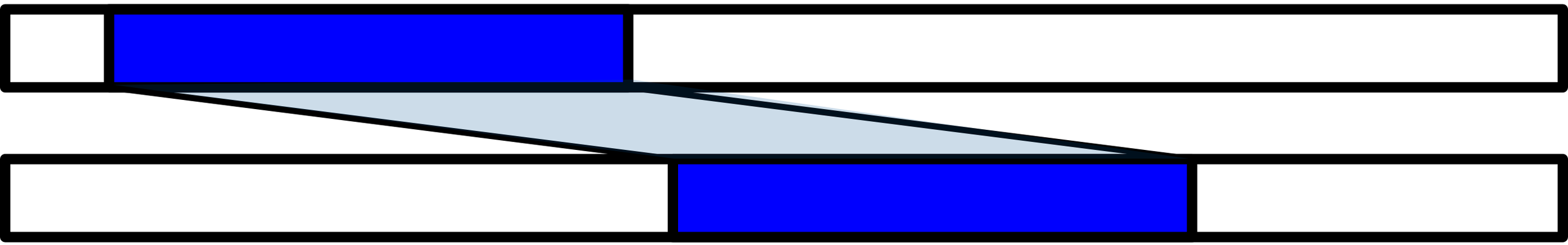
```
5’–AGCTTTGACGAGCTCCTTACGTGGTACGTGCTAA–3
   ||  |||||||  ||||||||||||| ||||||||
5’–AGAATTGACGATTTCCTTACGTGGTCCGTGCTAA–3’
```

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

Kinds of Alignments

Local



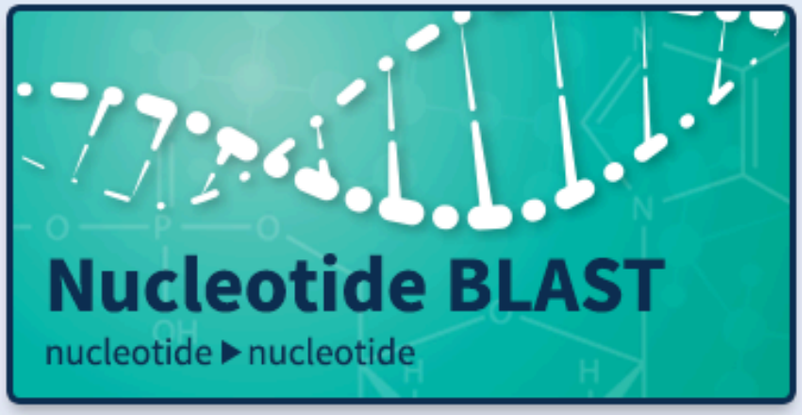
```
5’–AGCTTTGACGAGCTCCTTACGTGGTACGTGCTAA–3’
      |||||||  ||||||||||
    5’–CTCCTTAGATGGTACGTG–3’
```

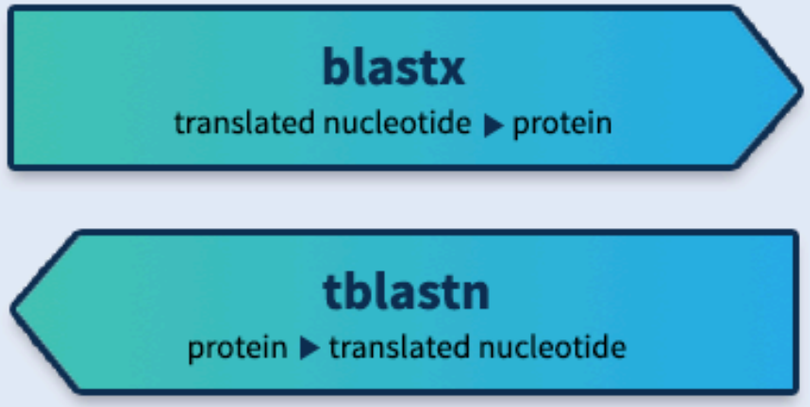
Local Sequence Alignments

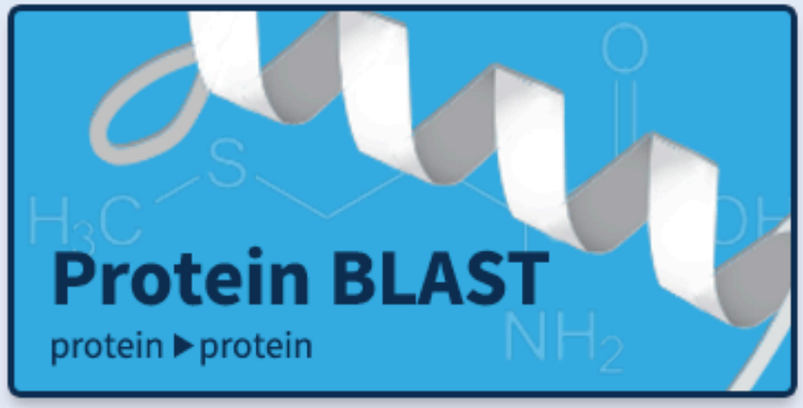
- Sequence comparison intended to find the most similar regions in the two sequences being aligned ('paired subsequences')
- Regions outside the area of local alignment are excluded
- More than one local alignment could be generated for any two sequences being compared
- Best for sequences that share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

Blast
Basic Local Alignment Search Tool

The Basic Local Alignment Search Tool (BLAST) finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.







Blast
Basic Local Alignment Search Tool

Specialized searches

SmartBLAST

Find proteins highly similar to your query

Primer-BLAST

Design primers specific to your PCR template

Global Align

Compare two sequences across their entire span (Needleman-Wunsch)

CD-search

Find conserved domains in your sequence

IgBLAST

Search immunoglobulins and T cell receptor sequences

VecScreen

Search sequences for vector contamination

CDART

Find sequences with similar conserved domain architecture

Multiple Alignment

Align sequences using domain and protein constraints

MOLE-BLAST

Establish taxonomy for uncultured or environmental sequences

Standalone and API BLAST

Download BLAST
Get BLAST databases and executables

Use BLAST API
Call BLAST from your application

Use BLAST in the cloud
Start an instance at a cloud provider

Blast
Basic Local Alignment Search Tool

>Query sequence
MSSAAAAAAGAAGGGALFQPQSVSTANSSSSNNNNSSTPAALATHSPTSNSPVSGASSASSLLTAAFGNL
FGGSSAKMLNELFGRQMKQAQDATSGLPQSLDNAMLAAAMETATSAELLIGSLNSTSKLLQQQHNNNSIA
PANSTPMSNGTNASISPGSAHSSSHSHQGVSPKGSRRVSACSDRSLEAAAADVAGGSPPRAASVSSLNGG
ASSGEQHQSQLQHDLVAHHMLRNILQGKKELMQLDQELRTAMQQQQQQLQEKEQLHSKLNNNNNNNIAAT
ANNNNNTTMESINLIDDSEMADIKIKSEPQTAPQPQQSPHGSSHSSRSGSGSGSHSSMASDGSLRRKSSD
SLDSHGAQDDAQDEEDAAPTGQRSESRAPEEPQLPTKKESVDDMLDEVELLGLHSRGSDMDSLASPSHSD
MMLLDKDDVLDEDDDDDCVEQKTSGSGCLKKPGMDLKRARVENIVSGMRCSPSSGLAQAGQLQVNGCKKR
KLYQPQQHAMERYVAAAAGLNFGLNLQSMMLDQEDSESNELESPQIQQKRVEKNALKSQLRSMQEQLAEM
QQKYVQLCSRMEQESECQELDQDQDVEQEQEPDNGSSDHIELSPSPTLTGDGDVSPNHKEETGQERPGSS
SPSPSPLKPKTSLGESSDSGANMLSQMMSKMMSGKLHNPLVGVGHPALPQGFPPLLQHMGDMSHAAAMYQ
QFFFEQEARMAKEAAEQQQQQQQQQQQQQQQQQEQQRRFEQEQQEQQRRKEEQQQQIQRQQQHLQQLQQ
QQMEQQHVATAAPRPQMHHPAPARLPTRMGGAAGHTALKSELSEKFQMLRANNNSSMMRMSGTDLEGLAD
VLKSEITTSLSALVDTIVTRFVHQRRLFSKQADSVTAAAEQLNKDLLLASQILDRKSPRTKVADRPQNGP
TPATQSAAAMFQAPKTPQGMNPVAAAALYNSMTGPFCLPPDQQQQQQTAQQQQSAQQQQQSSQQTQQQLE
QNEALSLVVTPKKKRHKVTDTRITPRTVSRILAQDGVVPPTGGPPSTPQQQQQQQQQQQQQQQQQQQQAS
NGGNSNATPAQSPTRSSGGAAYHPQPPPPPPPMMPVSLPTSVAIPNPSLHESKVFSPYSPFFNPHAAAGQ
ATAAQLHQHHQQHHPHHQSMQLSSSPPGSLGALMDSRDSPPLPHPPSMLHPALLAAAHHGGSPDYKTCLR
AVMDAQDRQSECNSADMQFDGMAPTISFYKQMQLKTEHQESLMAKHCESLTPLHSSTLTPMHLRKAKLMF
FWVRYPSSAVLKMYFPDIKFNKNNTAQLVKWFSNFREFYYIQMEKYARQAVTEGIKTPDDLLIAGDSELY
RVLNLHYNRNNHIEVPQNFRFVVESTLREFFRAIQGGKDTEQSWKKSIYKIISRMDDPVPEYFKSPNFLE
QLE

Blast
Basic Local Alignment Search Tool

Standard Protein BLAST

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

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

Enter Query Sequence

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

```
>Query sequence
MSSAAAAAAGAAGGGALFQPQSVSTANSSSSNNNNSSTPAALATHSPTSNS
PVSGASSASSLLTAAFGNL
FGGSSAKMLNELFGRQMKQAQDATSGLPQSLDNAMLAAAMETATSAELLIG
```

Query subrange

From

To

Or, upload file Choose File No file chosen

Job Title Query sequence

Enter a descriptive title for your BLAST search

Align two or more sequences

Choose Search Set

Databases Standard databases (nr etc.): New Experimental databases

Try experimental clustered nr database

For more info see What is clustered nr?

Compare Select to compare standard and experimental database

Experimental

Database Clustered nr(nr_clustered)

Organism Optional

Add organism

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown

Program Selection

Algorithm

- blastp (protein-protein BLAST)
- PSI-BLAST (Position-Specific Iterated BLAST)

Choose a BLAST algorithm

Blast
Basic Local Alignment Search Tool

— Algorithm parameters

General Parameters

Max target sequences: 10
Select the maximum number of aligned sequences to display

Short queries: ☑ Automatically adjust parameters for short input sequences

Expect threshold: 0.05

Word size: 6

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

Blast
Basic Local Alignment Search Tool

Clusters | Graphic Summary | Alignments | Taxonomy

Clusters producing significant alignments Download Select columns Show 10

select all *10 clusters selected* GenPept Graphics Distance tree of results Multiple alignment MSA Viewer

	New Cluster Composition Click the ⊞ to see the cluster contents	New Cluster Ancestor	Representative Sequence	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	⊞21 member(s), 11 organism(s)	flies	prospero, isoform H [Drosophila melanogaster]	2807	2807	100%	0.0	97.93%	1674	NP_001247044.1
☑	⊞3 member(s), 1 organism(s)	fruit fly	prospero, isoform L [Drosophila melanogaster]	2790	2790	100%	0.0	97.93%	1374	NP_788636.3
☑	⊞2 member(s), 1 organism(s)	fruit fly	homeodomain transcription factor Prospero [Drosophila melanogaster]	2751	2751	100%	0.0	96.72%	1403	AAF05703.1
☑	⊞3 member(s), 2 organism(s)	flies	homeobox protein prospero isoform X2 [Drosophila takahashii]	2107	2107	98%	0.0	91.17%	1698	XP_017008559.2
☑	⊞3 member(s), 2 organism(s)	flies	homeobox protein prospero isoform X2 [Drosophila kikkawai]	2008	2008	95%	0.0	91.39%	1690	XP_017016749.1
☑	⊞3 member(s), 2 organism(s)	flies	homeobox protein prospero isoform X2 [Drosophila bipectinata]	1960	1960	98%	0.0	86.75%	1728	XP_017090434.2
☑	⊞2 member(s), 2 organism(s)	flies	homeobox protein prospero isoform X5 [Drosophila pseudoobscura]	1746	1746	98%	0.0	76.98%	1694	XP_033232695.1
☑	⊞8 member(s), 2 organism(s)	flies	homeobox protein prospero isoform X4 [Drosophila guanche]	1721	1721	95%	0.0	78.78%	1672	XP_034131817.1
☑	⊞4 member(s), 1 organism(s)	flies	homeobox protein prospero isoform X3 [Drosophila busckii]	1588	1588	95%	0.0	72.48%	1657	XP_033149969.1
☑	⊞1 member(s), 1 organism(s)	flies	homeobox protein prospero isoform X4 [Drosophila virilis]	1578	1578	96%	0.0	69.97%	1739	XP_032289979.1

Blast
Basic Local Alignment Search Tool

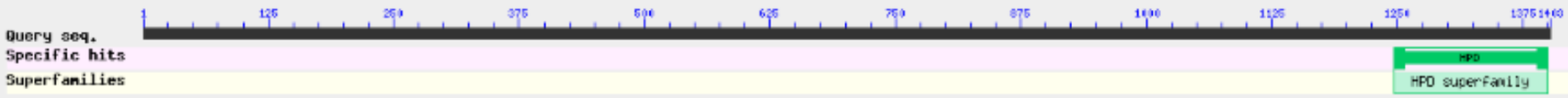
Clusters | **Graphic Summary** | Alignments | Taxonomy

hover to see the title *click to show alignments* *Show Conserved Domains*

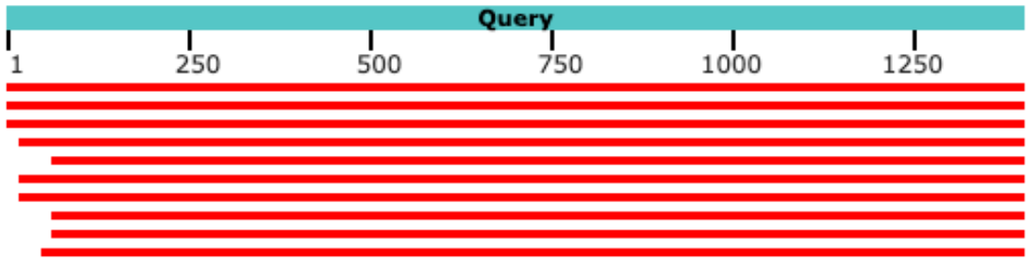
Alignment Scores < 40 | 40 - 50 | 50 - 80 | 80 - 200 | >= 200

10 clusters selected

Putative conserved domains have been detected, click on the image below for detailed results.



Distribution of the top 10 Blast Hits on 10 subject clusters



Blast
Basic Local Alignment Search Tool

Clusters | Graphic Summary | **Alignments** | Taxonomy

Alignment view: Pairwise | Restore defaults | Download

10 clusters selected

Download | GenPept | Graphics | ▼ Next | ▲ Previous | ◀ Descriptions

prospero, isoform H [Drosophila melanogaster]

Sequence ID: NP_001247044.1 **Length:** 1674 **Number of Matches:** 1

Range 1: 301 to 1674 GenPept | Graphics | ▼ Next Match | ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
2807 bits(7277)	0.0	Compositional matrix adjust.	1374/1403(98%)	1374/1403(97%)	29/1403(2%)

```
Query  1    MSSAAAAAAGAAGGGALFQPQSVSTANSSSSNNNNSSTPAALATHSPTSNSPVSGASSAS  60
            MSSAAAAAAGAAGGGALFQPQSVSTANSSSSNNNNSSTPAALATHSPTSNSPVSGASSAS
Sbjct  301  MSSAAAAAAGAAGGGALFQPQSVSTANSSSSNNNNSSTPAALATHSPTSNSPVSGASSAS  360

Query  61   SLLTAAFGNLFGGSSAKMLNELFGRQMKQAQDATSGLPQSLDNAMLAAAMETATSAELLI  120
            SLLTAAFGNLFGGSSAKMLNELFGRQMKQAQDATSGLPQSLDNAMLAAAMETATSAELLI
Sbjct  361  SLLTAAFGNLFGGSSAKMLNELFGRQMKQAQDATSGLPQSLDNAMLAAAMETATSAELLI  420

Query  121  GSLNSTSKLLQQQHNNNSIAPANSTPMSNGTNASISPGSAHSSSHSHQGVSPKGSRRVSA  180
            GSLNSTSKLLQQQHNNNSIAPANSTPMSNGTNASISPGSAHSSSHSHQGVSPKGSRRVSA
Sbjct  421  GSLNSTSKLLQQQHNNNSIAPANSTPMSNGTNASISPGSAHSSSHSHQGVSPKGSRRVSA  480

Query  181  CSDRSLEAAAADVAGGSPPRAASVSSLNGGASSGEQHQSQLQHDLVAHHMLRNILQGKKE  240
```

Related Information

Gene - associated gene details

Genome Data Viewer - aligned genomic context

Blast
Basic Local Alignment Search Tool

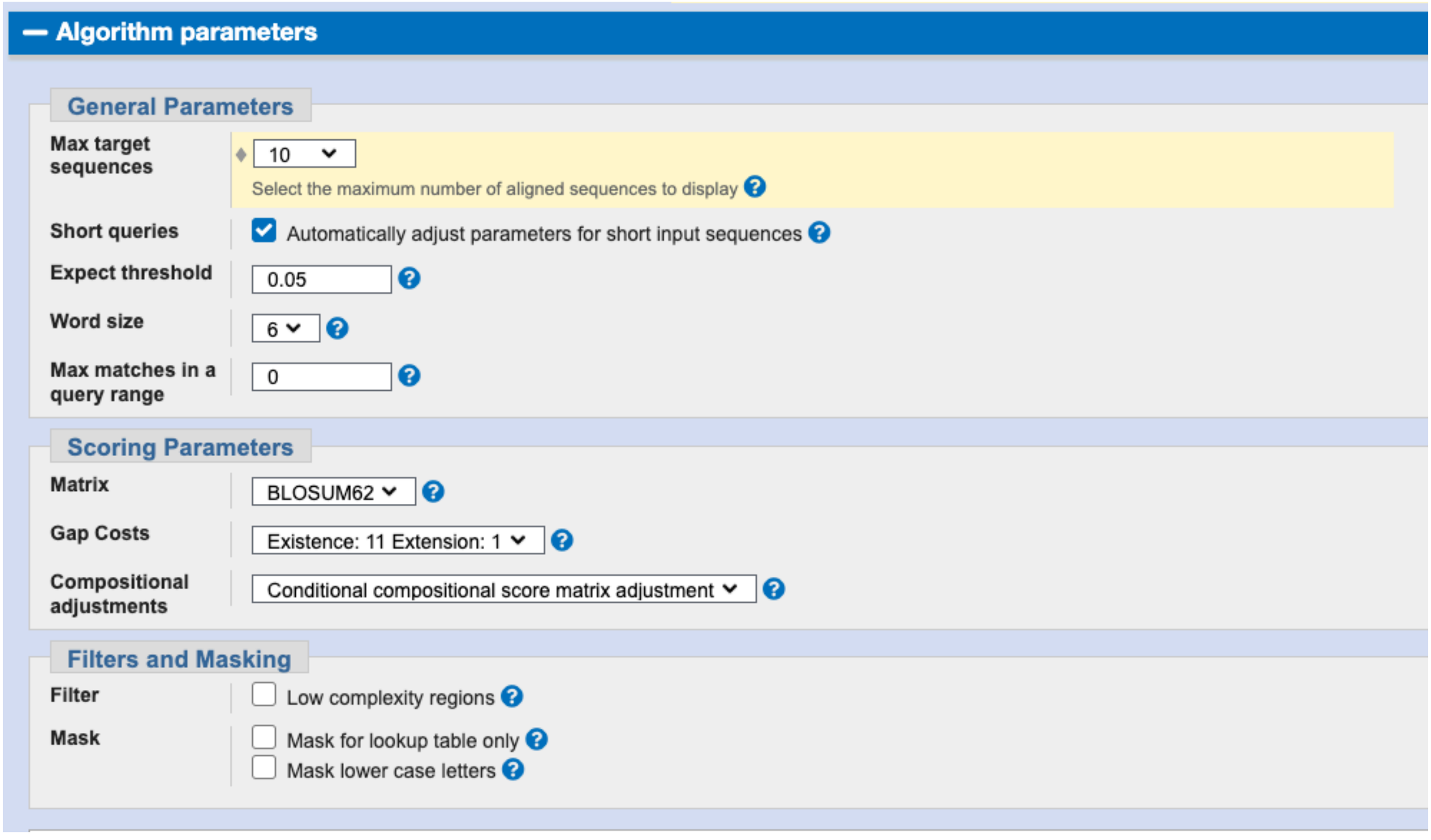
Clusters | Graphic Summary | Alignments | **Taxonomy**

Reports | **Lineage** | **Organism** | **Taxonomy**

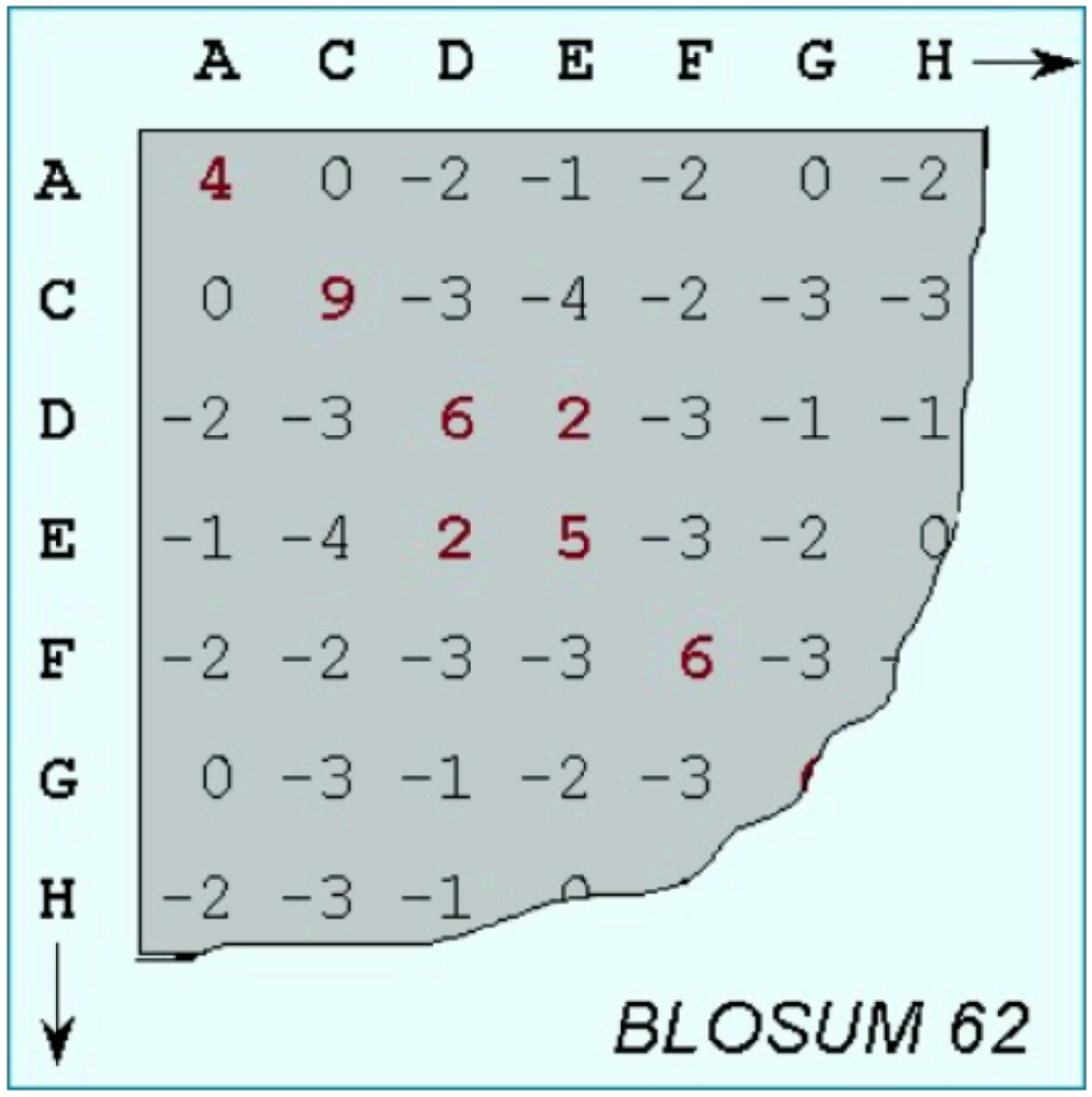
10 clusters selected

Organism	Blast Name	Score	Number of Hits	Description
Drosophila	flies		10	
. Sophophora	flies		8	
. . melanogaster group	flies		6	
. . . Drosophila melanogaster	flies	2807	3	Drosophila melanogaster hits
. . . Drosophila takahashii	flies	2107	1	Drosophila takahashii hits
. . . Drosophila kikkawai	flies	2008	1	Drosophila kikkawai hits
. . . Drosophila bipectinata	flies	1960	1	Drosophila bipectinata hits
. . Drosophila pseudoobscura	flies	1746	1	Drosophila pseudoobscura hits
. . Drosophila guanche	flies	1721	1	Drosophila guanche hits
. Drosophila busckii	flies	1588	1	Drosophila busckii hits
. Drosophila virilis	flies	1578	1	Drosophila virilis hits

About Configuring Blast Alignments



About Scoring Matrices



Link to original material

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?

Scoring Matrices

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	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

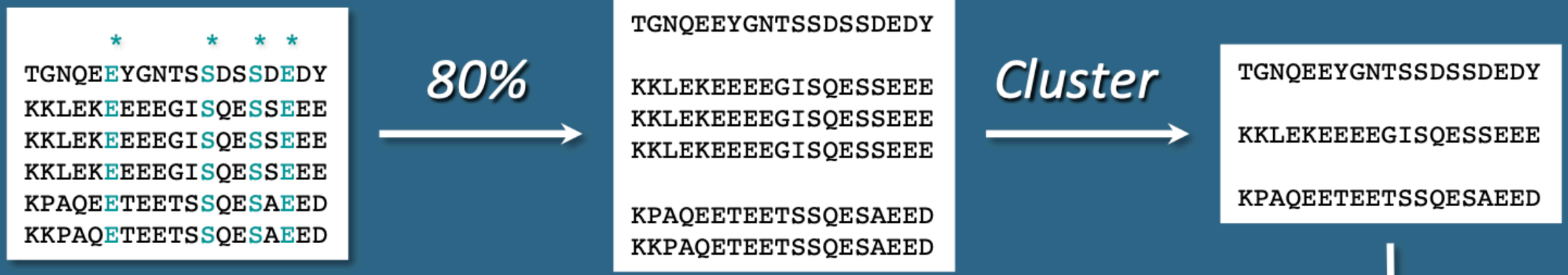
BLOSUM62

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



BLOSUM *n*

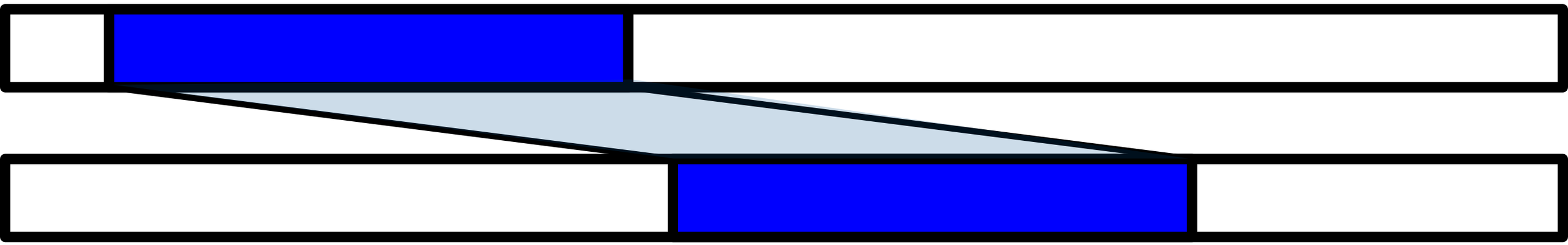
- Clustering reduces contribution of closely related sequences (less bias towards substitutions that occur in the most closely related members of a family)
- Reducing *n* yields more distantly related sequences
- Increasing *n* yields more closely related sequences

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

About Gaps

Local



5’–AGCTTTGACGAGCTCCTTACGTGGTACGTGCTAA–3’

5’–CTCCTTAGATGGTACGTG–3’

Gap

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*		

About Blast Algorithm

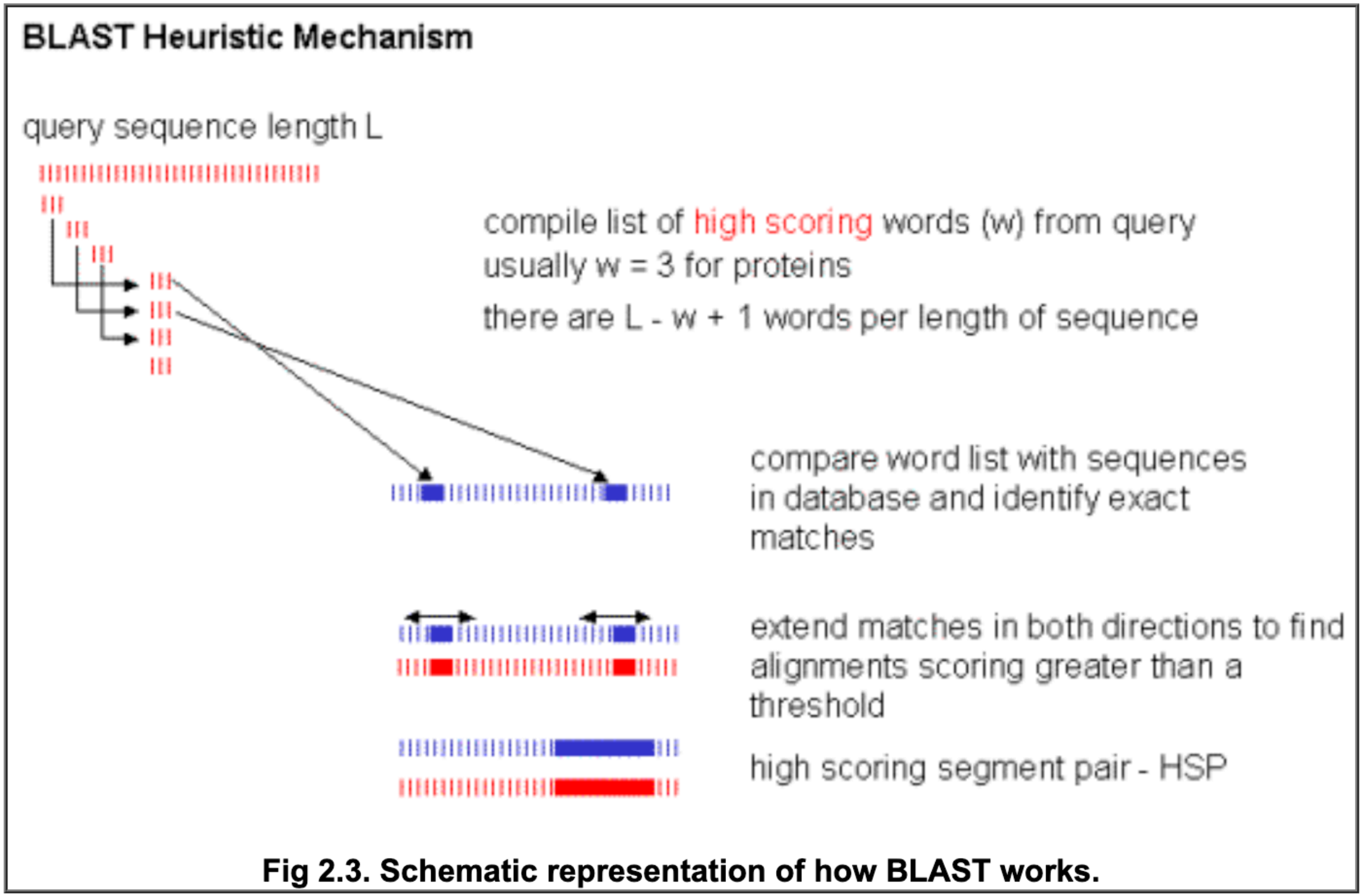


Fig 2.3. Schematic representation of how BLAST works.

Link to original material

BLAST

- 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

Altschul et al., J. Mol. Biol. 215: 403-410, 1990

Neighborhood Words

Query Word (W = 3)

Neighborhood Words

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

= 7 + 5 + 6

Neighborhood Score Threshold (T = 13)

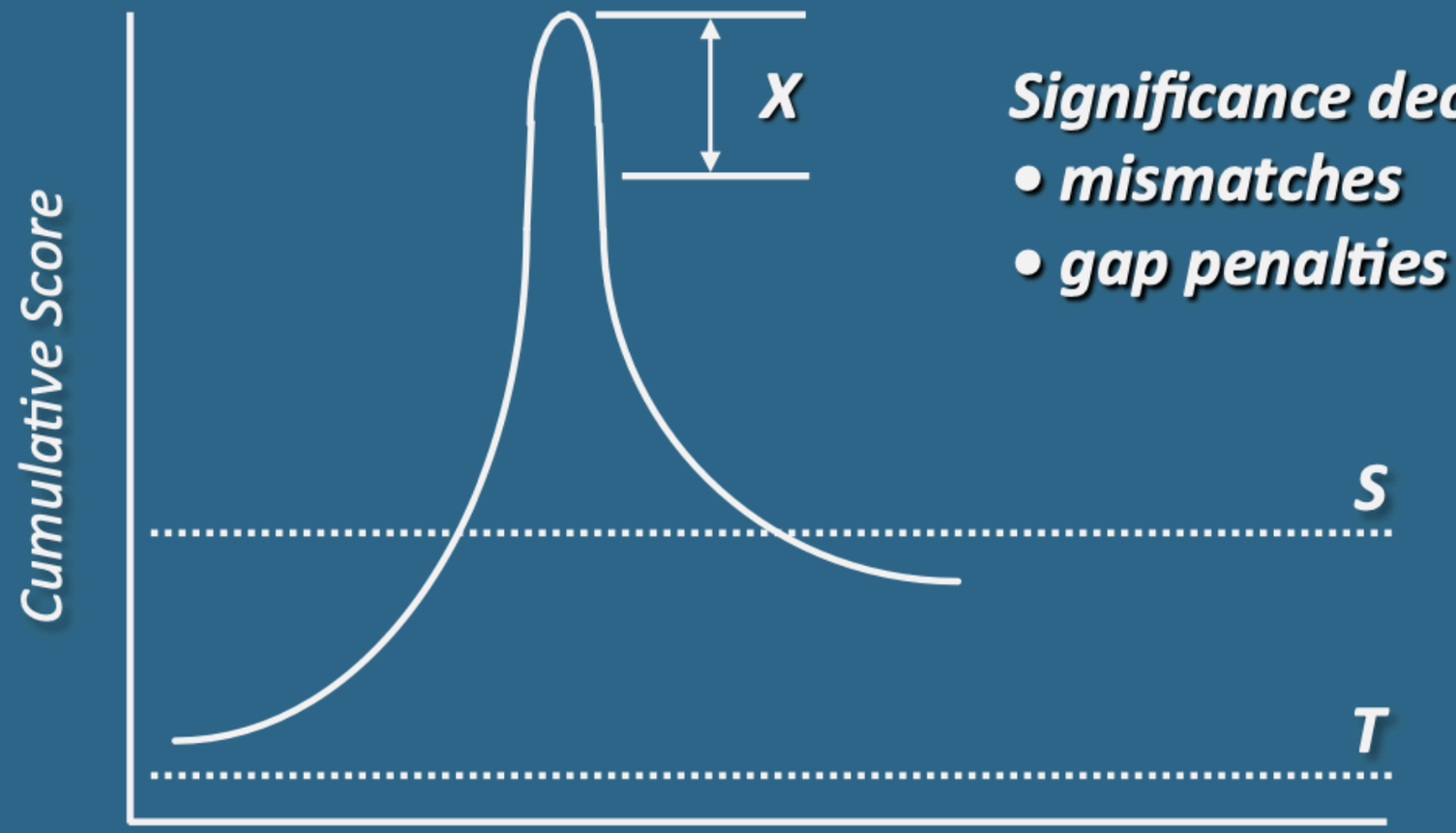
High-Scoring Segment Pairs

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  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
```

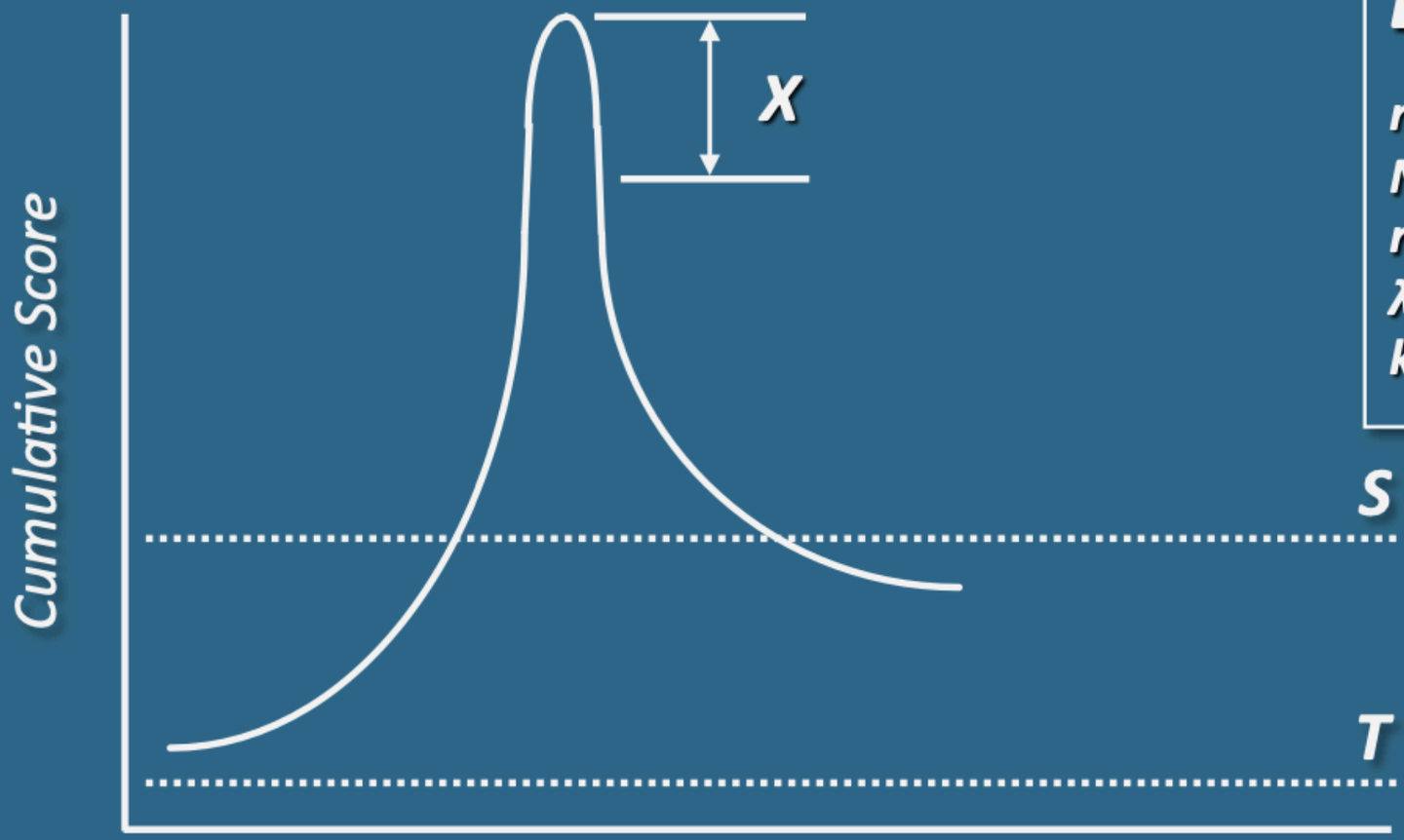
```
Score =   91
```

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

```
Score = 138
```

Scores and Probabilities

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

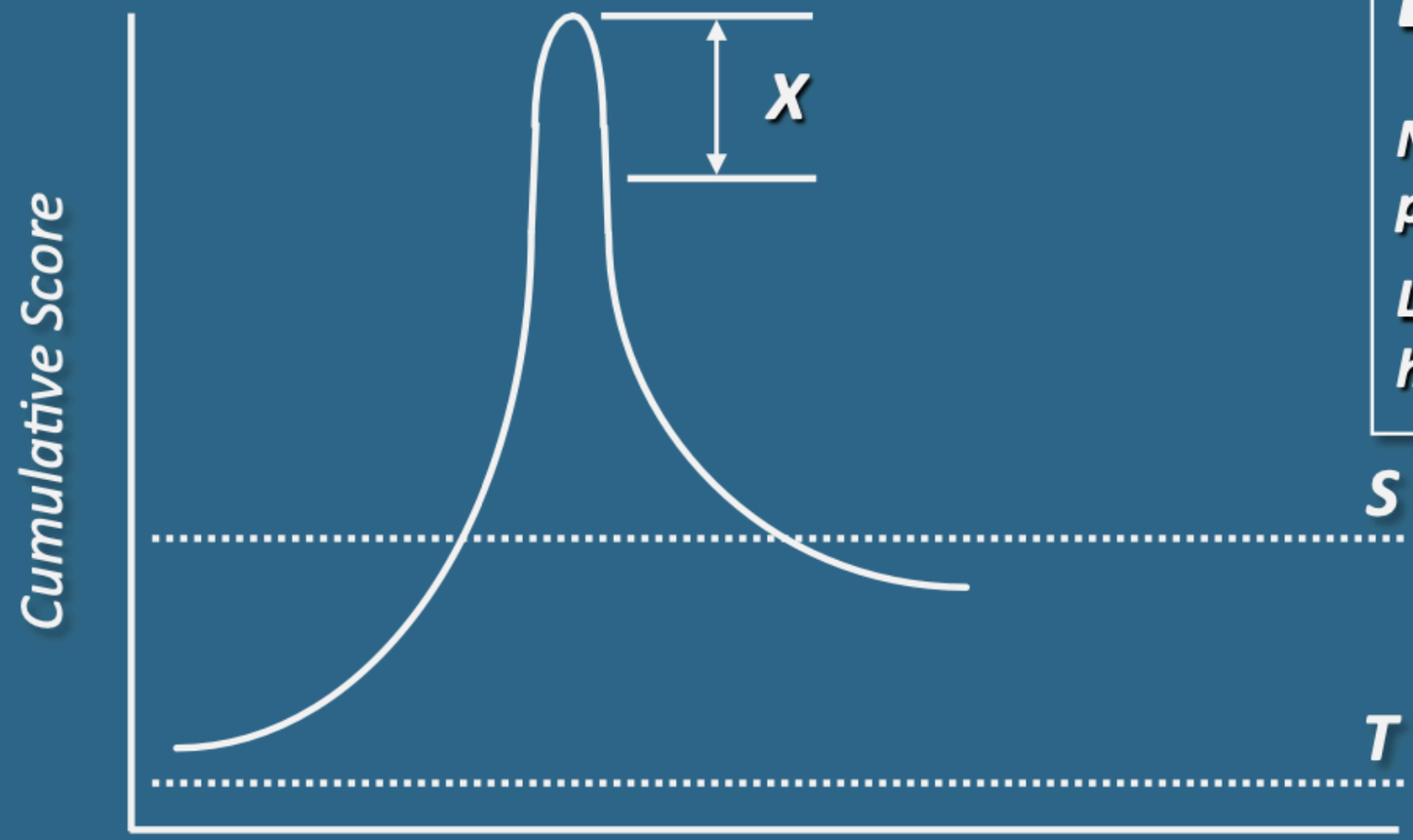


$$E = kmNe^{-\lambda S}$$

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

Scores and Probabilities

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



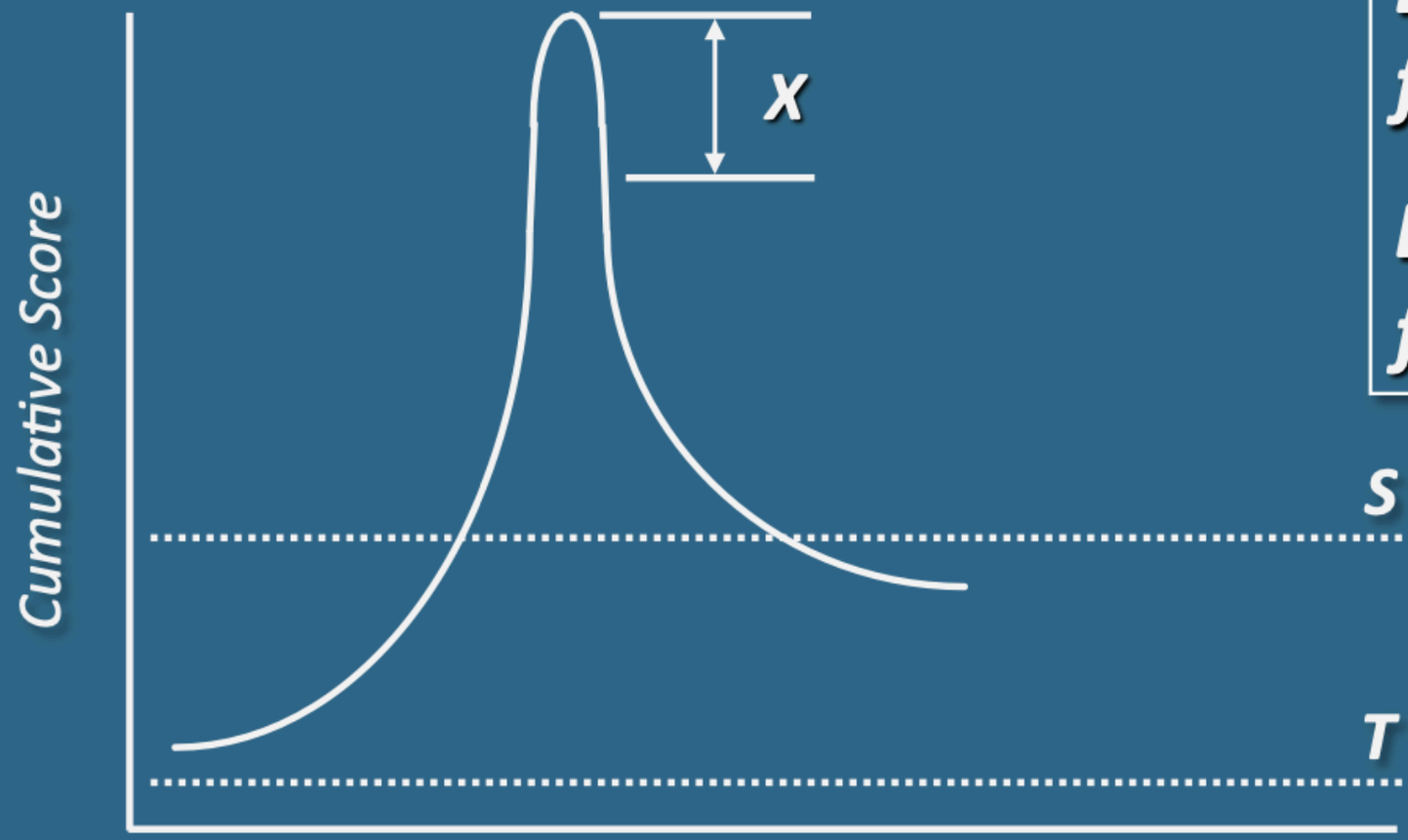
$$E = kmNe^{-\lambda S}$$

Number of HSPs found purely by chance

Lower values signify higher similarity

Scores and Probabilities

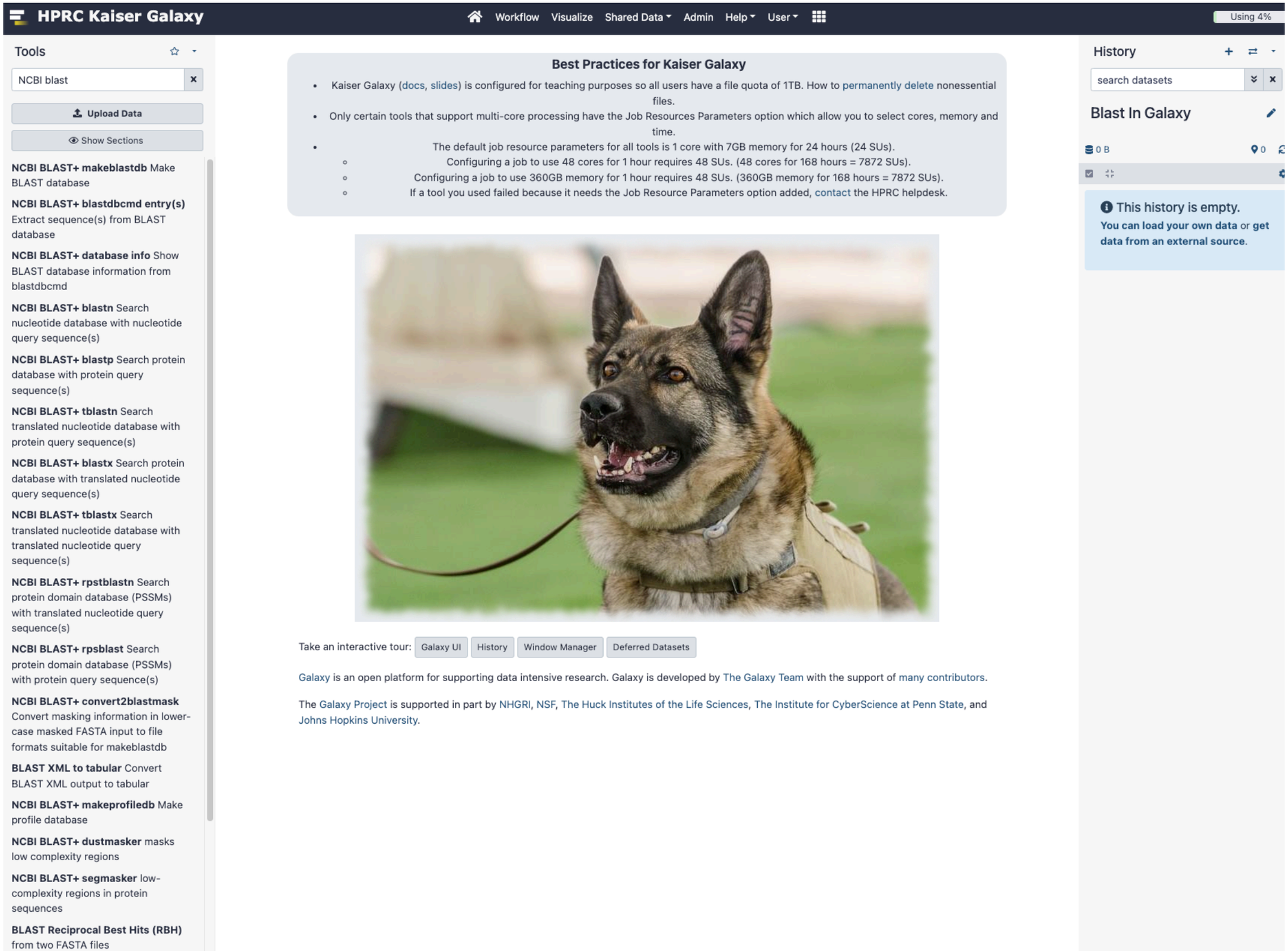
```
Query:    325   SLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEA   365
                +LA++L     TP G R++ +W+ +P+ D    + ER    + A
Sbjct:    290   TLASVLDCTVTPMGSRMLKRWLHMPVRDTRVLLERQQTIGA   330
```



E ≤ 10^{-6} for nucleotides

E ≤ 10^{-3} for proteins

About Blast In Galaxy



BLAST in Galaxy

Output formats of BLAST

- **First Option:**
  - Because Galaxy focuses on processing tabular data, the default output of this tool is tabular
  - The 12 column BLAST+ tabular output contains:

Column	NCBI Name	Description
1	qseqid	Query Seq-id (ID of your sequence)
2	sseqid	Subject Seq-id (ID of the database hit)
3	pident	Percentage of identical matches
4	length	Alignment length
5	mismatch	Number of mismatches
6	gapopen	Number of gap openings
7	qstart	Start of alignment in query
8	qend	End of alignment in query
9	sstart	Start of alignment in subject (database hit)
10	send	End of alignment in subject (database hit)
11	evalue	Expectation value (E-value)
12	bitscore	Bit score

BLAST in Galaxy

Output formats of BLAST

- **Second Option:**
  - The BLAST+ tools can optionally output additional columns of information, but this takes longer to calculate
  - Many commonly used extra columns are included by selecting the extended tabular output
  - The extra columns are included after the standard 12 columns
  - This is so that you can write workflow filtering steps that accept either the 12 or 25 column tabular BLAST output
  - Galaxy now uses this extended 25 column output by default:

Column	NCBI Name	Description
13	sallseqid	All subject Seq-id(s), separated by a ';'
14	score	Raw score
15	nident	Number of identical matches
16	positive	Number of positive-scoring matches
17	gaps	Total number of gaps
18	ppos	Percentage of positive-scoring matches
19	qframe	Query frame
20	sframe	Subject frame
21	qseq	Aligned part of query sequence
22	sseq	Aligned part of subject sequence
23	qlen	Query sequence length
24	slen	Subject sequence length
25	salltitles	All subject title(s), separated by a '<>'

BLAST in Galaxy

Output formats of BLAST

- **Third Option:**
  - To customize the tabular output by selecting which columns you want, from the standard set of 12, the default set of 25, or any of the additional columns BLAST+ offers (including species name)

- **Fourth Option:**
  - Request BLAST XML output, which is designed to be parsed by another program, and is understood by some Galaxy tools

- **Other Options:**
  - You can also choose several plain text or HTML output formats which are designed to be read by a person (not by another program)
  - The HTML versions use basic webpage formatting and can include links to the hits on the NCBI website
  - The pairwise output (the default on the NCBI BLAST website) shows each match as a pairwise alignment with the query
  - The two query anchored outputs show a multiple sequence alignment between the query and all the matches, and differ in how insertions are shown (marked as insertions or with gap characters added to the other sequences)

BLAST in Galaxy

A_Subject_01.fa

```
>NP_003131.1 sex determining region Y [Homo sapiens]
MQSYASAMLSVFNSDDYSPAVQENIPALRRSSSFLCTESCNSKYQCETGENSKGNVQDRVKRPMNAFIVW
SRDQRRKMALENPRMRNSEISKQLGYQWKMLTEAEKWPFFQEAQKLQAMHREKYPNYKYRPRRKAKMLPK
NCSLLPADPASVLCSEVQLDNRLYRDDCTKATHSRMEHQLGHLPPINAASSPQQRDRYSHWTKL
```

A_Query_01.fa

```
>NP_008872.1 SOX-10 [Homo sapiens]
MAEEQDLSEVELSPVGSEEPRCLSPGSAPSLGPDGGGGGSGLRASPGPGELGKVKKEQQDGEADDDKFPV
CIREAVSQVLSGYDWTLVPMPVRVNGASKSKPHVKRPMNAFMVWAQAARRKLADQYPHLHNAELSKTLGK
LWRLLNESDKRPFIEEAERLRMQHKKDHPDYKYQPRRRKNGKAAQGEAECPGGEAEQGGTAAIQAHYKSA
HLDHRHPGEGSPMSDGNPEHPSGQSHGPPTPPTTPKTELQSGKADPKRDGRSMGEGGKPHIDFGNVDIGE
ISHEVMSNMETFDVAELDQYLPPNGHPGHVSSYSAAGYGLGSALAVASGHSAWISKPPGVALPTVSPPGV
DAKAQVKTETAGPQGPPHYTDQPSTSQIAYTSLSLPHYGSAFPSISRPQFDYSDHQPSGPYYGHSGQASG
LYSAFSYMGPSQRPLYTAISDPSPSGPQSHSPTHWEQPVYTTLSRP
```

BLAST in Galaxy

blastp A_Query_01.fa vs 'A_Subject_01.fa'

NCBI BLAST+ blastp Search protein database with protein query sequence(s) (Galaxy Version 2.10.1+galaxy2)

Protein query sequence(s)

2: A_Query_01.fa

(-query)

Subject database/sequences

FASTA file from your history (see warning note below)

Protein FASTA subject file to use instead of a database

1: A_Subject_01.fa

(-subject)

Type of BLAST

- blastp - Traditional BLASTP to compare a protein query to a protein database
- blastp-short - BLASTP optimized for queries shorter than 30 residues
- blastp-fast - Use longer words for seeding, faster but less accurate

See help text for default parameter values for each BLAST type. (-task)

Set expectation value cutoff

0.001

(-evalue)

Output format

Tabular (standard 12 columns)

(-outfmt)

Advanced Options

Hide Advanced Options

Job Resource Parameters

Specify job resource parameters

Cores & Memory

1 core & 7GB memory

Number of processing cores & max total job memory

Time (hours)

24

Maximum job time

Execute

BLAST in Galaxy

blastp A_Query_01.fa vs 'A_Subject_01.fa'

qseqid	sseqid	pident	length	mismatch	gapopen	qstart	qend	sstart	send	evalue	bitscore
NP_008872.1	NP_003131.1	46.429	84	45	0	95	178	51	134	3.08e-27	94.7

BLAST in Galaxy

blastp A_Query_01.fa vs 'A_Subject_01.fa'

NCBI BLAST+ blastp Search protein database with protein query sequence(s) (Galaxy Version 2.10.1+galaxy2)

Protein query sequence(s)

2: A_Query_01.fa

(-query)

Subject database/sequences

FASTA file from your history (see warning note below)

Protein FASTA subject file to use instead of a database

1: A_Subject_01.fa

(-subject)

Type of BLAST

- blastp - Traditional BLASTP to compare a protein query to a protein database
- blastp-short - BLASTP optimized for queries shorter than 30 residues
- blastp-fast - Use longer words for seeding, faster but less accurate

See help text for default parameter values for each BLAST type. (-task)

Set expectation value cutoff

0.001

(-evalue)

Output format

Pairwise HTML

(-outfmt)

Advanced Options

Hide Advanced Options

Job Resource Parameters

Specify job resource parameters

Cores & Memory

1 core & 7GB memory

Number of processing cores & max total job memory

Time (hours)

24

Maximum job time

Execute

BLAST in Galaxy

blastp A_Query_01.fa vs 'A_Subject_01.fa'

```
Query= NP_008872.1 SOX-10 [Homo sapiens]

Length=466

                                                                      Score     E
Sequences producing significant alignments:                          (Bits)  Value

NP_003131.1 sex determining region Y [Homo sapiens]                   94.7    3e-27


> NP_003131.1 sex determining region Y [Homo sapiens]
Length=204

 Score = 94.7 bits (234),  Expect = 3e-27, Method: Compositional matrix adjust.
 Identities = 39/84 (46%), Positives = 62/84 (74%), Gaps = 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


Lambda      K        H        a          alpha
   0.310    0.130    0.399    0.792      4.96

Gapped
Lambda      K        H        a          alpha    sigma
   0.267   0.0410    0.140     1.90      42.6     43.6

Effective search space used: 77703


  Database: User specified sequence set (Input:
/scratch/user/galaxy/kaiser_galaxy/objects/6/a/3/dataset_6a38210a-
6a7f-4866-a61c-elea56df6a5f.dat).
    Posted date:  Unknown
  Number of letters in database: 204
  Number of sequences in database:  1



Matrix: BLOSUM62
Gap Penalties: Existence: 11, Extension: 1
Neighboring words threshold: 11
Window for multiple hits: 40
```

BLAST in Galaxy

blastp A_Query_01.fa vs 'A_Subject_01.fa'

NCBI BLAST+ blastp Search protein database with protein query sequence(s) (Galaxy Version 2.10.1+galaxy2)

Protein query sequence(s)

2: A_Query_01.fa

(-query)

Subject database/sequences

FASTA file from your history (see warning note below)

Protein FASTA subject file to use instead of a database

1: A_Subject_01.fa

(-subject)

Type of BLAST

- blastp - Traditional BLASTP to compare a protein query to a protein database
- blastp-short - BLASTP optimized for queries shorter than 30 residues
- blastp-fast - Use longer words for seeding, faster but less accurate

See help text for default parameter values for each BLAST type. (-task)

Set expectation value cutoff

0.001

(-evalue)

Output format

Query-anchored HTML

(-outfmt)

Advanced Options

Hide Advanced Options

Job Resource Parameters

Specify job resource parameters

Cores & Memory

1 core & 7GB memory

Number of processing cores & max total job memory

Time (hours)

24

Maximum job time

Execute

BLAST in Galaxy

blastp A_Query_01.fa vs 'A_Subject_01.fa'

BLAST Search Results

BLASTP 2.10.1+

Reference:
Stephen F. Altschul, Thomas L. Madden, Alejandro A. Schäffer,
Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997),
"Gapped BLAST and PSI-BLAST: a new generation of protein database
search programs", Nucleic Acids Res. 25:3389-3402.

**Reference for
composition-based statistics**:
Alejandro A. Schäffer, L. Aravind, Thomas L. Madden, Sergei
Shavirin, John L. Spouge, Yuri I. Wolf, Eugene V. Koonin, and
Stephen F. Altschul (2001), "Improving the accuracy of PSI-BLAST
protein database searches with composition-based statistics and
other refinements", Nucleic Acids Res. 29:2994-3005.

Database: User specified sequence set (Input:
/scratch/user/galaxy/kaiser_galaxy/objects/6/a/3/dataset_6a38210a-
6a7f-4866-a61c-e1ea56df6a5f.dat).
1 sequences; 204 total letters

Query= NP_008872.1 SOX-10 [Homo sapiens]

Length=466

Sequences producing significant alignments:	Score (Bits)	E Value
NP_003131.1 sex determining region Y [Homo sapiens]	94.7	3e-27

```
Query_1    95   NGASKSKPHVKRPMNAFMVWAQAARRKLADQYPHLHNAELSKTLGKLWRLLNESDKRPFI  154
Subject_1  51   NSKGNVQDRVKRPMNAFIVWSRDQRRKMALENPRMRNSEISKQLGYQWKMLTEAEKWPFF  110

Query_1    155  EEAERLRMQHKKDHPDYKYQPRRR  178
Subject_1  111  QEAQKLQAMHREKYPNYKYRPRRK  134
```

Lambda	K	H	a	alpha
0.310	0.130	0.399	0.792	4.96

Gapped Lambda	K	H	a	alpha	sigma
0.267	0.0410	0.140	1.90	42.6	43.6

Effective search space used: 77703

Database: User specified sequence set (Input:
/scratch/user/galaxy/kaiser_galaxy/objects/6/a/3/dataset_6a38210a-
6a7f-4866-a61c-e1ea56df6a5f.dat).
Posted date: Unknown
Number of letters in database: 204
Number of sequences in database: 1

Matrix: BLOSUM62
Gap Penalties: Existence: 11, Extension: 1
Neighboring words threshold: 11
Window for multiple hits: 40

BLAST in Galaxy

blastp A_Query_01.fa vs 'A_Subject_01.fa'

```
Query= NP_008872.1 SOX-10 [Homo sapiens]

Length=466

                                                                      Score     E
Sequences producing significant alignments:                          (Bits)  Value

NP_003131.1 sex determining region Y [Homo sapiens]                    94.7    3e-27


Query_1    95   NGASKSKPHVKRPMNAFMVWAQAARRKLADQYPHLHNAELSKTLGKLWRLLNESDKRPFI  154
Subject_1  51   NSKGNVQDRVKRPMNAFIVWSRDQRRKMALENPRMRNSEISKQLGYQWKMLTEAEKWPFF  110

Query_1    155  EEAERLRMQHKKDHPDYKYQPRRR  178
Subject_1  111  QEAQKLQAMHREKYPNYKYRPRRK  134


Lambda      K        H        a         alpha
   0.310    0.130    0.399    0.792     4.96

Gapped
Lambda      K        H        a         alpha    sigma
   0.267   0.0410    0.140     1.90     42.6     43.6

Effective search space used: 77703


  Database: User specified sequence set (Input:
/scratch/user/galaxy/kaiser_galaxy/objects/6/a/3/dataset_6a38210a-
6a7f-4866-a61c-e1ea56df6a5f.dat).
    Posted date:  Unknown
  Number of letters in database: 204
  Number of sequences in database:  1



Matrix: BLOSUM62
Gap Penalties: Existence: 11, Extension: 1
Neighboring words threshold: 11
Window for multiple hits: 40
```

BLAST in Galaxy

megablast B_Query_01.fa vs 'B_Subject_01.fa'

NCBI BLAST+ blastn Search nucleotide database with nucleotide query sequence(s) (Galaxy Version 2.10.1+galaxy2)

Nucleotide query sequence(s)

3: B_Query_01.fa

(-query)

Subject database/sequences

FASTA file from your history (see warning note below)

Nucleotide FASTA subject file to use instead of a database

1: B_Subject_01.fa

(-subject)

Type of BLAST

- ◉ megablast - Traditional megablast used to find very similar (e.g., intraspecies or closely related species) sequences
- ○ blastn - Traditional BLASTN requiring an exact match of 11, for somewhat similar sequences
- ○ blastn-short - BLASTN program optimized for sequences shorter than 50 bases
- ○ dc-megablast - Discontiguous megablast used to find more distant (e.g., interspecies) sequences

(-task)

Set expectation value cutoff

0.001

(-evalue)

Output format

Tabular (standard 12 columns)

(-outfmt)

Advanced Options

Hide Advanced Options

Job Resource Parameters

Specify job resource parameters

Cores & Memory

1 core & 7GB memory

Number of processing cores & max total job memory

Time (hours)

24

Maximum job time

Execute

BLAST in Galaxy

megablast B_Query_01.fa vs 'B_Subject_01.fa'

B_Query_01	B_Subject_01	99.980	4938	1	0	3104	8041	22175	27112	0.0	9114
B_Query_01	B_Subject_01	99.968	3111	1	0	10844	13954	84073	87183	0.0	5740
B_Query_01	B_Subject_01	100.000	1160	0	0	1	1160	423	1582	0.0	2143
B_Query_01	B_Subject_01	100.000	1119	0	0	1992	3110	18186	19304	0.0	2067
B_Query_01	B_Subject_01	100.000	429	0	0	8206	8634	40775	41203	0.0	793
B_Query_01	B_Subject_01	100.000	357	0	0	9174	9530	49092	49448	0.0	660
B_Query_01	B_Subject_01	99.234	261	1	1	1265	1524	4990	5250	3.54e-133	470
B_Query_01	B_Subject_01	100.000	247	0	0	10454	10700	80602	80848	2.76e-129	457
B_Query_01	B_Subject_01	100.000	201	0	0	9952	10152	65230	65430	1.03e-103	372
B_Query_01	B_Subject_01	100.000	191	0	0	8815	9005	43655	43845	3.72e-98	353
B_Query_01	B_Subject_01	100.000	185	0	0	8632	8816	42340	42524	8.06e-95	342
B_Query_01	B_Subject_01	86.379	301	36	5	12474	12771	28977	28679	2.92e-89	324
B_Query_01	B_Subject_01	100.000	173	0	0	9003	9175	48436	48608	3.78e-88	320
B_Query_01	B_Subject_01	100.000	169	0	0	10150	10318	65662	65830	6.32e-86	313
B_Query_01	B_Subject_01	100.000	158	0	0	9529	9686	56315	56472	8.23e-80	292
B_Query_01	B_Subject_01	100.000	150	0	0	9682	9831	56866	57015	2.30e-75	278
B_Query_01	B_Subject_01	100.000	149	0	0	10699	10847	82811	82959	8.29e-75	276
B_Query_01	B_Subject_01	100.000	141	0	0	10316	10456	65921	66061	2.32e-70	261
B_Query_01	B_Subject_01	82.143	308	50	5	12473	12778	63521	63217	8.35e-70	259
B_Query_01	B_Subject_01	81.877	309	48	6	12473	12777	21275	20971	3.88e-68	254
B_Query_01	B_Subject_01	81.911	293	43	8	72	360	55778	55492	1.09e-63	239
B_Query_01	B_Subject_01	80.906	309	51	7	53	361	55030	54730	3.91e-63	237
B_Query_01	B_Subject_01	100.000	125	0	0	9830	9954	62583	62707	1.82e-61	231
B_Query_01	B_Subject_01	100.000	117	0	0	1714	1830	12412	12528	5.10e-57	217
B_Query_01	B_Subject_01	100.000	113	0	0	1881	1993	16834	16946	8.53e-55	209
B_Query_01	B_Subject_01	100.000	111	0	0	1514	1624	10989	11099	1.10e-53	206
B_Query_01	B_Subject_01	99.091	110	1	0	1157	1266	2333	2442	1.85e-51	198
B_Query_01	B_Subject_01	78.289	304	57	8	62	363	62156	62452	4.00e-48	187
B_Query_01	B_Subject_01	78.716	296	52	6	72	361	66744	66454	4.00e-48	187
B_Query_01	B_Subject_01	78.351	291	58	5	63	351	44567	44854	5.17e-47	183
B_Query_01	B_Subject_01	100.000	98	0	0	8040	8137	30472	30569	1.86e-46	182
B_Query_01	B_Subject_01	100.000	71	0	0	8136	8206	32741	32811	1.90e-31	132
B_Query_01	B_Subject_01	85.484	124	18	0	12649	12772	42162	42285	6.83e-31	130
B_Query_01	B_Subject_01	100.000	52	0	0	1623	1674	12014	12065	6.93e-21	97.1
B_Query_01	B_Subject_01	100.000	52	0	0	1829	1880	15356	15407	6.93e-21	97.1
B_Query_01	B_Subject_01	81.818	110	16	4	11696	11804	78312	78418	1.16e-18	89.8
B_Query_01	B_Subject_01	78.014	141	30	1	11699	11839	55794	55655	4.17e-18	87.9
B_Query_01	B_Subject_01	100.000	45	0	0	1672	1716	12154	12198	5.39e-17	84.2

BLAST in Galaxy

megablast B_Query_01.fa vs 'B_Subject_01.fa'

NCBI BLAST+ blastn Search nucleotide database with nucleotide query sequence(s) (Galaxy Version 2.10.1+galaxy2)

Nucleotide query sequence(s)

3: B_Query_01.fa

(-query)

Subject database/sequences

FASTA file from your history (see warning note below)

Nucleotide FASTA subject file to use instead of a database

1: B_Subject_01.fa

(-subject)

Type of BLAST

- megablast - Traditional megablast used to find very similar (e.g., intraspecies or closely related species) sequences
- blastn - Traditional BLASTN requiring an exact match of 11, for somewhat similar sequences
- blastn-short - BLASTN program optimized for sequences shorter than 50 bases
- dc-megablast - Discontiguous megablast used to find more distant (e.g., interspecies) sequences

(-task)

Set expectation value cutoff

1e-250

(-evalue)

Output format

Tabular (standard 12 columns)

(-outfmt)

Advanced Options

Hide Advanced Options

Job Resource Parameters

Specify job resource parameters

Cores & Memory

1 core & 7GB memory

Number of processing cores & max total job memory

Time (hours)

24

Maximum job time

Execute

BLAST in Galaxy

megablast B_Query_01.fa vs 'B_Subject_01.fa'

B_Query_01	B_Subject_01	99.980	4938	1	0	3104	8041	22175	27112	0.0	9114
B_Query_01	B_Subject_01	99.968	3111	1	0	10844	13954	84073	87183	0.0	5740
B_Query_01	B_Subject_01	100.000	1160	0	0	1	1160	423	1582	0.0	2143
B_Query_01	B_Subject_01	100.000	1119	0	0	1992	3110	18186	19304	0.0	2067

BLAST in Galaxy

megablast B_Query_01.fa vs 'B_Subject_01.fa'

NCBI BLAST+ blastn Search nucleotide database with nucleotide query sequence(s) (Galaxy Version 2.10.1+galaxy2)

Nucleotide query sequence(s)

3: B_Query_01.fa

(-query)

Subject database/sequences

FASTA file from your history (see warning note below)

Nucleotide FASTA subject file to use instead of a database

1: B_Subject_01.fa

(-subject)

Type of BLAST

- megablast - Traditional megablast used to find very similar (e.g., intraspecies or closely related species) sequences
- blastn - Traditional BLASTN requiring an exact match of 11, for somewhat similar sequences
- blastn-short - BLASTN program optimized for sequences shorter than 50 bases
- dc-megablast - Discontiguous megablast used to find more distant (e.g., interspecies) sequences

(-task)

Set expectation value cutoff

1e-250

(-evalue)

Output format

Pairwise HTML

(-outfmt)

Advanced Options

Hide Advanced Options

Job Resource Parameters

Specify job resource parameters

Cores & Memory

1 core & 7GB memory

Number of processing cores & max total job memory

Time (hours)

24

Maximum job time

Execute

BLAST in Galaxy

megablast B_Query_01.fa vs 'B_Subject_01.fa'

```
Query= B_Query_01

Length=13954
                                                                      Score     E
Sequences producing significant alignments:                          (Bits)  Value

B_Subject_01                                                          9114    0.0


> B_Subject_01
Length=87183

 Score = 9114 bits (4935),  Expect = 0.0
 Identities = 4937/4938 (99%), Gaps = 0/4938 (0%)
 Strand=Plus/Plus

Query  3104   TTCAGGTTTATTGCATTCTTCTGTGAAAAGAAGCTGTTCACAGAATGATTCTGAAGAACC  3163
              || ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  22175  TTTAGGTTTATTGCATTCTTCTGTGAAAAGAAGCTGTTCACAGAATGATTCTGAAGAACC  22234

Query  3164   AACTTTGTCCTTAACTAGCTCTTTTGGGACAATTCTGAGGAAATGTTCTAGAAATGAAAC  3223
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  22235  AACTTTGTCCTTAACTAGCTCTTTTGGGACAATTCTGAGGAAATGTTCTAGAAATGAAAC  22294

Query  3224   ATGTTCTAATAATACAGTAATCTCTCAGGATCTTGATTATAAAGAAGCAAAATGTAATAA  3283
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  22295  ATGTTCTAATAATACAGTAATCTCTCAGGATCTTGATTATAAAGAAGCAAAATGTAATAA  22354

Query  3284   GGAAAAACTACAGTTATTTATTACCCCAGAAGCTGATTCTCTGTCATGCCTGCAGGAAGG  3343
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  22355  GGAAAAACTACAGTTATTTATTACCCCAGAAGCTGATTCTCTGTCATGCCTGCAGGAAGG  22414

Query  3344   ACAGTGTGAAAATGATCCaaaaagcaaaaaaGTTTCAGATATAAAAGAAGAGGTCTTGGC  3403
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  22415  ACAGTGTGAAAATGATCCAAAAAGCAAAAAAGTTTCAGATATAAAAGAAGAGGTCTTGGC  22474

Query  3404   TGCAGCATGTCACCCAGTACAACATTCAAAAGTGGAATACAGTGATACTGACTTTCAATC  3463
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  22475  TGCAGCATGTCACCCAGTACAACATTCAAAAGTGGAATACAGTGATACTGACTTTCAATC  22534

Query  3464   CCAGAAAAGTCTTTTATATGATCATGAAAATGCCAGCACTCTTATTTTAACTCCTACTTC  3523
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  22535  CCAGAAAAGTCTTTTATATGATCATGAAAATGCCAGCACTCTTATTTTAACTCCTACTTC  22594

Query  3524   CAAGGATGTTCTGTCAAACCTAGTCATGATTTCTAGAGGCAAAGAATCATACAAAATGTC  3583
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  22595  CAAGGATGTTCTGTCAAACCTAGTCATGATTTCTAGAGGCAAAGAATCATACAAAATGTC  22654

Query  3584   AGACAAGCTCAAAGGTAACAATTATGAATCTGATGTTGAATTAACCAAAAATATTCCCAT  3643
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  22655  AGACAAGCTCAAAGGTAACAATTATGAATCTGATGTTGAATTAACCAAAAATATTCCCAT  22714

Query  3644   GGAAAAGAATCAAGATGTATGTGCTTTAAATGAAAATTATAAAAACGTTGAGCTGTTGCC  3703
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  22715  GGAAAAGAATCAAGATGTATGTGCTTTAAATGAAAATTATAAAAACGTTGAGCTGTTGCC  22774

Query  3704   ACCTGAAAAATACATGAGAGTAGCATCACCTTCAAGAAAGGTACAATTCAACCAAAACAC  3763
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  22775  ACCTGAAAAATACATGAGAGTAGCATCACCTTCAAGAAAGGTACAATTCAACCAAAACAC  22834

Query  3764   AAATCTAAGAGTAATCCaaaaaaaTCAAGAAGAAACTACTTCAATTTCAAAAATAACTGT  3823
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  22835  AAATCTAAGAGTAATCCAAAAAAATCAAGAAGAAACTACTTCAATTTCAAAAATAACTGT  22894

Query  3824   CAATCCAGACTCTGAAGAACTTTTCTCAGACAATGAGAATAATTTTGTCTTCCAAGTAGC  3883
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
```

BLAST in Galaxy

megablast B_Query_01.fa vs 'B_Subject_01.fa'

NCBI BLAST+ blastn Search nucleotide database with nucleotide query sequence(s) (Galaxy Version 2.10.1+galaxy2)

Nucleotide query sequence(s)

3: B_Query_01.fa

(-query)

Subject database/sequences

FASTA file from your history (see warning note below)

Nucleotide FASTA subject file to use instead of a database

1: B_Subject_01.fa

(-subject)

Type of BLAST

- megablast - Traditional megablast used to find very similar (e.g., intraspecies or closely related species) sequences
- blastn - Traditional BLASTN requiring an exact match of 11, for somewhat similar sequences
- blastn-short - BLASTN program optimized for sequences shorter than 50 bases
- dc-megablast - Discontiguous megablast used to find more distant (e.g., interspecies) sequences

(-task)

Set expectation value cutoff

1e-250

(-evalue)

Output format

Query-anchored HTML

(-outfmt)

Advanced Options

Hide Advanced Options

Job Resource Parameters

Specify job resource parameters

Cores & Memory

1 core & 7GB memory

Number of processing cores & max total job memory

Time (hours)

24

Maximum job time

Execute

BLAST in Galaxy

megablast B_Query_01.fa vs 'B_Subject_01.fa'

```
Query= B_Query_01

Length=13954

                                                                      Score     E
Sequences producing significant alignments:                          (Bits)  Value

B_Subject_01                                                          9114    0.0


Query_1    1      GGTCTTCCAGTTCAGCCTGCGAGGAAGACAGGTGATCCGAATCCTAAGAATGCAAAAGAT  60
Subject_1  423    GGTCTTCCAGTTCAGCCTGCGAGGAAGACAGGTGATCCGAATCCTAAGAATGCAAAAGAT  482

Query_1    61     GGGCCGGGTGTGGTGGCTCATGCCTGTAATCCCAGCGCTTTGGGAGGCCGAGGCAGGCAG  120
Subject_1  483    GGGCCGGGTGTGGTGGCTCATGCCTGTAATCCCAGCGCTTTGGGAGGCCGAGGCAGGCAG  542

Query_1    121    ATCACCTGAGGTCGGGAGGTTGAGACCAGACTGACCAACAACGGAGAAACCCCGTCTCTA  180
Subject_1  543    ATCACCTGAGGTCGGGAGGTTGAGACCAGACTGACCAACAACGGAGAAACCCCGTCTCTA  602

Query_1    181    CTTAAAAATGCAAAGTTAGCCGTGCGTGGTGGCCCATGCCTGTATTCCCAGCTACTCGGG  240
Subject_1  603    CTTAAAAATGCAAAGTTAGCCGTGCGTGGTGGCCCATGCCTGTATTCCCAGCTACTCGGG  662

Query_1    241    AGGCTGAGGCAGGAGAACCACTTGATCCCTGGAGGCGGAAGTTGCGGTGAGCGGAGATTG  300
Subject_1  663    AGGCTGAGGCAGGAGAACCACTTGATCCCTGGAGGCGGAAGTTGCGGTGAGCGGAGATTG  722

Query_1    301    CGCCATTGCACACCAGCCCGGGCCACAAGAGCGAAACTCCGTCTCaaaaaaaaaaGCAAA  360
Subject_1  723    CGCCATTGCACACCAGCCCGGGCCACAAGAGCGAAACTCCGTCTCAAAAAAAAAAGCAAA  782

Query_1    361    AGATACTACCAAGCCCTGCGGAGCAAGGTACCTCACACTTCATGAGCGAGTTAAGATGGG  420
Subject_1  783    AGATACTACCAAGCCCTGCGGAGCAAGGTACCTCACACTTCATGAGCGAGTTAAGATGGG  842

Query_1    421    TTTCACAATTTTTCAAGCAAGGAAACGGGCTCGGAGGTCTTGAACACCTGCTACCCAATA  480
Subject_1  843    TTTCACAATTTTTCAAGCAAGGAAACGGGCTCGGAGGTCTTGAACACCTGCTACCCAATA  902

Query_1    481    GCAGAACAGCTACTGGAACTAAAATCCTCTGATTTCAAATAACAGCCCCGCCCACTACCA  540
Subject_1  903    GCAGAACAGCTACTGGAACTAAAATCCTCTGATTTCAAATAACAGCCCCGCCCACTACCA  962

Query_1    541    CTAAGTGAAGTCATCCACAACCACACACCGACCACTCTAAGCTTTTGTAAGATCGGCTCG  600
Subject_1  963    CTAAGTGAAGTCATCCACAACCACACACCGACCACTCTAAGCTTTTGTAAGATCGGCTCG  1022

Query_1    601    CTTTGGGGAACAGGTCTTGAGAGAACATCCCTTTTAAGGTCAGAACAAAGGTATTTCATA  660
Subject_1  1023   CTTTGGGGAACAGGTCTTGAGAGAACATCCCTTTTAAGGTCAGAACAAAGGTATTTCATA  1082

Query_1    661    GGTCCCAGGTCGTGTCCCGAGGGCGCCCACCCAAACATGAGCTGGAGCAAAAAGAAAGGG  720
Subject_1  1083   GGTCCCAGGTCGTGTCCCGAGGGCGCCCACCCAAACATGAGCTGGAGCAAAAAGAAAGGG  1142

Query_1    721    ATGGGGGACTTGGAGTAGGCATAGGGGCGGCCCCTCCAAGCAGGGTGGCCTGGGACTCTT  780
Subject_1  1143   ATGGGGGACTTGGAGTAGGCATAGGGGCGGCCCCTCCAAGCAGGGTGGCCTGGGACTCTT  1202

Query_1    781    AAGGGTCAGCGAGAAGAGAACACACACTCCAGCTCCCGCTTTATTCGGTCAGATACTGAC  840
Subject_1  1203   AAGGGTCAGCGAGAAGAGAACACACACTCCAGCTCCCGCTTTATTCGGTCAGATACTGAC  1262

Query_1    841    GGTTGGGATGCCTGACAAGGAATTTCCTTTCGCCACACTGAGAAATACCCGCAGCGGCCC  900
Subject_1  1263   GGTTGGGATGCCTGACAAGGAATTTCCTTTCGCCACACTGAGAAATACCCGCAGCGGCCC  1322

Query_1    901    ACCCAGGCCTGACTTCCGGGTGGTGCGTGTGCTGCGTGTCGCGTCACGGCGTCACGTGGC  960
Subject_1  1323   ACCCAGGCCTGACTTCCGGGTGGTGCGTGTGCTGCGTGTCGCGTCACGGCGTCACGTGGC  1382

Query_1    961    CAGCGCGGGCTTGTGGCGCGAGCTTCTGAAACTAGGCGGCAGAGGCGGAGCCGCTGTGGC  1020
Subject_1  1383   CAGCGCGGGCTTGTGGCGCGAGCTTCTGAAACTAGGCGGCAGAGGCGGAGCCGCTGTGGC  1442

Query_1    1021   ACTGCTGCGCCTCTGCTGCGCCTCGGGTGTCTTTTGCGGCGGTGGGTCGCCGCCGGGAGA  1080
Subject_1  1443   ACTGCTGCGCCTCTGCTGCGCCTCGGGTGTCTTTTGCGGCGGTGGGTCGCCGCCGGGAGA  1502

Query_1    1081   AGCGTGAGGGGACAGATTTGTGACCGGCGCGGTTTTTGTCAGCTTACTCCGGCCAAAAAA  1140
```