Evolutionary Genetics
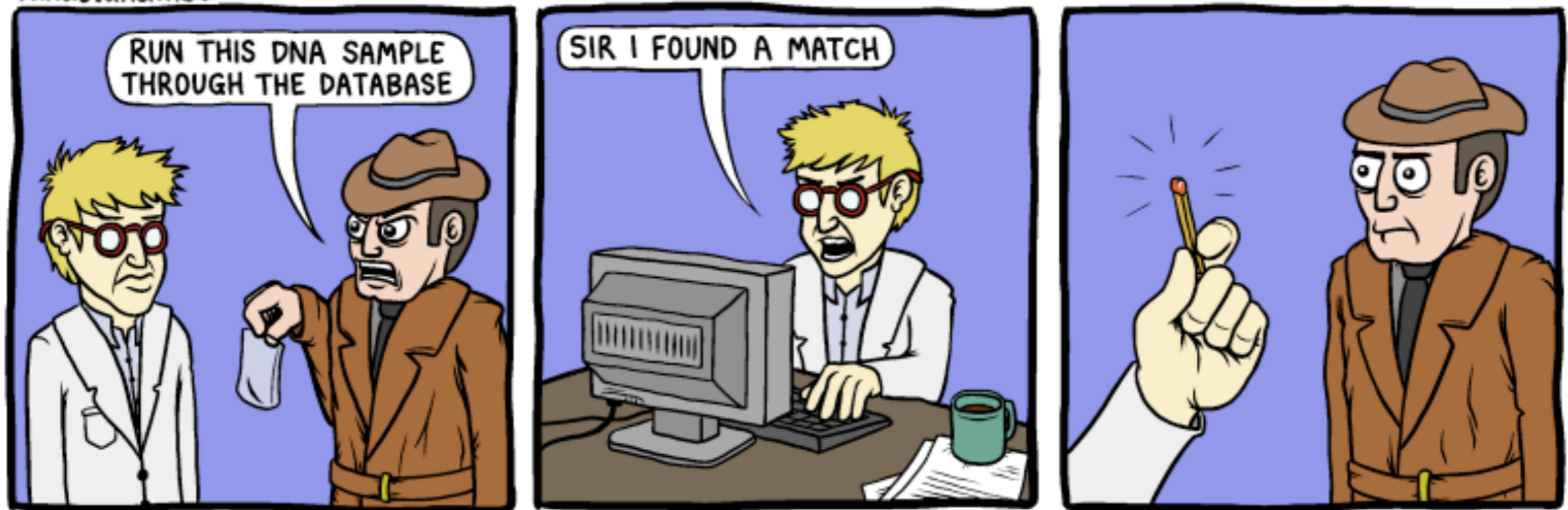
LV 25600-01 | Lecture with exercises | 4KP

BLAST is the acronym for "**B**asic **L**ocal **A**lignment **S**earch **T**ool", which is a **local** alignment search tool first described by Altschul et **al**. (1990). NCBI started providing sequence alignment service to the public using BLAST in 1992, first through its blast email server (decommissioned in 2002) and later through the web (1997).
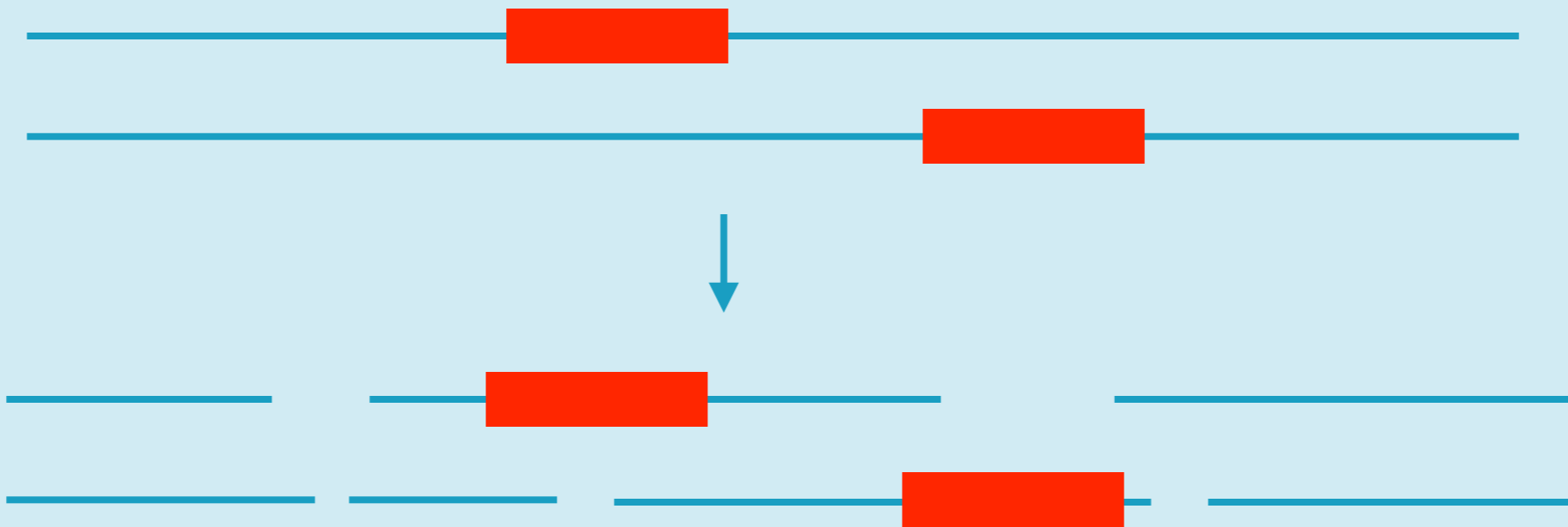

An Essential Guide to the Basic Local Alignment Search Tool

BLAST

O'REILLY®

Ian Korf, Mark Yandell & Joseph Bedell

BLAST
Authors: I Korf, M Yandell, J Bedell
Publisher: O'Reilly Media
Release Date: July 2003
Pages: 362

Global Alignment

Local Alignment

Match   Mismatch   Indel*   Gap

```
Seq1 : TTGCACGGCTTGGTCCA-GTGCGGTTTAC
       ||||||||||x|||||||||||||  |||
Seq2 : TTGCACGGCTCGGTCCACGTGC----TAC
```

```
Seq1 : ...................-...........
Seq2 : ..........C...............----...
Cons : TTGCACGGCTtGGTCCAcGTGCggttTAC
```
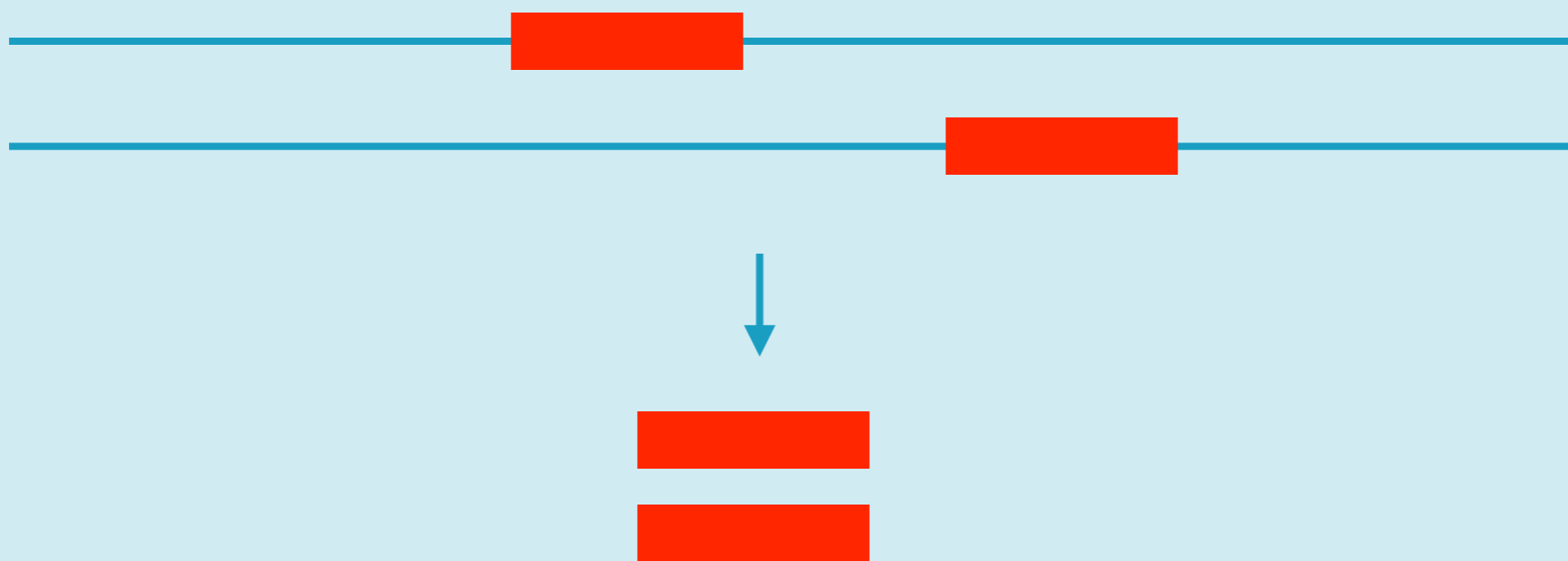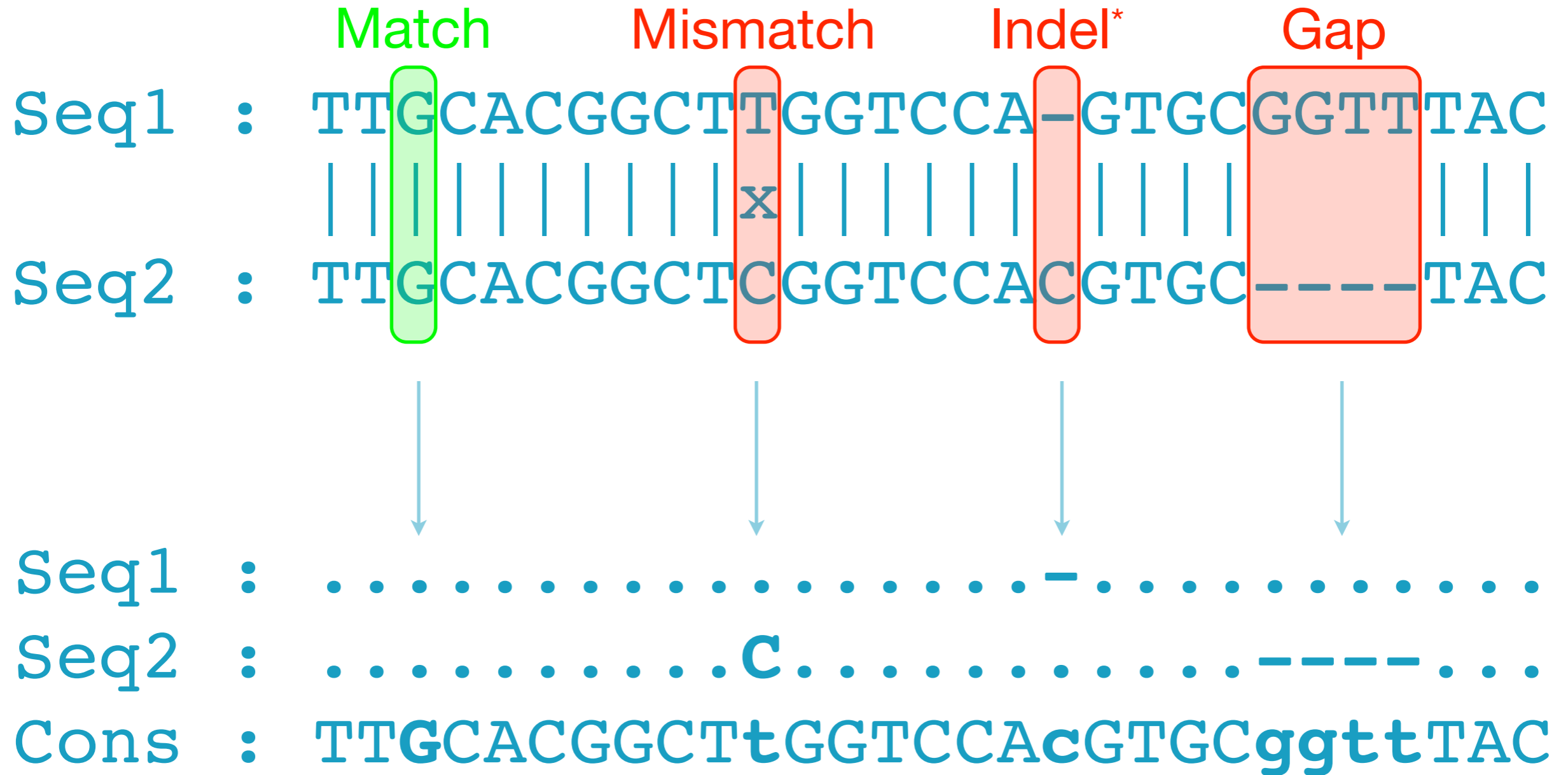
* indels: insertions & deletions

BLAST finds the optimal alignment by using the **"word matching" algorithm**, in which BLAST does the search in several distinctive phases: 1) **generating overlapping words** from the input query, 2) scanning the database for **word matches** (hits), and 3) **extending word hits** to produce (local) alignments through multiple steps of extension.

During the first phase, BLAST breaks the input query into short overlapping segments (words/**seeds**). In the second phase BLAST takes those query words and **scans the target database** for initial matches. The nucleotide BLAST algorithm looks for any single exact word match. The protein BLAST algorithm uses a scoring threshold cutoff to identify matches. In addition, protein BLAST algorithm also requires two word hits within a certain distance in order to proceed to the next step.

ATGCGGTCACGTCACG > query sequence

ATGCG > word 1

TGCGG > word 2

GCGGT > word 3

CGGTC > word 4

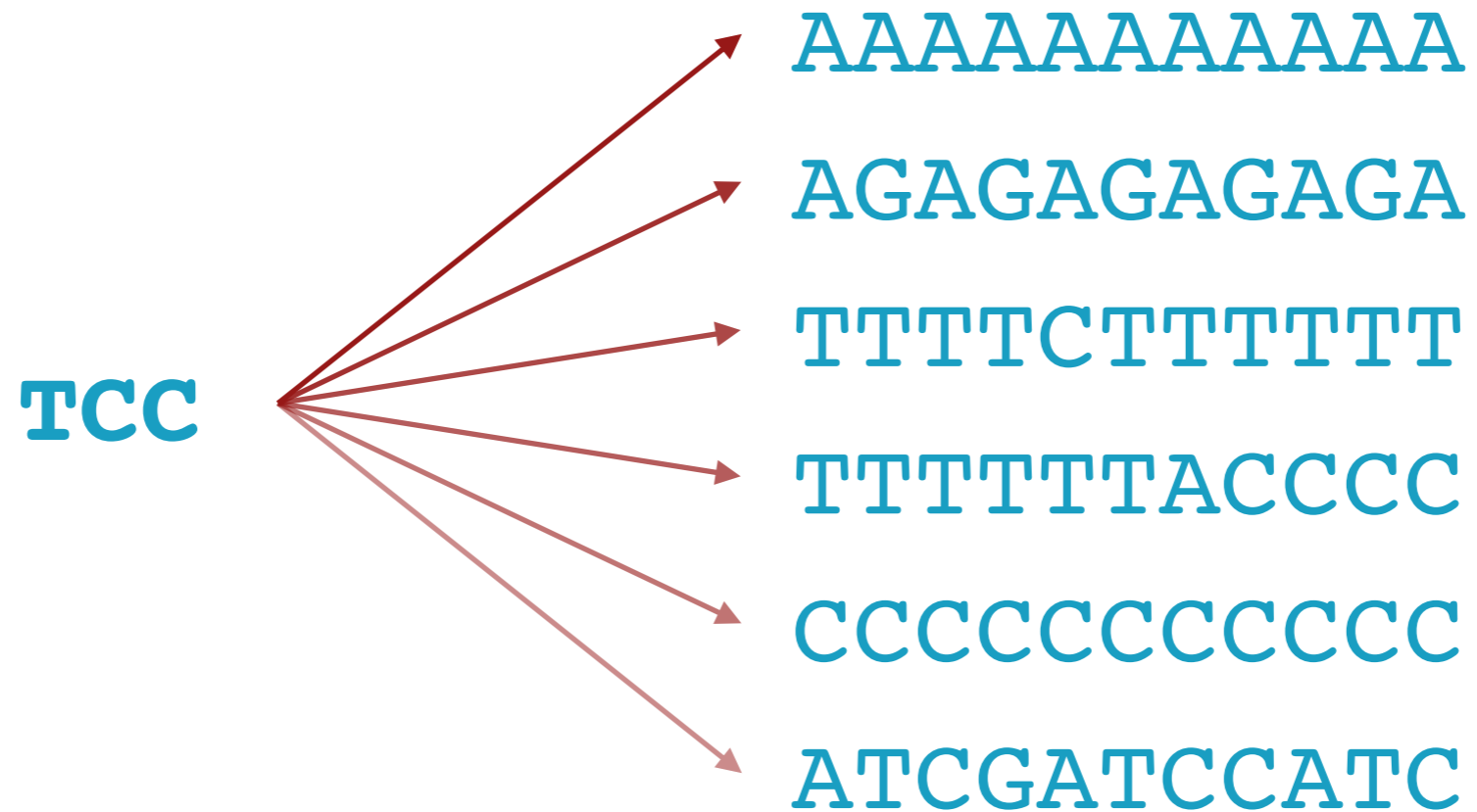**GGTCA** > word 5

GTCAC > word 6

TCC ?

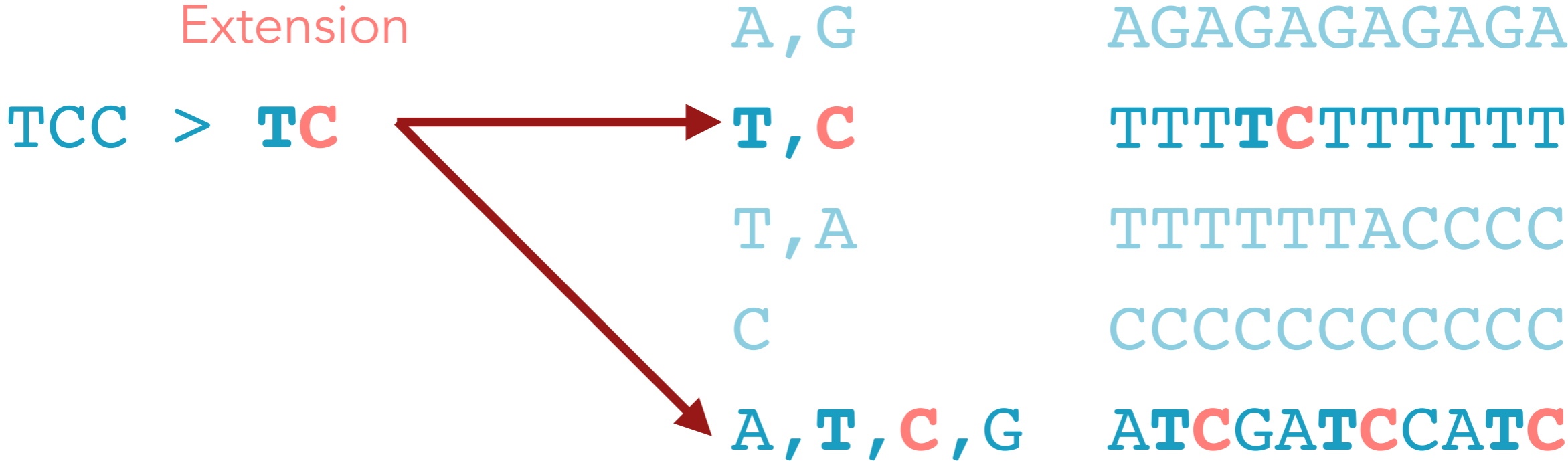AAAAAAAAAAA

AGAGAGAGAGA
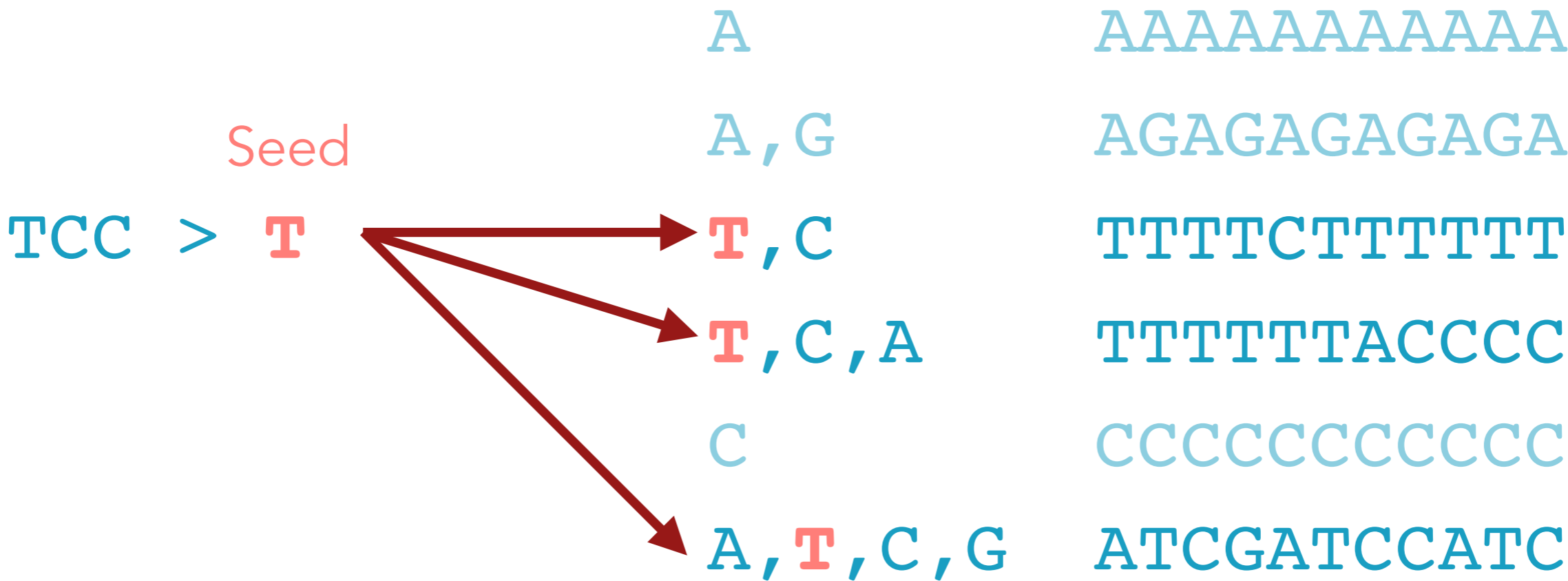
TTTTCTTTTTT

TTTTTTACCCC

CCCCCCCCCCC

ATCGATCCATC

TCC

- AAAAAAAAAAA
- AGAGAGAGAGA
- TTTTCTTTTTT
- TTTTTTACCCC
- CCCCCCCCCC
- ATCGATCCATC

Index

A	AAAAAAAAAAAA
A,G	AGAGAGAGAGA
T,C	TTTTCTTTTTT
T,C,A	TTTTTTACCCC
C	CCCCCCCCCCC
A,T,C,G	ATCGATCCATC

```
                              A            AAAAAAAAAAA

            Seed              A,G          AGAGAGAGAGA

TCC  >  T                     T,C          TTTTCTTTTTTT

                              T,C,A        TTTTTTACCCC

                              C            CCCCCCCCCCC

                              A,T,C,G      ATCGATCCATC
```

A   AAAAAAAAAAAA

A,G   AGAGAGAGAGA

Extension

TCC > **TC**   **T,C**   **TTTTCTTTTTT**

T,A   TTTTTTACCCC

C   CCCCCCCCCCC

A,**T**,**C**,G   AT**C**GA**TC**CAT**C**

AAAAAAAAAAA

Extension

AGAGAGAGAGA

TCC

TTTTCTTTTTT

TTTTTTACCCC

CCCCCCCCCCC

ATCGA**TCC**ATC

Global Alignment

```
Query >    1 ------TCC--- 11
Gaps >          |||< Matches
Subject >  1 ATCGATCCATC 11
```

Local Alignment

Identities **3**/**3** (**100%**)

```
Query   1 TCC 3
          |||
Subject 6 TCC 9
```

Bioinformatics ▶ BLAST Search

Bioinformatics ▶ BLAST Search

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

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

GEO

Find matches to gene
expression profiles

IgBLAST

Search immunoglobulins
and T cell receptor
sequences

VecScreen

Search sequences for vector
contamination

CDART

Find sequences with similar
conserved domain
architecture

Targeted Loci

Search markers for
phylogenetic analysis

Multiple Alignment

Align sequences using
domain and protein
constraints

BioAssay

Search protein or nucleotide
targets in PubChem
BioAssay

MOLE-BLAST

Establish taxonomy for
uncultured or enviromental
sequences

Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance.   Learn more
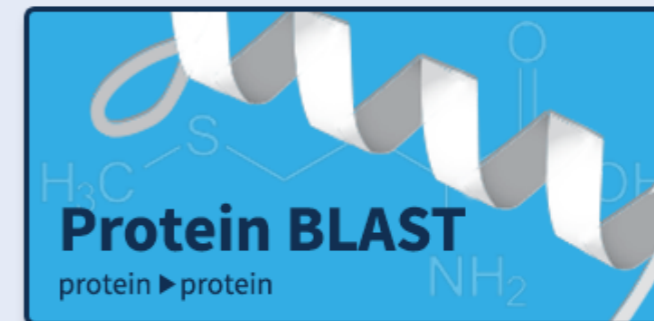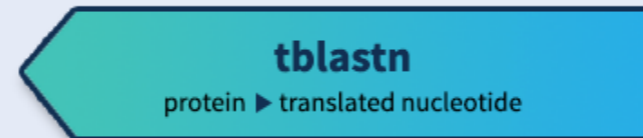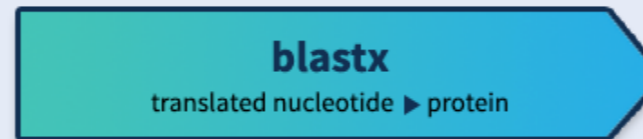
N E W S

BLAST+ 2.5.0 released

The new version offers support for HTTPS, accession.version as the primary sequence identifier, support for composition-based statistics with RPSTBLASTN, and a new taxonomic organism report.
Fri, 23 Sep 2016 17:00:00 EST

📄 More BLAST news...

Web BLAST

Nucleotide BLAST
nucleotide ▶ nucleotide

blastx
translated nucleotide ▶ protein

tblastn
protein ▶ translated nucleotide

Protein BLAST
protein ▶ protein

BLAST Genomes

Enter organism common name, scientific name, or tax id	Search

Human   Mouse   Rat   Microbes

search	input		query		database
blastn	nt	➡	nt	➡	nt
blastp	pr	➡	pr	➡	pr
blastx	nt	➡	pr (6)	➡	pr
tblastn	pr	➡	pr	➡	pr (6)
tblastx	nt	➡	pr (6)	➡	pr (6)

blastn compares nucleotide queries to a nucleotide database

blastp compares protein queries to a protein database

blastx compares a nucleotide query translated in all six reading frames against a protein database

tblastn compares a protein query against a nucleotide sequence database dynamically translated in all six reading frames

tblastx compares a nucleotide query in all six reading frames against a nucleotide sequence database in all six reading frames

DNA ⟶ NCBI
DNA ⟶ Protein ⟶ BLAST

Nucleotide–nucleotide searches are beneficial because no information is lost in the alignment. When a codon is translated from nucleotides to amino acids, approximately 69% of the complexity is lost (4^3=64 possible nucleotide combinations mapped to 20 amino acids). In contrast, however, **the true physical relationship between two coding sequences is best captured in the translated view**. Matrices that take into account physical properties, such as PAM and BLOSUM, can be used to add power to the search. Additionally, in a nucleotide search, there are only four possible character states compared to 20 in an amino acid search. Thus the **probability of a match** due to chance versus a match due to common ancestry (identify in state versus identical by descent) is higher.

▶ **Setting up a BLAST search**

Step 1. Plan the search
Step 2. Enter the query sequence
Step 3. Choose the appropriate search parameters
Step 4. Submit the query

▶ **Deciphering the BLAST output**

Step 1. Examine the alignment scores and statistics
Step 2. Examine the alignments
Step 3. Review search details to plan the next step
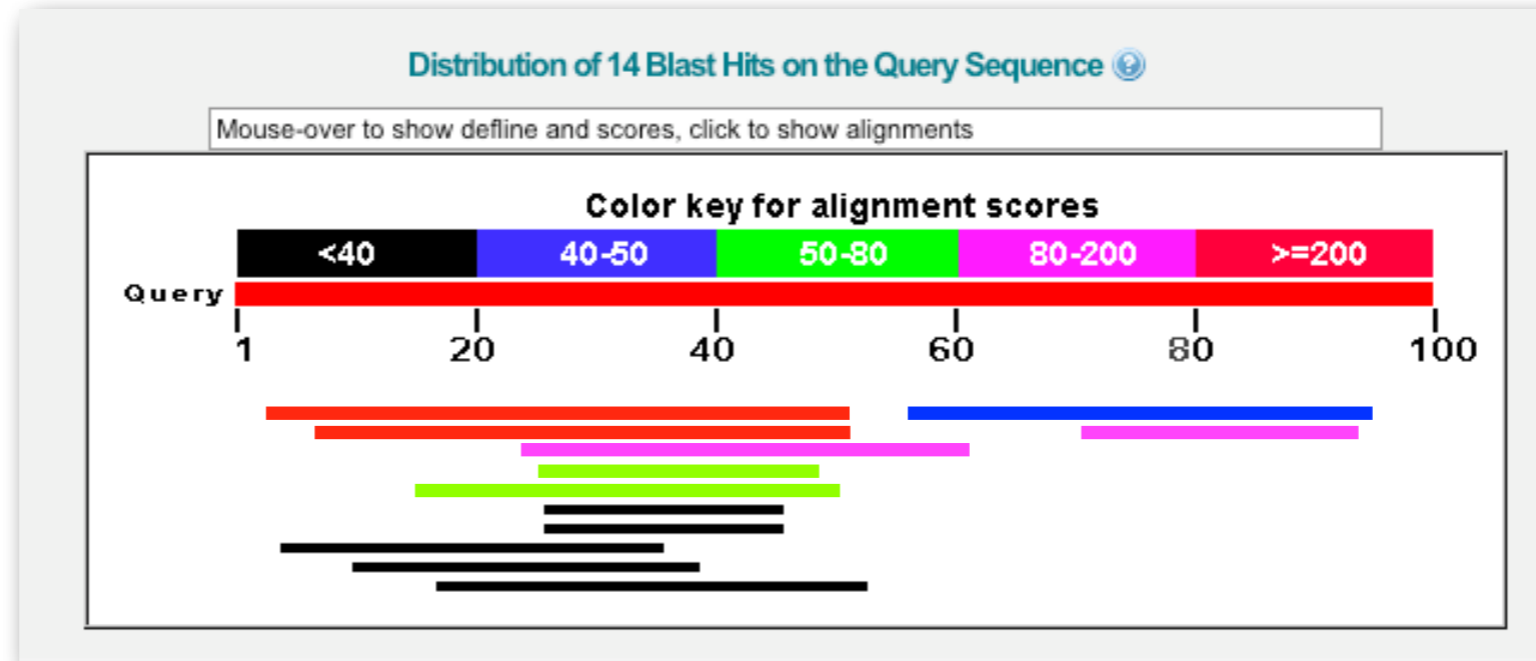
▶ **Post-BLAST analysis**

Perform a PSI-BLAST analysis
Create a multiple alignment
Try motif searching with PHI-BLAST

Graphic Summary

Distribution of 14 Blast Hits on the Query Sequence

Mouse-over to show defline and scores, click to show alignments

Color key for alignment scores

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

Bit Score ▶

click here to see the nr entry (accession) ▼

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
DQ487112.1	Panax ginseng dehydrin 7 (Dhn7) mRNA, complete cds	39.9	39.9	21%	2.3	100%	
DQ487106.1	Panax ginseng dehydrin 1 (Dhn1) mRNA, complete cds	39.9	39.9	21%	2.3	100%	
AC238433.1	Mus musculus BAC clone RP24-160E3 from chromosome 9, complete :	38.1	38.1	20%	8.2	100%	
AC215885.3	Mus musculus BAC clone RP23-36L10 from chromosome 9, complete :	38.1	38.1	20%	8.2	100%	
CU467051.7	Pig DNA sequence from clone CH242-177E21 on chromosome 2, comp	38.1	38.1	28%	8.2	90%	
NM_001079232.1	Xenopus (Silurana) tropicalis T-cell activation RhoGTPase activating pr	38.1	38.1	20%	8.2	100%	UGM

click here to see the corresponding alignment ▲

Nucleotide Alignment

```
> emb|FN568088.1|  Homo sapiens SRY gene for sex determining region Y, individual
TH7
Length=615

 Score =  848 bits (459),  Expect = 0.0
 Identities = 480/497 (97%), Gaps = 2/497 (0%)
 Strand=Plus/Minus

Query  1    CTACARCTTTGTCCAGTGGCTGTAGCGGTCCCGTTGCTGCGGTGAGCTGGCTGCGTTGAT  60
            ||||| ||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  615  CTACAGCTTTGTCCAGTGGCTGTAGCGGTCCCGTTGCTGCGGTGAGCTGGCTGCGTTGAT  556

Query  61   GGGCGGTAAGTGGCCTAGCTGGTGCTCCATTCTTGAGTGTGTGGCTTTCGTACAGTCATC  120
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  555  GGGCGGTAAGTGGCCTAGCTGGTGCTCCATTCTTGAGTGTGTGGCTTTCGTACAGTCATC  496

Query  121  CCTGTACAACCTGTTGTCCAGTTGCACTTCGCTGCAGAGTACCGAAGCGGGATCTGCGGG  180
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  495  CCTGTACAACCTGTTGTCCAGTTGCACTTCGCTGCAGAGTACCGAAGCGGGATCTGCGGG  436

Query  181  AAGCAAACTGCAATTCTTCGGCAGCATNTTCGCCTTCCGACGAGGTCGATACTTATAATT  240
            |||||||||||||||||||||||||||| |||||||||||||||||||||||||||||||
Sbjct  435  AAGCAAACTGCAATTCTTCGGCAGCATCTTCGCCTTCCGACGAGGTCGATACTTATAATT  376

Query  241  CGGGTATTTCTCTCTGTGCATGGCCTGTAATTTCTGTGCCTCCTGGAAGAATGGCCATTT  300
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  375  CGGGTATTTCTCTCTGTGCATGGCCTGTAATTTCTGTGCCTCCTGGAAGAATGGCCATTT  316

Query  301  TTCGGCTTCAGTAAGCATTTTCCACTGGTATCCCAGCTGCTTGCTGATCTCTGAAAGTTT  360
            |||||||||||||||||||||||||||||||||||||||||||||||||||||||  ||||
Sbjct  315  TTCGGCTTCAGTAAGCATTTTCCACTGGTATCCCAGCTGCTTGCTGATCTCTGA--GTTT  258

Query  361  CGCATTCTGGGATTCTCTAGAGCCATCTTGCGCCTCTGATCGCGAGACCACACGNNGAAT  420
            |||||||||||||||||||||||||||||||||||||||||||||||||||||| ||||
Sbjct  257  CGCATTCTGGGATTCTCTAGAGCCATCTTGCGCCTCTGATCGCGAGACCACACGATGAAT  198

Query  421  GCGTTCATGGGTCGCTTCACTCTATCCTGGNNNNNNNNNNNNTTACTGTTTTCTCCCGTTTCA  480
            ||||||||||||||||||||||||||||||      ||||||||||||||||||||||||
Sbjct  197  GCGTTCATGGGTCGCTTCACTCTATCCTGGACGTTGCCTTTACTGTTTTCTCCCGTTTCA  138

Query  481  RRCTGATACTTAGAGTT  497
              |||||||||||||||
Sbjct  137  CACTGATACTTAGAGTT  121
```

query sequence

database sequence

Alignment - The process of lining up two or more sequences to achieve **maximal levels of identity** (and conservation, in the case of amino acid sequences) for the purpose of assessing the **degree of similarity** and the **possibility of homology**.
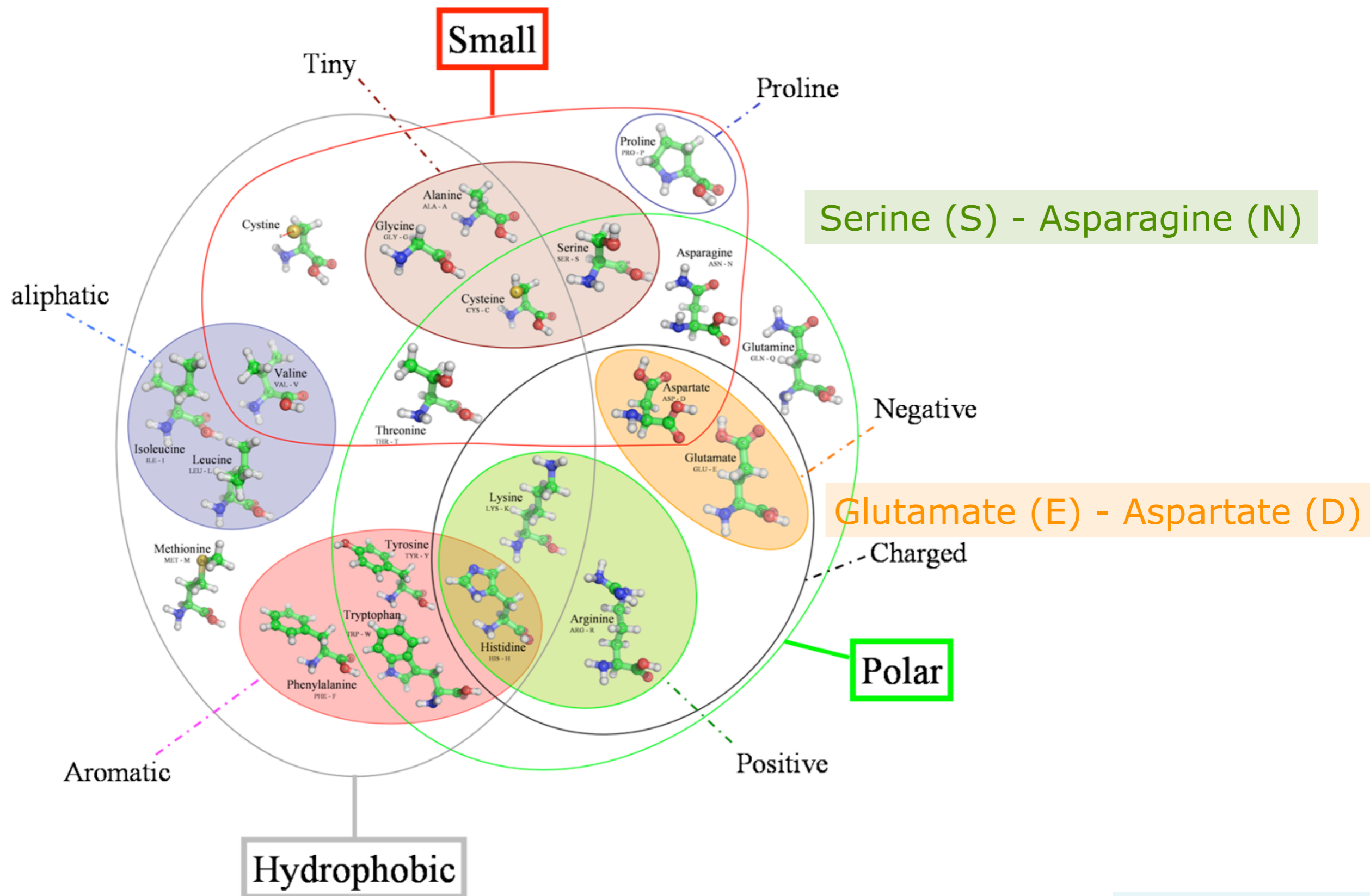
Identity - The extent to which two (nucleotide or amino acid) sequences are invariant.

$$PID = \frac{\text{\# of identical aa or nt}}{\text{\# of total aa or nt}} \times 100$$

Protein Alignment

```
>☐emb|CBH40193.1|    sex determining region Y [Homo sapiens]
Length=204

 Score =  330 bits (845),   Expect = 5e-89
 Identities = 157/164 (96%), Positives = 157/164 (96%), Gaps = 0/164 (0%)
 Frame = +1

Query  1     NSKYQXETGENSXXXXQDRVKRPMNAFXVWSRDQRRKMALENPRMRNSEISKQLGYQWKM  180
             NSKYQ ETGENS    QDRVKRPMNAF VWSRDQRRKMALENPRMRNSEISKQLGYQWKM
Sbjct  41    NSKYQCETGENSKGNVQDRVKRPMNAFIVWSRDQRRKMALENPRMRNSEISKQLGYQWKM  100

Query  181   LTEAEKWPFFQEAQKLQAMHREKYPNYKYRPRRKAXMLPKNCSLLPADPASVLCSEVQLD  360
             LTEAEKWPFFQEAQKLQAMHREKYPNYKYRPRRKA MLPKNCSLLPADPASVLCSEVQLD
Sbjct  101   LTEAEKWPFFQEAQKLQAMHREKYPNYKYRPRRKAKMLPKNCSLLPADPASVLCSEVQLD  160

Query  361   NRLYRDDCTKATHSRMEHQLGHLPPINAASSPQQRDRYSHWTKL  492
             NRLYRDDCTKATHSRMEHQLGHLPPINAASSPQQRDRYSHWTKL
Sbjct  161   NRLYRDDCTKATHSRMEHQLGHLPPINAASSPQQRDRYSHWTKL  204
```

query sequence

database sequence

matching sequence

Protein Alignment - **Identities**

151 - 4 mis-matches = 147 identities

Range 1: 1 to 151 GenPept   Graphics   ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
301 bits(770)	3e-102	Compositional matrix adjust.	147/151(97%)	150/151(99%)	0/151(0%)

```
Query    6   MKKIDVKILDARVGKDFPLPTYATPGSAGLDLRACIDDVMEIAPGTTTLIPTGLAIHIAD   65
             MKKIDVKILDARVGK FPLPTYATPGSAGLDLRACI+DVMEIAPGTTTLIPTGLAIHIAD
Sbjct    1   MKKIDVKILDARVGKAFPLPTYATPGSAGLDLRACIEDVMEIAPGTTTLIPTGLAIHIAD   60

Query   66   PSLAAVILPRSGLGHKHGIVLGNLVGLIDADYQGQLMVSVWNRGQESFTLQPGDRMAQLV  125
             P+LAAVILPRSGLGHKHGIVLGNLVGLIDADYQGQLMVSVWNRGQESFTLQPGDRMAQLV
Sbjct   61   PNLAAVILPRSGLGHKHGIVLGNLVGLIDADYQGQLMVSVWNRGQESFTLQPGDRMAQLV  120

Query  126   FVPVVQAEFNLVEEFDASLRGEGGFGHSGRQ      156
             FVPVVQAEFNLV+EFDASLRGEGGFGHSGRQ
Sbjct  121   FVPVVQAEFNLVDEFDASLRGEGGFGHSGRQ      151
```

4 mis-matches

Protein Alignment - **Positives**

147 identities + 3 similar = 150 positives

Range 1: 1 to 151 GenPept   Graphics   ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
301 bits(770)	3e-102	Compositional matrix adjust.	147/151(97%)	150/151(99%)	0/151(0%)

```
Query  6    MKKIDVKILDARVGKDFPLPTYATPGSAGLDLRACIDVMEIAPGTTTLIPTGLAIHIAD   65
            MKKIDVKILDARVGK FPLPTYATPGSAGLDLRACI+DVMEIAPGTTTLIPTGLAIHIAD
Sbjct  1    MKKIDVKILDARVGKAFPLPTYATPGSAGLDLRACIEDVMEIAPGTTTLIPTGLAIHIAD   60

Query  66   PSLAAVILPRSGLGHKHGIVLGNLVGLIDADYQGQLMVSVWNRGQESFTLQPGDRMAQLV   125
            P+LAAVILPRSGLGHKHGIVLGNLVGLIDADYQGQLMVSVWNRGQESFTLQPGDRMAQLV
Sbjct  61   PNLAAVILPRSGLGHKHGIVLGNLVGLIDADYQGQLMVSVWNRGQESFTLQPGDRMAQLV   120

Query  126  FVPVVQAEFNLVEIFDASLRGEGGFGHSGRQ   156
            FVPVVQAEFNLV+IFDASLRGEGGFGHSGRQ
Sbjct  121  FVPVVQAEFNLVDIFDASLRGEGGFGHSGRQ   151
```

Glutamate (E) - Aspartate (D)

Serine (S) - Asparagine (N)

Alignment - The process of lining up two or more sequences to achieve **maximal levels of identity** (and conservation, in the case of amino acid sequences) for the purpose of assessing the **degree of similarity** and the **possibility of homology**.

Identity - The extent to which two (nucleotide or amino acid) sequences are invariant.

$$PID = \frac{\# \text{ of identical aa or nt}}{\# \text{ of total aa or nt}} \times 100$$

Similarity - The extent to which nucleotide or protein sequences are related. The extent of similarity between two sequences can be based on percent sequence identity and/or conservation. In BLAST similarity refers to a positive matrix score.

$$\%Similarity = \frac{\# \text{ of identical aa or nt} + \# \text{ of similar aa or nt substitutions}}{\# \text{ of total aa or nt}} \times 100$$

Homology - Similarity attributed to descent from a common ancestor.

Serine (S) - Asparagine (N)

Glutamate (E) - Aspartate (D)

Bioinformatics ▶ BLAST Search

Bioinformatics ▶ BLAST Search

< Edit Search   Save Search   Search Summary ⌄   ❓ How to read this report?   ▶ BLAST Help Videos   ↺ Back to Traditional Results Page

| | Your search parameters were adjusted to search for a short input sequence. |

Filter Results

Job Title	Seq1
RID	RXGRMSJU014 *Search expires on 10-09 15:42 pm* Download All ⌄
Program	BLASTN ❓ Citation ⌄
Database	nt See details ⌄
Query ID	lcl|Query_25269
Description	Seq1
Molecule type	nucleic acid
Query Length	22
Other reports	Distance tree of results MSA viewer ❓

Organism   *only top 20 will appear*   ☐ exclude

[Type common name, binomial, taxid or group name]

➕ Add organism

Percent Identity		**E value**		**Query Coverage**	
[ ] to [ ]		[ ] to [ ]		[ ] to [ ]	

Filter   **Reset**

Descriptions	Graphic Summary	Alignments	Taxonomy

Sequences producing significant alignments   Download ⌄   Manage Columns ⌄   Show [100 ⌄] ❓

☑ select all *100 sequences selected*   GenBank   Graphics   Distance tree of results

|	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☑	Homo sapiens clone NA12878_chr21_27696869_27696870 genomic sequence	44.1	44.1	100%	0.035	100.00%	KY429753.1
☑	Homo sapiens clone CHM1_3_139044444_139044445 genomic sequence	44.1	44.1	100%	0.035	100.00%	KY429400.1
☑	Homo sapiens clone CHM1_8_84691462_84691463 genomic sequence	44.1	44.1	100%	0.035	100.00%	KY429510.1
☑	Eukaryotic synthetic construct chromosome 20	42.1	213	95%	0.14	100.00%	CP034499.1
☑	Eukaryotic synthetic construct chromosome 18	42.1	188	95%	0.14	100.00%	CP034496.1
☑	Eukaryotic synthetic construct chromosome 17	42.1	170	95%	0.14	100.00%	CP034495.1
☑	Eukaryotic synthetic construct chromosome 16	42.1	102	95%	0.14	100.00%	CP034494.1
☑	Pongo abelii chromosome 5 clone CH276-75J1, complete sequence	42.1	42.1	95%	0.14	100.00%	AC275833.1
☑	Pongo abelii chromosome 5 clone CH276-124H21, complete sequence	42.1	42.1	95%	0.14	100.00%	AC275818.1

E value

Expectation value. The number of different alignments with scores equivalent to or better than S that are expected to occur in a database search by chance. **The lower the E value, the more significant the score.**

Query `ATGCTGGTA`

BLAST

Hit A: e-value 0.0000432

Hit B: e-value 0.0000081

Query `ATGCTGGTA`

BLAST

Hit A: e-value 0.0000432

```
ATG
|||
ATG
```

Hit B: e-value 0.0000081

```
ATGCTGGTA
|||  |||||
ATGGTGGTA
```

Graphic Summary

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
DQ487112.1	Panax ginseng dehydrin 7 (Dhn7) mRNA, complete cds	39.9	39.9	21%	2.3	100%	
DQ487106.1	Panax ginseng dehydrin 1 (Dhn1) mRNA, complete cds	39.9	39.9	21%	2.3	100%	
AC238433.1	Mus musculus BAC clone RP24-160E3 from chromosome 9, complete s	38.1	38.1	20%	8.2	100%	
AC215885.3	Mus musculus BAC clone RP23-36L10 from chromosome 9, complete s	38.1	38.1	20%	8.2	100%	
CU467051.7	Pig DNA sequence from clone CH242-177E21 on chromosome 2, comp	38.1	38.1	28%	8.2	90%	
NM_001079232.1	Xenopus (Silurana) tropicalis T-cell activation RhoGTPase activating pr	38.1	38.1	20%	8.2	100%	UGM

E value

Expectation value. The number of different alignments with scores equivalent to or better than S that are expected to occur in a database search by chance. **The lower the E value, the more significant the score.**

Bit score

The value S' is derived from the raw alignment score S in which the statistical properties of the scoring system used have been taken into account. Because bit scores have been normalized with respect to the scoring system, they **can be used to compare alignment scores from different searches**.

Descriptions | **Graphic Summary** | Alignments | Taxonomy

Distribution of the top 182 Blast Hits on 100 subject sequences
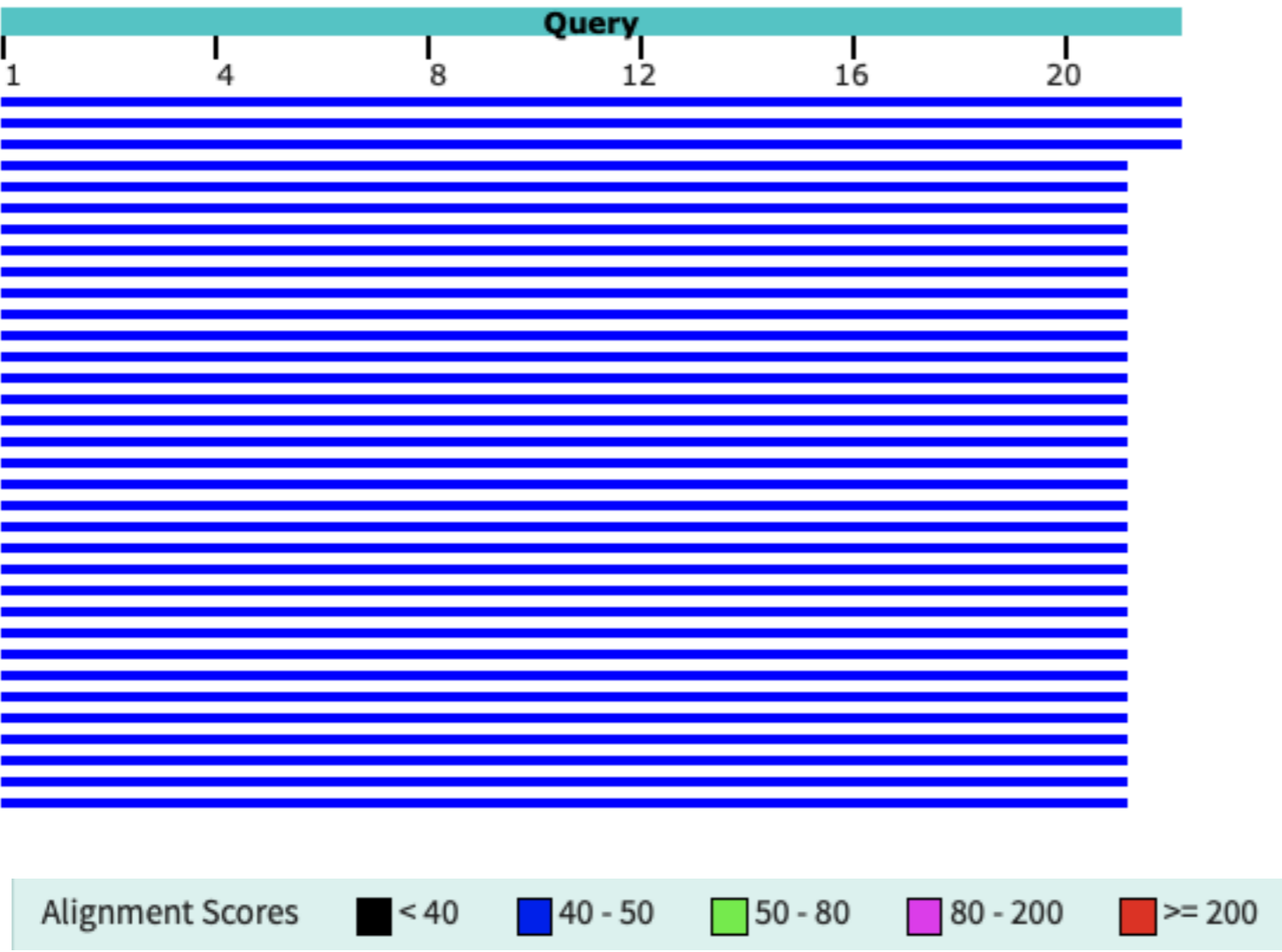
Bioinformatics ▶ BLAST Search

Descriptions	Graphic Summary	**Alignments**	Taxonomy

⬇ Download ⌄   GenBank Graphics   ▼ Next ▲ Previous ◄Descriptions

Homo sapiens clone NA12878_chr21_27696869_27696870 genomic sequence

Sequence ID: KY429753.1   Length: **3003**   Number of Matches: **1**

Range 1: 60 to 81 GenBank   Graphics   ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
44.1 bits(22)	0.035	22/22(100%)	0/22(0%)	Plus/Minus

```
Query   1    TGCACATGTACCTAAAACTTAG   22
             ||||||||||||||||||||||
Sbjct   81   TGCACATGTACCTAAAACTTAG   60
```

⬇ Download ⌄   GenBank Graphics   ▼ Next ▲ Previous ◄Descriptions

Homo sapiens clone CHM1_3_139044444_139044445 genomic sequence

Sequence ID: KY429400.1   Length: **6061**   Number of Matches: **1**

Range 1: 52 to 73 GenBank   Graphics   ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
44.1 bits(22)	0.035	22/22(100%)	0/22(0%)	Plus/Minus

```
Query   1    TGCACATGTACCTAAAACTTAG   22
             ||||||||||||||||||||||
Sbjct   73   TGCACATGTACCTAAAACTTAG   52
```

Descriptions	Graphic Summary	Alignments	**Taxonomy**

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

100 sequences selected ❓

Organism	Blast Name	Score	Number of Hits	Description
root			104	
. cellular organisms			94	
. . Eukaryota	eukaryotes		92	
. . . Bilateria	animals		90	
. . . . Euteleostomi	vertebrates		88	
. . . . . Amniota	vertebrates		80	
. . . . . . Boreoeutheria	placentals		78	
. . . . Euarchontoglires	placentals		73	
. . . . Hominoidea	primates		64	
. . . . Hominidae	primates		63	
. . . . Homininae	primates		57	
. . . . Homo sapiens	primates	44.1	53	Homo sapiens hits
. . . . Pan troglodytes	primates	42.1	4	Pan troglodytes hits
. . . . Pongo abelii	primates	42.1	6	Pongo abelii hits
. . . . Nomascus leucogenys	primates	36.2	1	Nomascus leucogenys hits
. . . . Mus musculus	rodents	38.2	4	Mus musculus hits
. . . . Onychomys torridus	rodents	38.2	1	Onychomys torridus hits
. . . . Galeopterus variegatus	placentals	36.2	1	Galeopterus variegatus hits
. . . . Acomys russatus	rodents	36.2	3	Acomys russatus hits
. . . . Rousettus aegyptiacus	bats	38.2	2	Rousettus aegyptiacus hits
. . . . Canis lupus familiaris	carnivores	36.2	2	Canis lupus familiaris hits
. . . . Felis catus	carnivores	36.2	1	Felis catus hits
. . . . . . Anas platyrhynchos	birds	40.1	1	Anas platyrhynchos hits
. . . . . . Streptopelia turtur	birds	36.2	1	Streptopelia turtur hits
. . . . . Danio kyathit	bony fishes	38.2	2	Danio kyathit hits
. . . . . Sparus aurata	bony fishes	36.2	1	Sparus aurata hits
. . . . . Danio rerio	bony fishes	36.2	3	Danio rerio hits
. . . . . Epinephelus fuscoguttatus	bony fishes	36.2	1	Epinephelus fuscoguttatus hits
. . . . . Poecilia reticulata	bony fishes	34.2	1	Poecilia reticulata hits
. . . . Belonocnema treatae	wasps, ants, and bees	36.2	1	Belonocnema treatae hits
. . . . Carposina sasakii	moths	34.2	1	Carposina sasakii hits
. . . Raphanus sativus	eudicots	36.2	1	Raphanus sativus hits
. . . Medicago truncatula	eudicots	36.2	1	Medicago truncatula hits
. . Acinetobacter seifertii	g-proteobacteria	36.2	2	Acinetobacter seifertii hits
. eukaryotic synthetic construct	other sequences	42.1	10	eukaryotic synthetic construct hits

BLAST in Terminal

Blast on fasta file - for smaller references

```
blastn –db SUBJECT.fa –evalue 0.0001 –query QUERY.fa –outfmt 6 –out RES.blast
```

Index reference (subject) first and blast against index db

```
makeblastdb –dbtype nucl –in REF.fa –title "REF" –logfile REF.log
```

```
blastn –db REF –evalue 0.0001 –query Q.fa –outfmt 6 –out Q_dbREF.blast
```

BLAST in R

Packages with blast functions:

```
blastSeq {hoardeR}
blastSequences {annotate}
rBLAST (GitHub)
```

Blast via system command:

```
system(command = "/path/to/blast/blastn -db REF -query Q.fa -outfmt 6 -evalue 1e-6)
```

```
system2(
    command = "/path/to/blast/blastn",
    args = c("-db REF -query Q.fa -outfmt 6 -evalue 10e-6))
```

BLAT (BLAST-like alignment tool)

MegaBLAST (BLAT variant)

UBLAST (USEARCH BLAST alternative)

Primary Databases

GenBank® : NIH genetic sequence database, an annotated collection of all publicly available DNA sequences. http://www.ncbi.nlm.nih.gov/genbank/

ENA - The European Nucleotide Archive (ENA) captures and presents information relating to experimental workflows that are based around nucleotide sequencing. http://www.ebi.ac.uk/ena/
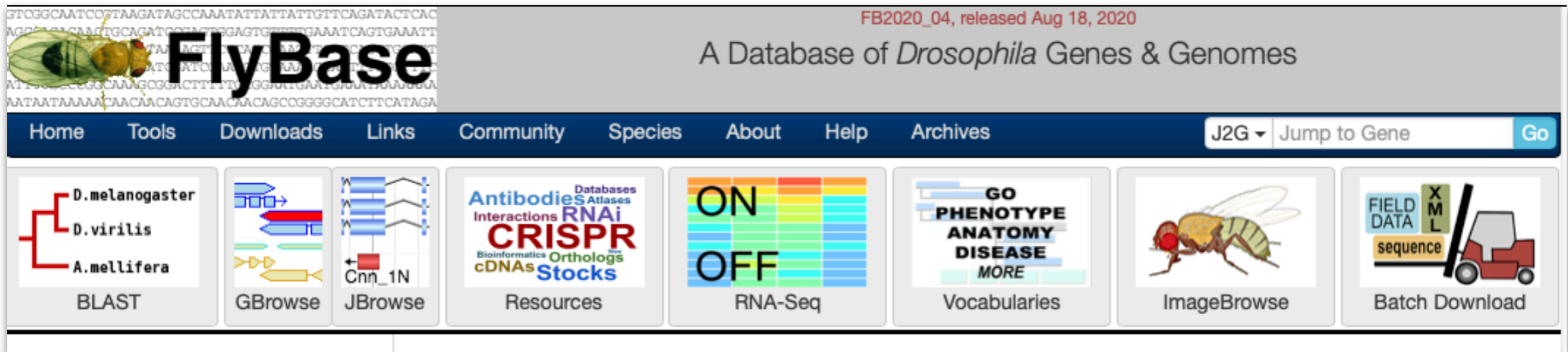
DDBJ - DNA Data Bank of Japan was established 1986. http://www.ddbj.nig.ac.jp/

INSDC - The International Nucleotide Sequence Databases (INSD) have been developed and maintained collaboratively between DDBJ, ENA, and GenBank for over 18 years. http://insdc.org/

Primary Databases

User

submit

retrieve

Error Rate
unknown

Secondary Databases

User

retrieve

Error Rate
lower

Self-Study Guide

Bioinformatics ▶ BLAST Search

>Dmag_B24_ORF0007_contigh23_2356_3466

```
ATGTGAACAAGTCTGAGAGATTCATCAACCGAATGTATATGAAAGCGGTGGTCCAGTCTG
ATCGGGCGGGCTTTCGATCAACAACAACAACAACAACAACAACAAGTACGATCGATCTAA
CTAGCTGACTAGCTGGACTGACTAGCTACTACGTACACGATCATATAATCGCGCGCGGCC
CCCTATATAGCTACGATGCATCGTATATAAATATTCTTATCTCCCTTA
```

Fasta header

Sequence

```
NCBI fasta headers:
>gi|224922792|ref|NM_000860.4| Homo sapiens hydroxyprostaglandin
dehydrogenase 15-(NAD) (HPGD), transcript variant 1, mRNA

Your header:
>Code_Species_Location/Gene/Coordinates
```

Sequences are expected to be represented in the standard IUB/IUPAC amino acid and nucleic acid codes, with these exceptions: lower-case letters are accepted and are mapped into upper-case; a single hyphen or dash can be used to represent a gap of indeterminate length; and in amino acid sequences, U and * are acceptable letters (see below). Before submitting a request, any numerical digits in the query sequence should either be removed or replaced by appropriate letter codes (e.g., N for unknown nucleic acid residue or X for unknown amino acid residue). The nucleic acid codes supported are:

```
A   adenosine        C   cytidine          G   guanine
T   thymidine        N   A/G/C/T (any)     U   uridine
K   G/T (keto)       S   G/C (strong)      Y   T/C (pyrimidine)
M   A/C (amino)      W   A/T (weak)        R   G/A (purine)
B   G/T/C            D   G/A/T             H   A/C/T
V   G/C/A            -   gap of indeterminate length
```

For those programs that use amino acid query sequences (BLASTP and TBLASTN), the accepted amino acid codes are:

```
A   alanine                   P   proline
B   aspartate/asparagine      Q   glutamine
C   cystine                   R   arginine
D   aspartate                 S   serine
E   glutamate                 T   threonine
F   phenylalanine             U   selenocysteine
G   glycine                   V   valine
H   histidine                 W   tryptophan
I   isoleucine                Y   tyrosine
K   lysine                    Z   glutamate/glutamine
L   leucine                   X   any
M   methionine                *   translation stop
N   asparagine                -   gap of indeterminate length
```

Choose Database (Subject)

Megablast is intended for comparing a query to closely related sequences and works best if the target percent identity is 95% or more but is very fast.

Discontiguous megablast uses an initial seed that ignores some bases (allowing mismatches) and is intended for cross-species comparisons.

BlastN is slow, but allows a word-size down to seven bases.

Change Default Search Parameter

Bioinformatics ▶ BLAST Search

▼ **Algorithm parameters**   Note: Parameter values that differ from the default are highlighted in yellow and marked with ♦ sign

General Parameters

Max target sequences — `[ 100 ▲▼ ]`
Select the maximum number of aligned sequences to display ❓

Short queries — ☑ Automatically adjust parameters for short input sequences ❓

Expect threshold — `[ 10 ]` ❓

Word size — `[ 28 ▲▼ ]` ❓

Max matches in a query range — `[ 0 ]` ❓

Scoring Parameters

Match/Mismatch Scores — `[ 1,-2 ▲▼ ]` ❓

Gap Costs — `[ Linear ▲▼ ]` ❓

Match/mismatch ratio	Similarity (%)
0.33 (1/−3)	99
−0.5 (1/−2)	95
−1 (1/−1)	75

Filters and Masking

Filter — ☑ Low complexity regions ❓
☐ Species-specific repeats for: `[ Human ▲▼ ]` ❓

Mask — ☑ Mask for lookup table only ❓
☐ Mask lower case letters ❓

BLAST — Search **database Nucleotide collection (nr/nt)** using **Megablast (Optimize for highly similar sequences)**
☐ Show results in a new window

HS20 | UniBS | JCW

When choosing a matrix, it is important to consider the alternatives. Do not simply choose the default setting without some initial consideration.

Alignment size	Best at detecting	Similarity (%)	PAM	BLOSUM
Short	Similarity within a **species**	75–90	PAM30	BLOSUM95
"	Similarity within a **genus**	60–75	PAM70	BLOSUM85
Medium	Similarity within a **family**	50–60	PAM120	**BLOSUM80**
"	The **largest range** of similarity	40–50	PAM160	**BLOSUM62**
Long	Similarity within a **class**	30–40	PAM250	**BLOSUM45**
"	Similarity within the **twilight zone**	20–30		BLOSUM30

The matrices highlighted in bold are available through NCBI's BLAST web interface. **BLOSUM62** has been shown to provide the best results in BLAST searches overall due to its ability to detect large ranges of similarity. Nevertheless, the other matrices have their strengths. For example, if your goal is to only detect sequences of high similarity to infer homology within a species, the PAM30, BLOSUM90, and PAM70 matrices would provide the best results.

Percent **A**ccepted **M**utation (**PAM**) - A unit introduced by Margaret Dayhoff et al. (1978) to quantify the amount of evolutionary change in a protein sequence. 1.0 PAM unit, is the amount of evolution which will change, on average, 1% of amino acids in a protein sequence. A PAM(x) substitution matrix is a look-up table in which scores for each amino acid substitution have been calculated based on the frequency of that substitution in closely related proteins that have experienced a certain amount (x) of evolutionary divergence.

The PAM matrices imply a **Markov chain model** of protein mutation. The PAM matrices are normalized so that, for instance, the PAM1 matrix gives substitution probabilities for sequences that have experienced one point mutation for every hundred amino acids. The mutations may overlap so that the sequences reflected in the PAM250 matrix have experienced 250 mutation events for every 100 amino acids, yet only 80 out of every 100 amino acids have been affected.
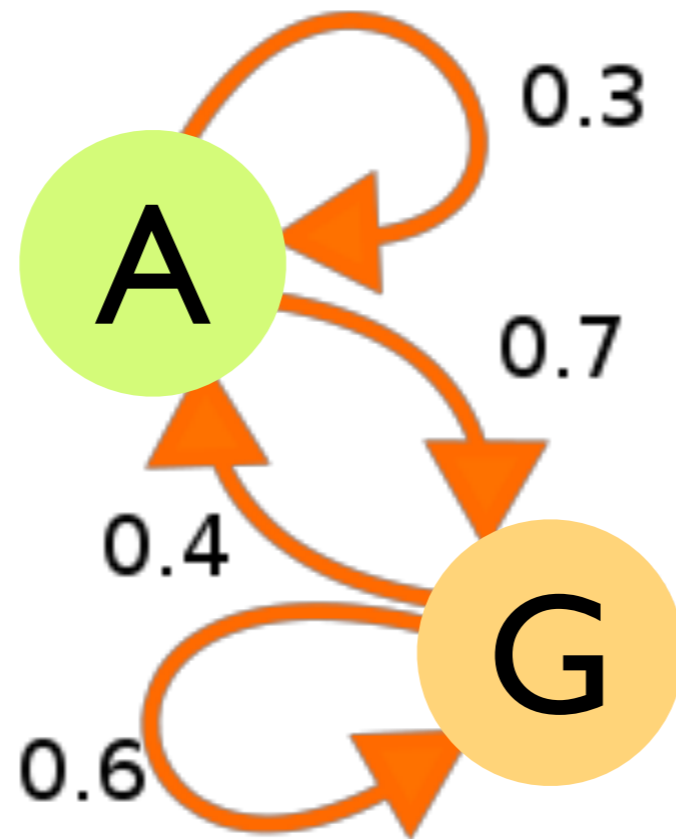
A **Markov chain**, named for Andrey Markov, is a mathematical system that undergoes transitions from one state to another in a chainlike manner. It is a **random process** characterized as memoryless: the next state depends only on the current state and not on the entire past. This specific kind of "memorylessness" is called the Markov property. Markov chains have many applications as statistical models of real-world processes.

Blocks **Su**bstitution **M**atrix (**BLOSUM**). A substitution matrix in which scores for each position are derived from observations of the frequencies of substitutions in blocks of local alignments in related proteins. Each matrix is tailored to a particular evolutionary distance. In the BLOSUM62 matrix, for example, the alignment from which scores were derived was created using sequences sharing no more than 62% identity. Sequences more identical than 62% are represented by a single sequence in the alignment so as to avoid over-weighting closely related family members. (Henikoff and Henikoff 1992)

The BLOSUM62 matrix

$$S_{ij} = \left(\frac{1}{\lambda}\right) \log\left(\frac{p_{ij}}{p_i * q_j}\right)$$

p_{ij} is the probability of two amino acids i and j replacing each other in a homologous sequence, and q_i and q_j are the background probabilities of finding the amino acids i and j in any protein sequence at random. The factor λ is a scaling factor, set such that the matrix contains easily computable integer values.

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	-1	5																		
Asn	-2	0	6																	
Asp	-2	-2	1	6																
Cys	0	-3	-3	-3	9															
Gln	-1	1	0	0	-3	5														
Glu	-1	0	0	2	-4	2	5													
Gly	0	-2	0	-1	-3	-2	-2	6												
His	-2	0	1	-1	-3	0	0	-2	8											
Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

Bioinformatics ▶ BLAST Search

Note: Parameter values that differ from the default are highlighted in yellow and marked with ♦ sign

▼ Algorithm parameters

General Parameters

Max target sequences [100 ▲▼]
Select the maximum number of aligned sequences to display ⊚

Short queries ☑ Automatically adjust parameters for short input sequences ⊚

Expect threshold [10] ⊚

Word size [28 ▲▼] ⊚

Max matches in a query range [0] ⊚

Scoring Parameters

Match/Mismatch Scores [1,-2 ▲▼] ⊚

Gap Costs [Linear ▲▼] ⊚

Filters and Masking

Filter ☑ Low complexity regions ⊚
☐ Species-specific repeats for: [Human ▲▼] ⊚

Mask ☑ Mask for lookup table only ⊚
☐ Mask lower case letters ⊚

BLAST Search **database Nucleotide collection (nr/nt)** using **Megablast (Optimize for highly similar sequences)**
☐ Show results in a new window

▼ **Algorithm parameters**

▶ Note: Parameter values that differ from the default are highlighted in yellow and marked with ♦ sign

General Parameters

Max target sequences	[100 ▲▼]
	Select the maximum number of aligned sequences to display ⓗ
Short queries	☑ Automatically adjust parameters for short input sequences ⓗ
Expect threshold	[10] ⓗ
Word size	[28 ▲▼] ⓗ
Max matches in a query range	[0] ⓗ

Scoring Parameters

Match/Mismatch Scores	[1,−2 ▲▼] ⓗ
Gap Costs	[Linear ▲▼] ⓗ

Filters and Masking

Filter	☑ Low complexity regions ⓗ
	☐ Species-specific repeats for: [Human ▲▼] ⓗ
Mask	☑ Mask for lookup table only ⓗ
	☐ Mask lower case letters ⓗ

BLAST

Search **database Nucleotide collection (nr/nt)** using **Megablast (Optimize for highly similar sequences)**

☐ Show results in a new window