Pairwise alignment

Ali Etemadi

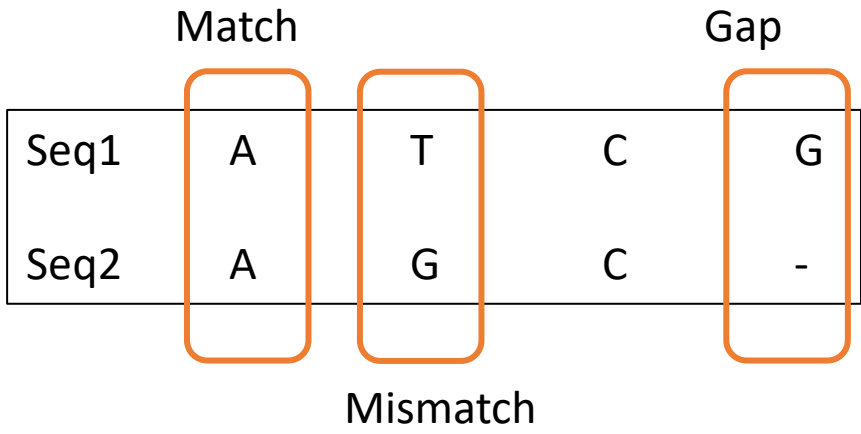
Tehran University of Medical Sciences

Introduction

- - The fundamental question in molecular biology: Is a gene or protein **related** to any other?
- - Relatedness at the sequence level implies **homology**, suggesting shared **evolutionary** ancestry.
- - Relatedness also implies potential **shared functions** among proteins.

What we can do with Sequence Analysis:

- - Analysis of DNA and protein sequences reveals **shared domains and motifs** among molecules.
- - Alignment of sequences facilitates the exploration of **relatedness** between proteins and genes.
- - Crucial for deciphering intra-organismal and inter-organismal protein relationships, advancing our **comprehension of life processes**.

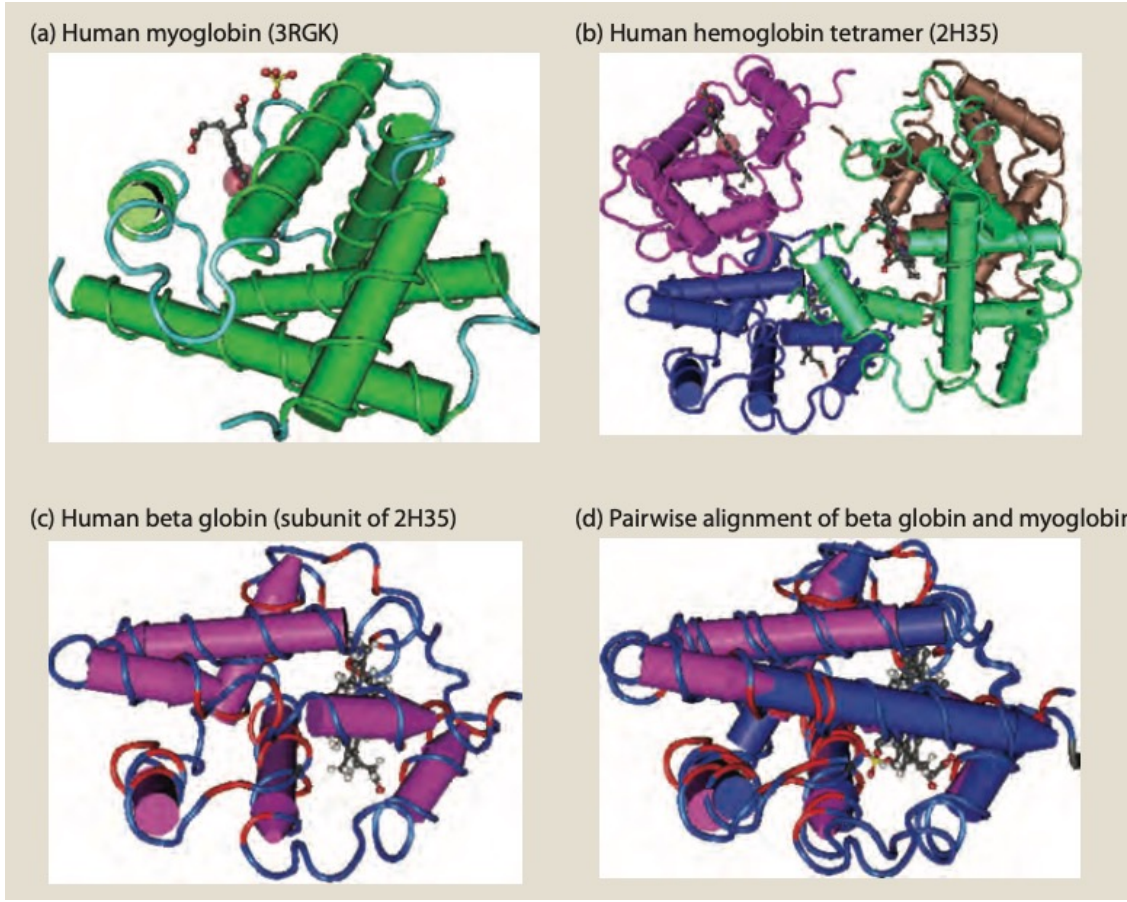


pairwise alignment

- **lining up two sequences** to achieve **maximal levels of identity** (and maximal levels of conservation in the case of amino acid alignments).
- to assess the **degree of similarity** and the **possibility of homology** between two molecules.
- If the amount of sequence identity is sufficient, then the two sequences are probably homologous. It is never correct to say that two proteins share a certain percent homology, because they are **either homologous or not**.
- Similarly, it is not appropriate to describe two sequences as “highly homologous;” instead, it can be said that they share a high degree of similiarity.
- the strongest evidence to determine whether two proteins are homologous comes from structural studies in combination with **evolutionary analyses**

Comparing Protein and DNA Sequence Alignment

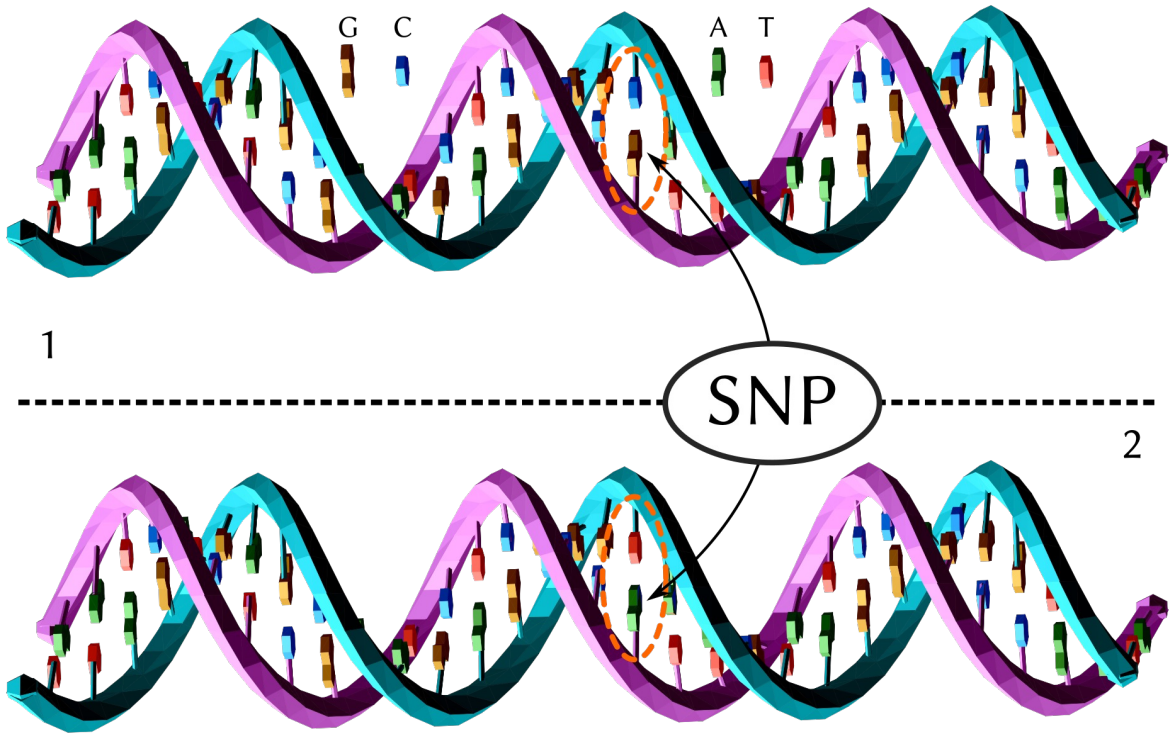
- - Protein sequence alignment is often **more informative** than DNA sequence alignment.
- - Many changes in DNA sequences, particularly at the third position of a codon, do not alter the specified amino acid due to **codon redundancy**.
- - Amino acids often share related **biophysical properties**, making it challenging to differentiate between them solely based on DNA sequence.
- - Example: Lysine and arginine are both **basic amino acids**.
- - Ability to identify **homologous sequences**, which is often not feasible through DNA sequence comparisons.
- - Example: **TBLASTN** tool from NCBI BLAST website enables the search for **related proteins** using a protein sequence query, translating DNA sequences into potential protein sequences.



share very similar three-dimensional structures. However, pairwise alignment of the amino acid sequences of these proteins reveals that the proteins share very limited amino acid identity.

Importance of DNA Sequence Comparisons:

- - Searching for polymorphisms.
- - Analyzing cloned cDNA fragments.
- - Comparing regulatory regions



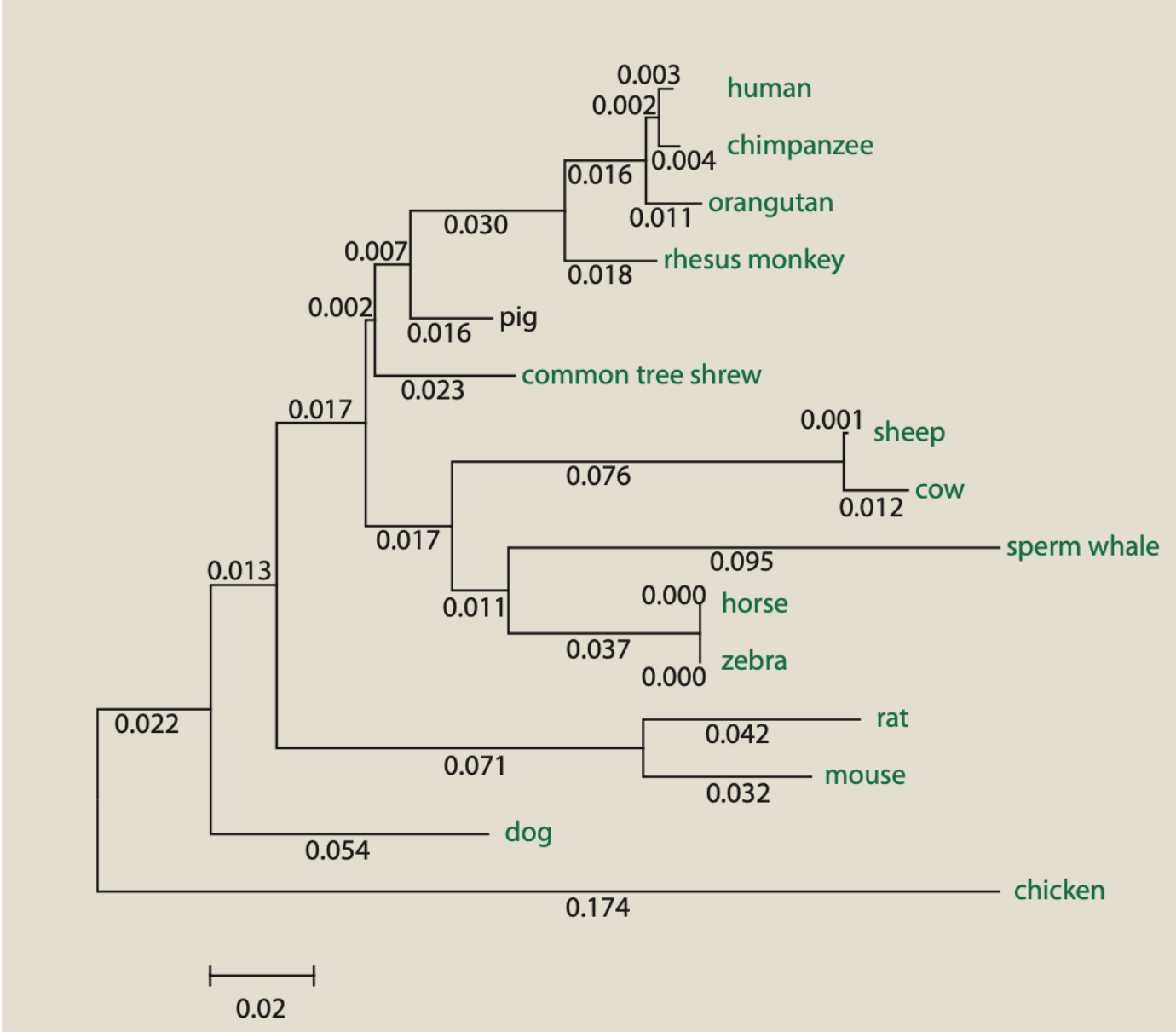
https://en.wikipedia.org/wiki/Single-nucleotide_polymorphism

Definitions: Homology, Similarity, Identity

- **Introduction to Homology:**
  - Homology refers to the **evolutionary** relationship between sequences.
  - Two sequences are considered homologous if they share a **common ancestry**.
  - Homologous proteins typically share a significantly related **three-dimensional** structure.
- **Quantitative Measures:**
  - Homology is **qualitative**; sequences are either homologous or not.
  - Identity and similarity are **quantitative** measures describing the relatedness of sequences.
  - Notably, homologous sequences may not share statistically significant identity.
- **Example from the Globin Family:**
  - Consideration of human myoglobin and beta globin as distantly related proteins.
  - Despite divergence, they share a similar three-dimensional structure.

Orthology vs. Paralogy

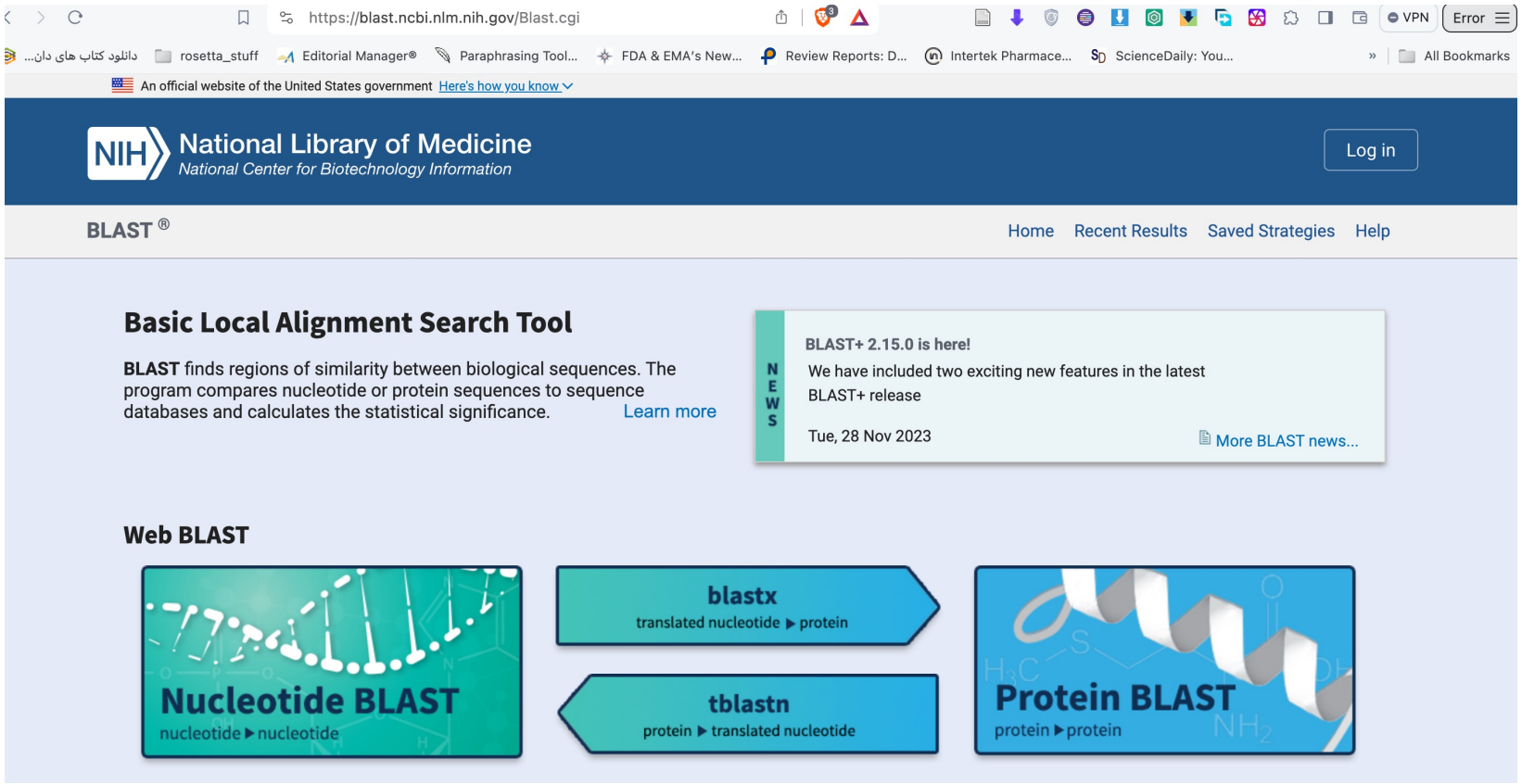
- **Orthology vs. Paralogy:**
  - Orthologs: Homologous sequences in different species arising from a common ancestral gene during speciation.
  - Paralogs: Homologous sequences resulting from gene duplication within the same species.
- **Functional Implications:**
  - Orthologs are presumed to have similar biological functions, while paralogs may have diverged functions.
  - Notably, homologous proteins may have entirely distinct functions.
- **Identification of Homologs:**
  - Homologous sequences are identified through database searches based on significant alignment scores.
  - Homologous proteins may have distinct functions despite sharing evolutionary ancestry.

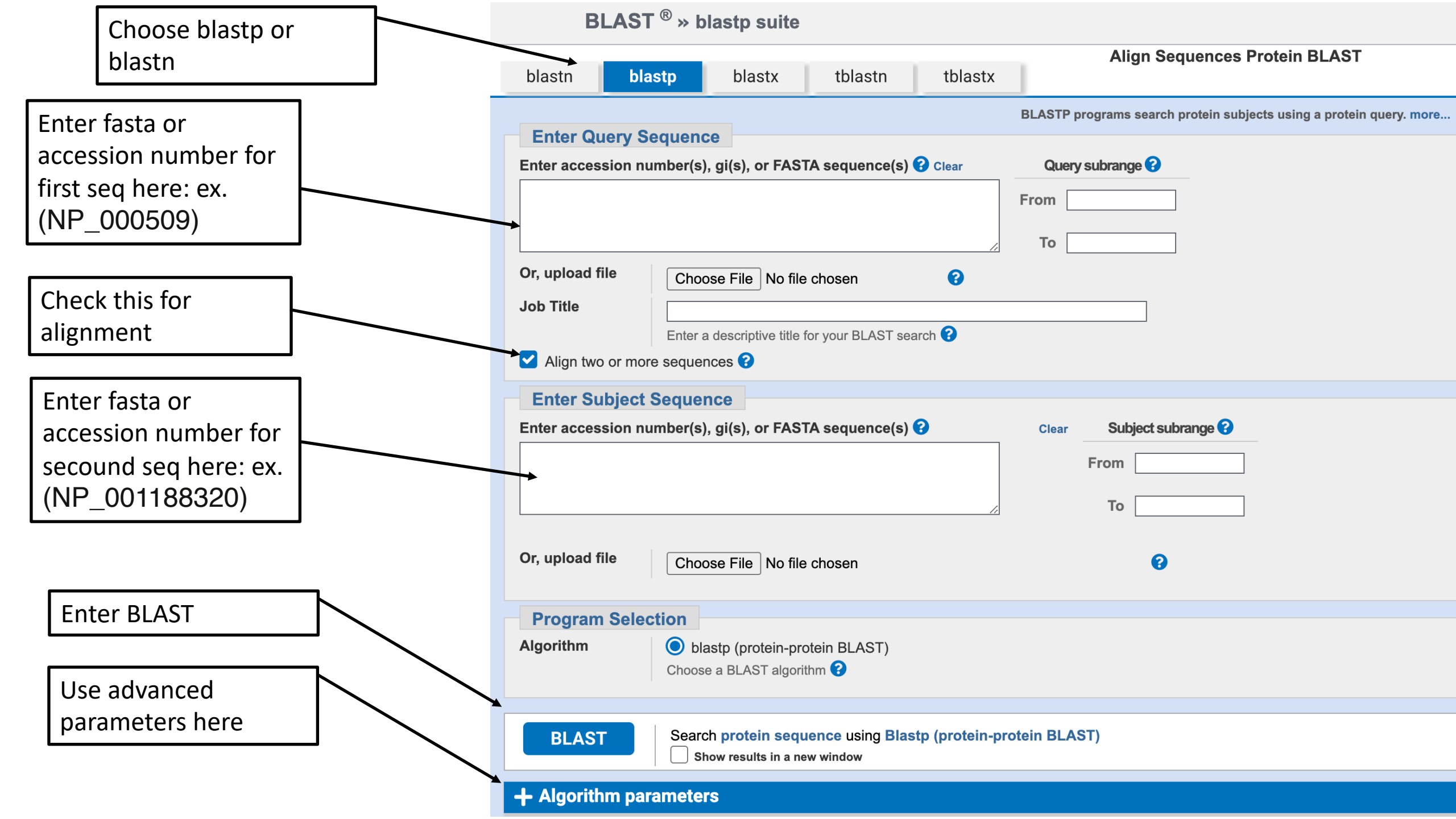


A group of myoglobin orthologs, visualized by multiply aligning the sequences then creating a phylogenetic tree by neighbor-joining

pairwise alignment

- NCBI BLASTP tool (for proteins) or BLASTN (for nucleotides)




Choose blastp or blastn
Enter fasta or accession number for first seq here: ex. (NP_000509)
Check this for alignment
Enter fasta or accession number for secound seq here: ex. (NP_001188320)
Enter BLAST
Use advanced parameters here
BLAST ® » blastp suite
Align Sequences Protein BLAST
blastn
blastp
blastx
tblastn
tblastx
BLASTP programs search protein subjects using a protein query. more...
Enter Query Sequence
Enter accession number(s), gi(s), or FASTA sequence(s)
Clear
Query subrange
From
To
Or, upload file
Choose File
No file chosen
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
Subject subrange
From
To
Or, upload file
Choose File
No file chosen
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

Job Title	**NP_000509:hemoglobin subunit beta [Homo sapiens]**
RID	0640PYRH114 *Search expires on 03-28 02:12 am* Download All
Program	Blast 2 sequences Citation
Query ID	NP_000509.1 (amino acid)
Query Descr	hemoglobin subunit beta [Homo sapiens]
Query Length	147
Subject ID	NP_001188320.1 (amino acid)
Subject Descr	hemoglobin, beta adult s chain [Mus musculus]
Subject Length	147
Other reports	Multiple alignment MSA viewer

Filter Results

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

Filter Reset

Descriptions | Graphic Summary | Alignments | Dot Plot

Sequences producing significant alignments

Download | Select columns | Show 100

select all *1 sequences selected*

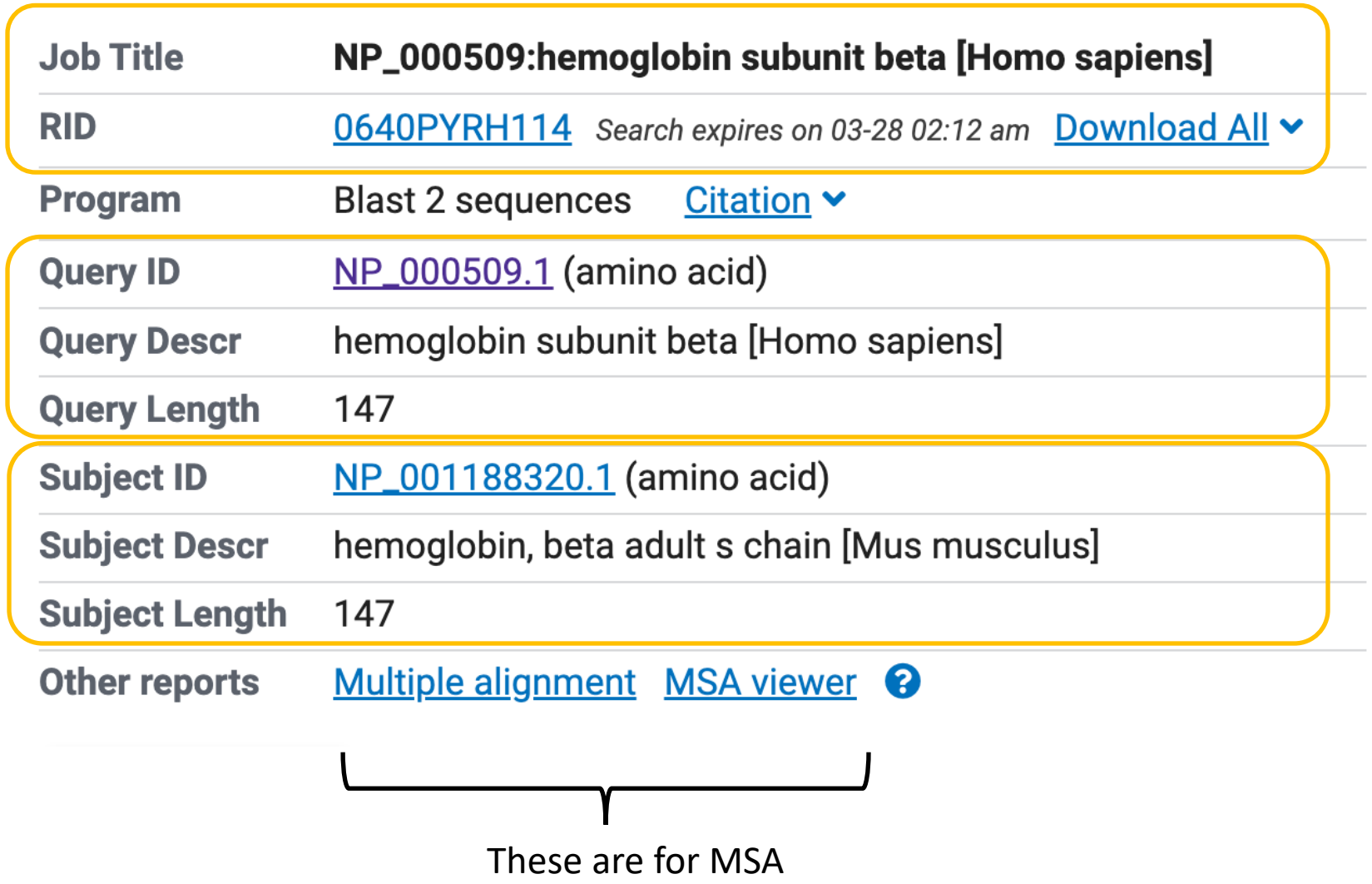
GenPept Graphics Multiple alignment MSA Viewer

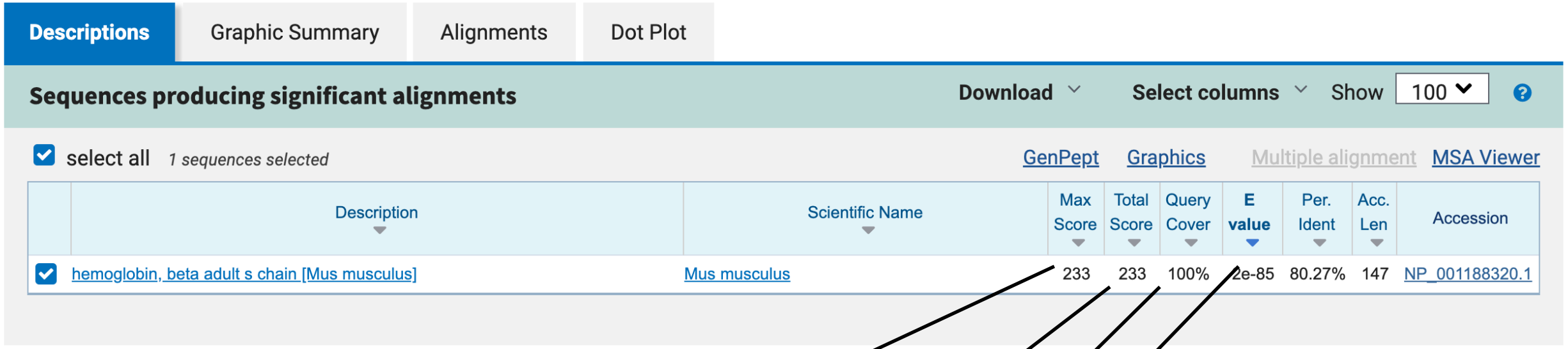
	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	hemoglobin, beta adult s chain [Mus musculus]	Mus musculus	233	233	100%	2e-85	80.27%	147	NP_001188320.1

Temporarily job name and ID

Info for first sequence

Info for second sequence





Max[imum] **Score:** the highest alignment score calculated from the sum of the rewards for matched aa and penalities for mismatches and gaps.

•**Total Score:** the sum of alignment scores of all segments from the same subject sequence.

•**Query Cover**[age]: the percent of the query length that is included in the aligned segments.

•**E**[xpect] **Value:** the number of alignments expected by chance with the calculated score or better. The expect value is the default sorting metric; for significant alignments the E value should be very close to zero.

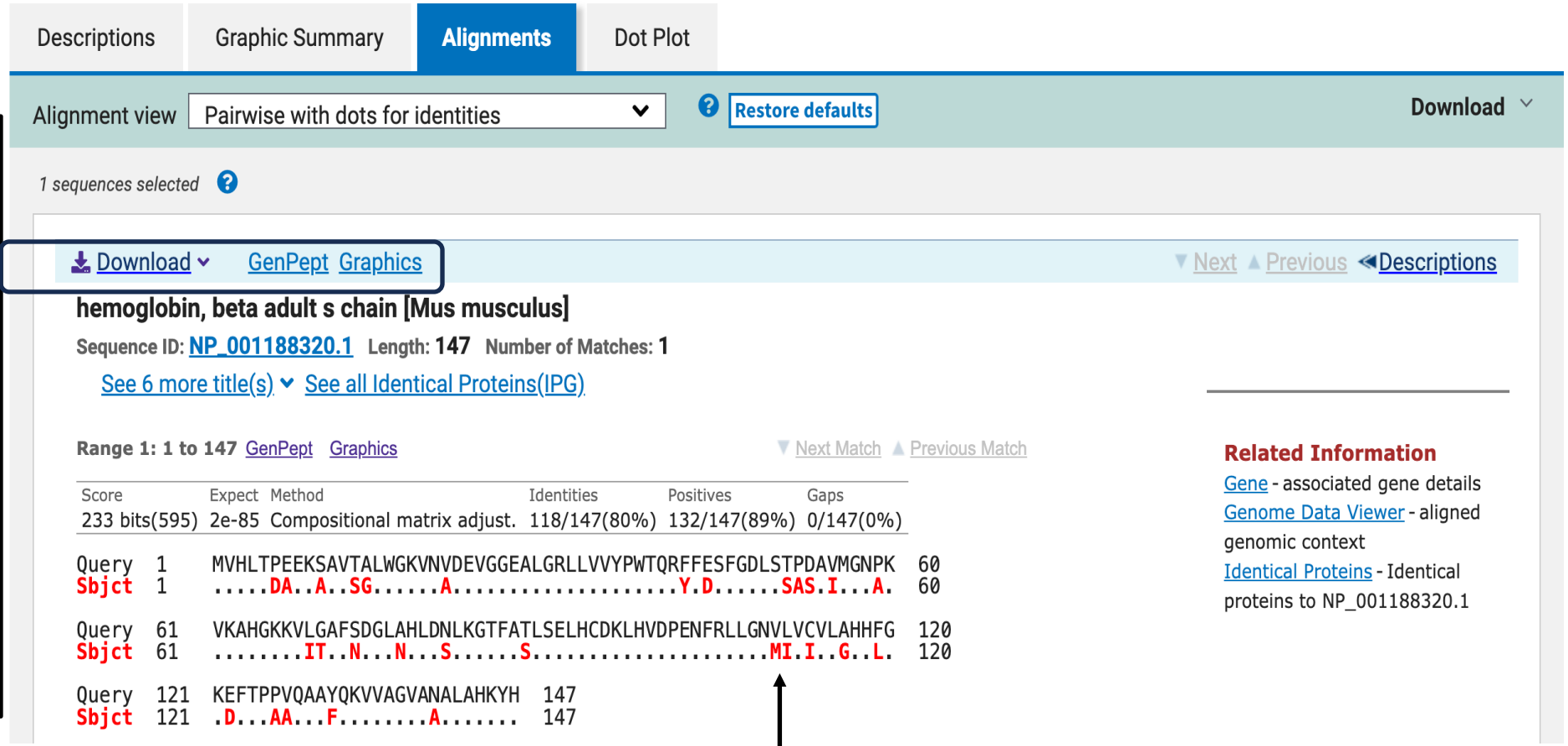
•**Ident**[ity]: the highest percent identity for a set of aligned segments to the same subject sequence.

•**Acc**[ession] **Len**[gth]: the number of nucleotides or amino acids in the result sequence identified by the accession number

•**Accession** [number]: a unique identifier assigned to records in the NCBI databases

Viewing your results
Under the **Alignments** tab next to **Alignment view** select **Pairwise with dots for identities**.

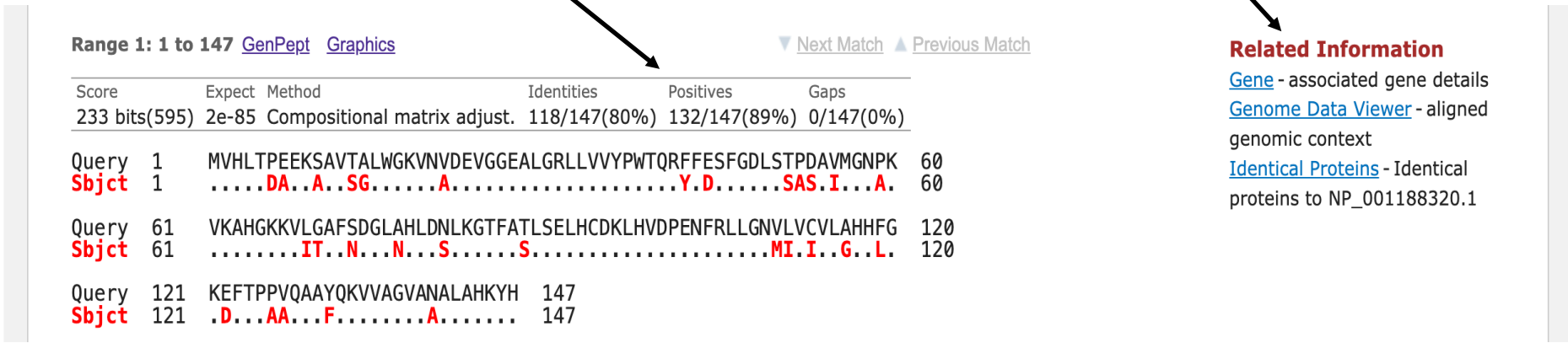
Download this matched database sequence.
GenePept retrieves the record from the NCBI Protein database.
Graphics links to a graphical sequence viewer-based display of the alignment between this database sequence and the query, anchored by that database sequence.



Dots for identities,
any differing amino acid in the subject sequence will be displayed in red

the number of amino acids that are either identical between the query and the subject sequence or have similar chemical properties. Amino acids with similar properties include the basic amino acids (K, R, H), acidic amino acids (D, E), hydroxylated amino acids (S, T), and hydrophobic amino acids (W, F, Y, L, I, V, M, A).

link to other resources with additional information derived from the matched database sequence.



Range 1: 1 to 147 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
233 bits(595)	2e-85	Compositional matrix adjust.	118/147(80%)	132/147(89%)	0/147(0%)

```
Query  1    MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK  60
Sbjct  1    .....DA..A..SG......A.....................Y.D......SAS.I...A.  60

Query  61   VKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG  120
Sbjct  61   ........IT..N...N...S......S.....................MI.I..G..L.  120

Query  121  KEFTPPVQAAYQKVVAGVANALAHKYH  147
Sbjct  121  .D...AA...F........A.......  147
```

Related Information

Gene - associated gene details

Genome Data Viewer - aligned genomic context

Identical Proteins - Identical proteins to NP_001188320.1

Alignment view Pairwise

Restore defaults

Download GenPept Graphics

hemoglobin, beta adult s chain [Mus musculus]

Sequence ID: NP_001188320.1 **Length:** 147 **Number of Matches:** 1

See 6 more title(s) See all Identical Proteins(IPG)

Range 1: 1 to 147 GenPept Graphics Next Match Previous Match

Score	Expect	Method	Identities	Positives	Gaps
233 bits(595)	2e-85	Compositional matrix adjust.	118/147(80%)	132/147(89%)	0/147(0%)

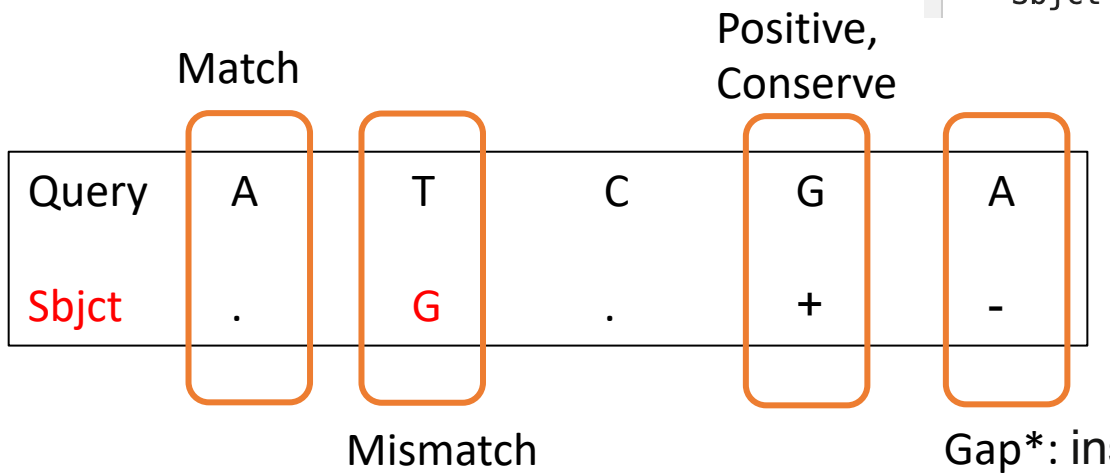
```
Query  1    MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK  60
            MVHLT  EK+AV+ LWGKVN DEVGGEALGRLLVVYPWTQR+F+SFGDLS+  A+MGN K
Sbjct  1    MVHLTDAEKAAVSGLWGKVNADEVGGEALGRLLVVYPWTQRYFDSFGDLSSASAIMGNAK  60

Query  61   VKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG  120
            VKAHGKKV+ AF+DGL HLD+LKGTFA+LSELHCDKLHVDPENFRLLGN++V VL HH G
Sbjct  61   VKAHGKKVITAFNDGLNHLDSLKGTFASLSELHCDKLHVDPENFRLLGNMIVIVLGHHLG  120

Query  121  KEFTPPVQAAYQKVVAGVANALAHKYH  147
            K+FTP  QAA+QKVVAGVA ALAHKYH
Sbjct  121  KDFTPAAQAAFQKVVAGVAAALAHKYH  147
```

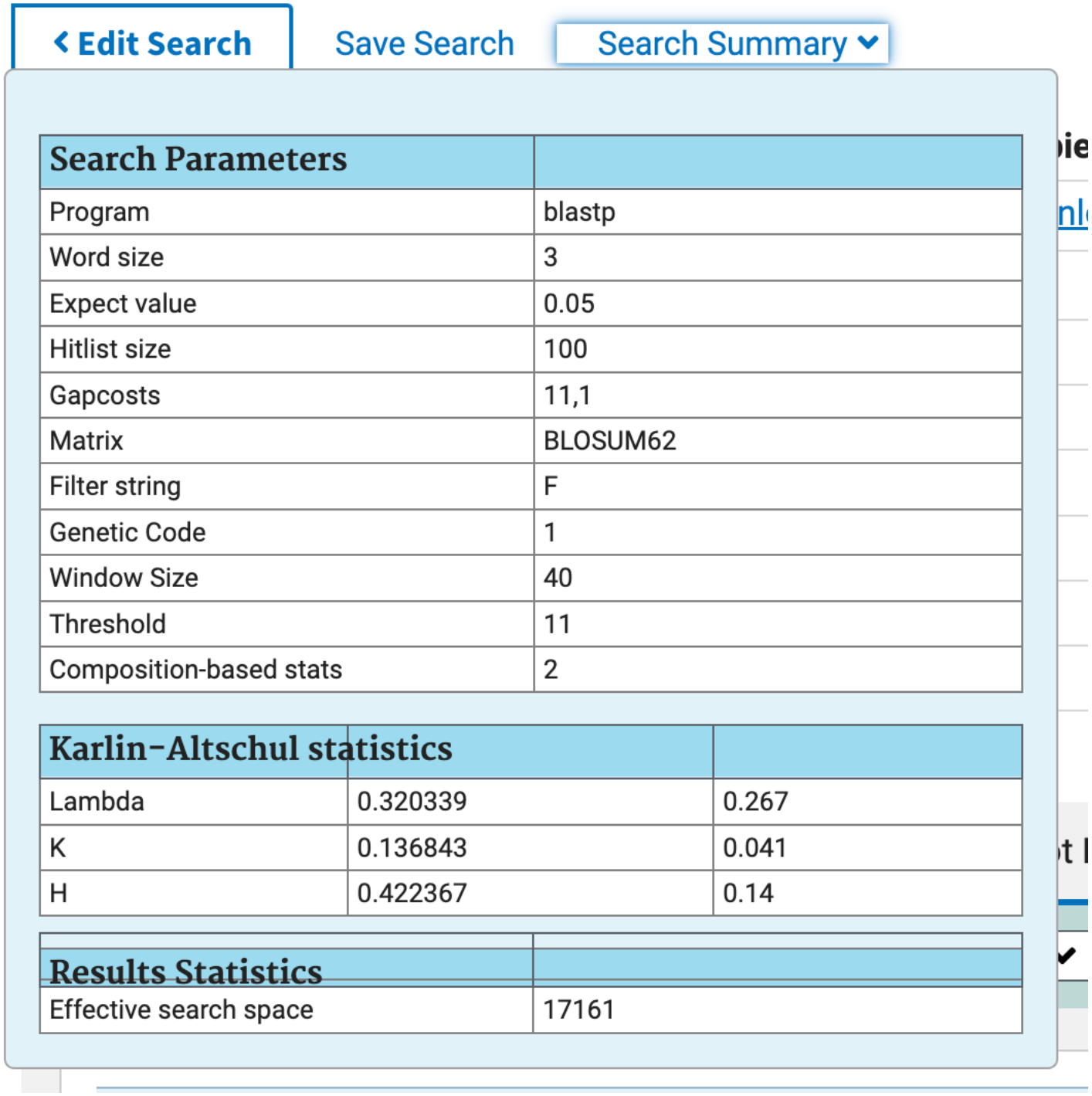
A space in the matching sequence represents amino acids that are completely different.

In addition, the „+" character denotes amino acids that are different between the query and subject sequences but the two residues have similar chemical properties.



Search summary
Default search mode

- **Saving your results**
- To save your search queries and settings, click on the **Save Search** link, then log in to My NCBI using the **Sign in** or **Register** link at the upper right. Once you do this, your search strategies should appear in the **Saved Search Strategies** tab.



Search Parameters	
Program	blastp
Word size	3
Expect value	0.05
Hitlist size	100
Gapcosts	11,1
Matrix	BLOSUM62
Filter string	F
Genetic Code	1
Window Size	40
Threshold	11
Composition-based stats	2

Karlin-Altschul statistics		
Lambda	0.320339	0.267
K	0.136843	0.041
H	0.422367	0.14

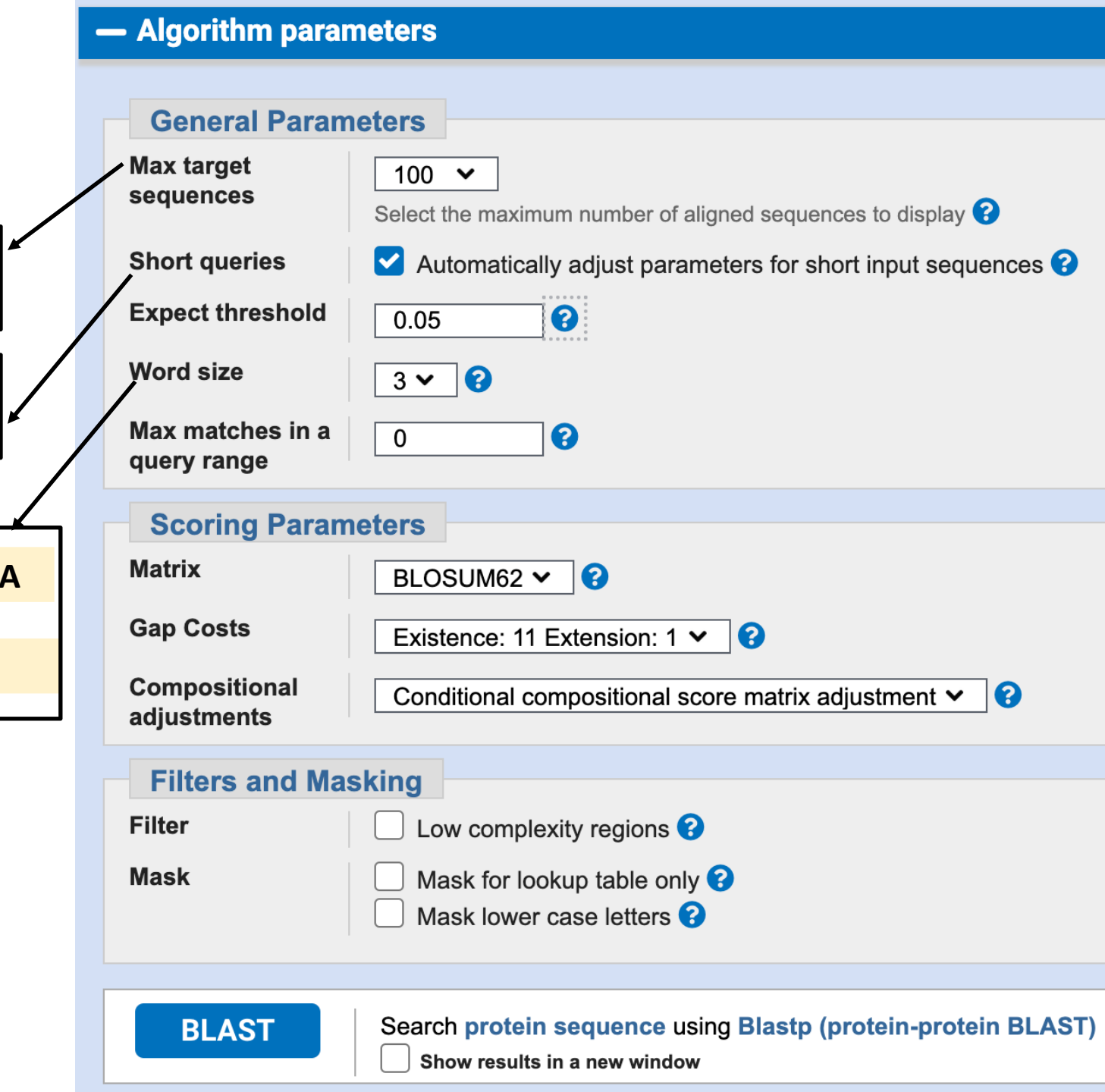
Results Statistics	
Effective search space	17161

Algorithm parameters

Select the maximum number of aligned sequences to display

Automatically adjust word size and other parameters to improve results for short queries

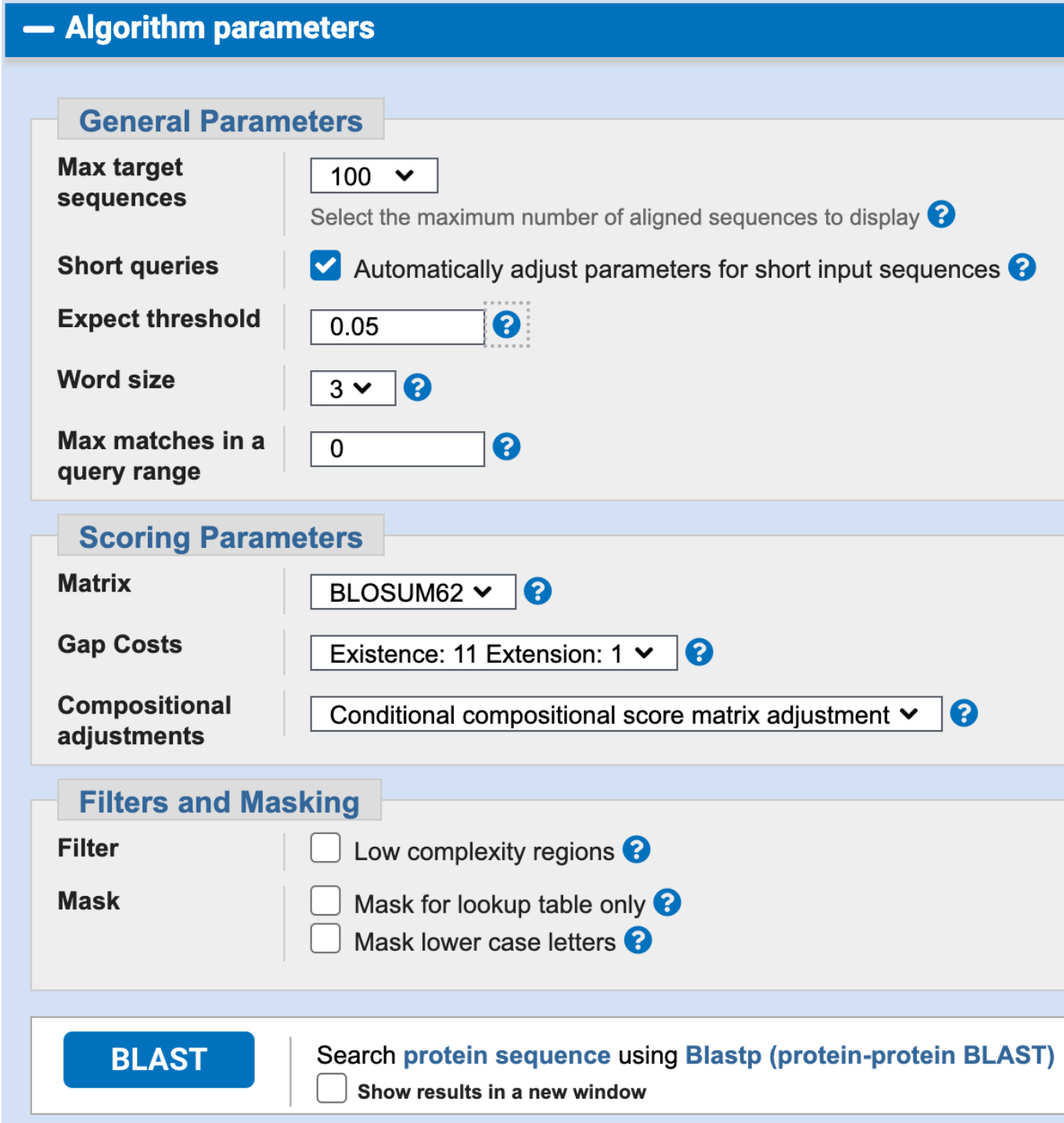
Query 1	V	N	V	D	E	V	G	G	E	A
Sbjct 1				D	E	V				



Algorithm parameters: Expect Value (E)

- the number of hits expected to occur by chance when searching a database of a specific size.
- - It decreases exponentially as the match score (S) increases, representing the random background noise.

- - Interpretation of E-value:
- - Lower E-values indicate more significant matches.
- - An E-value of 1 suggests that in a database of the same size, one expects to see 1 match with a similar or higher score purely by chance.

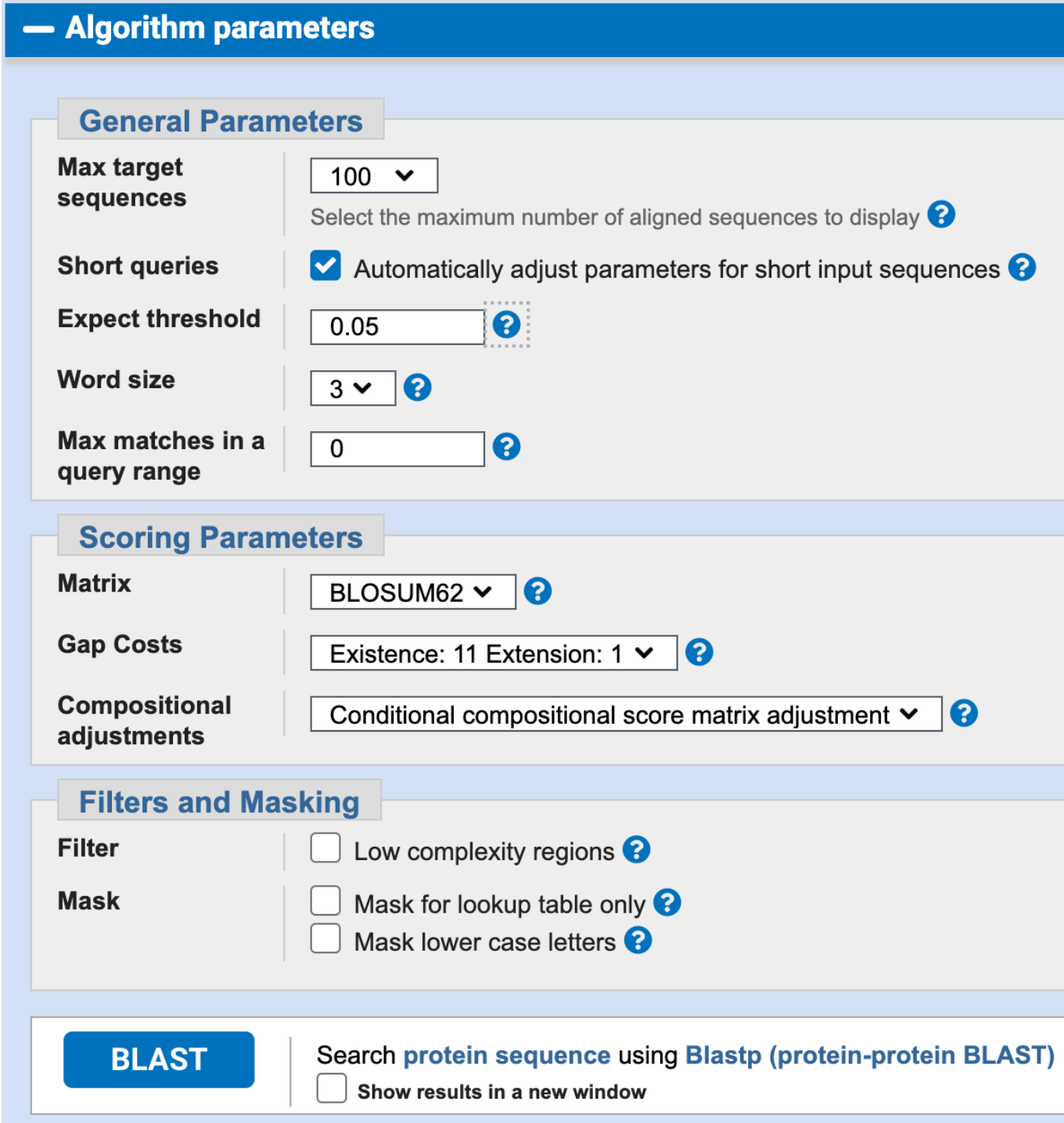


Considerations for Short Alignments

- - Virtually identical short alignments tend to have relatively high E-values.
- - Shorter sequences have a higher probability of occurring in the database by chance, affecting E-value calculations.

- **- Use of E-value for Significance Threshold:**
- - E-value serves as a convenient method to set a significance threshold for reporting results.

- **- Effect of Lower E-value Thresholds:**
- - Lowering the E-value threshold increases stringency, resulting in fewer chance matches being reported.



Understanding Gaps in Sequence Alignment

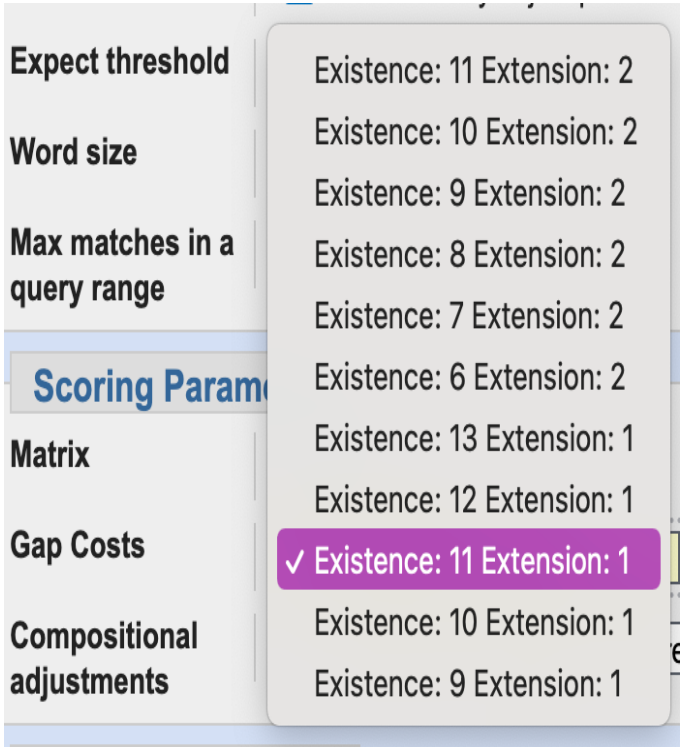
Pairwise alignment helps identify mutations causing sequence **divergence** during evolution.

- Types of Mutations:

- **Substitutions**: Mutation changing the codon for one amino acid into another.
- **Insertions and deletions:** Addition or removal of residues, represented by dashes in sequences.

- Scoring Scheme and Gap Penalties:

- Two gap penalties:
  1. Penalty (score −a) for creating a gap.
  2. Penalty (score −b) for each residue that a gap extends.
- Gap of length k receives a penalty of −(a + bk).
- Example: For a gap of length 1, the score is −(a + b).



Limitations of Traditional Identity Scoring (0 0r 1) in Protein Single Amino Acid Mutation

Equal Weighting: Traditional scoring methods assign equal weight to all amino acid substitutions, regardless of their functional or structural significance.

Complexity of Proteins: Proteins consist of 20 amino acids, each with unique chemical properties and functional roles, making them more complex than nucleotides.

Functional Significance: Substitutions can have varying functional consequences; conservative changes may be functionally neutral, while non-conservative ones could disrupt critical interactions.

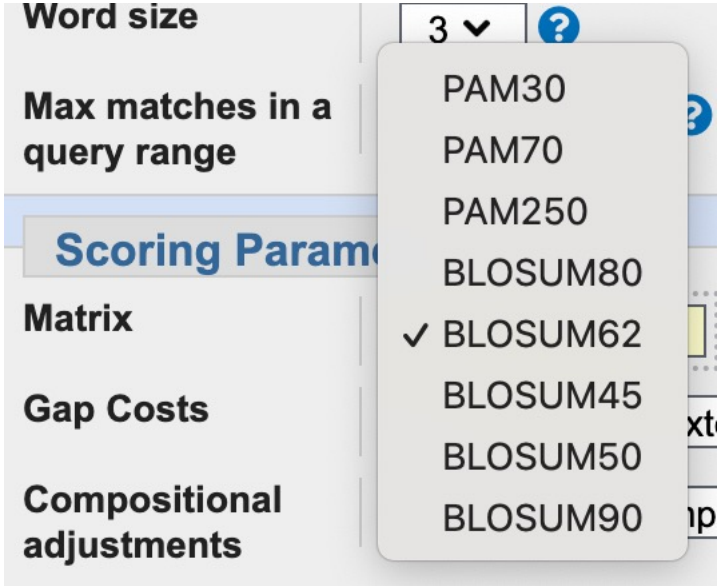
Frequency Bias: Certain substitutions occur more frequently due to evolutionary constraints or mutational biases, but traditional methods do not consider their prevalence.

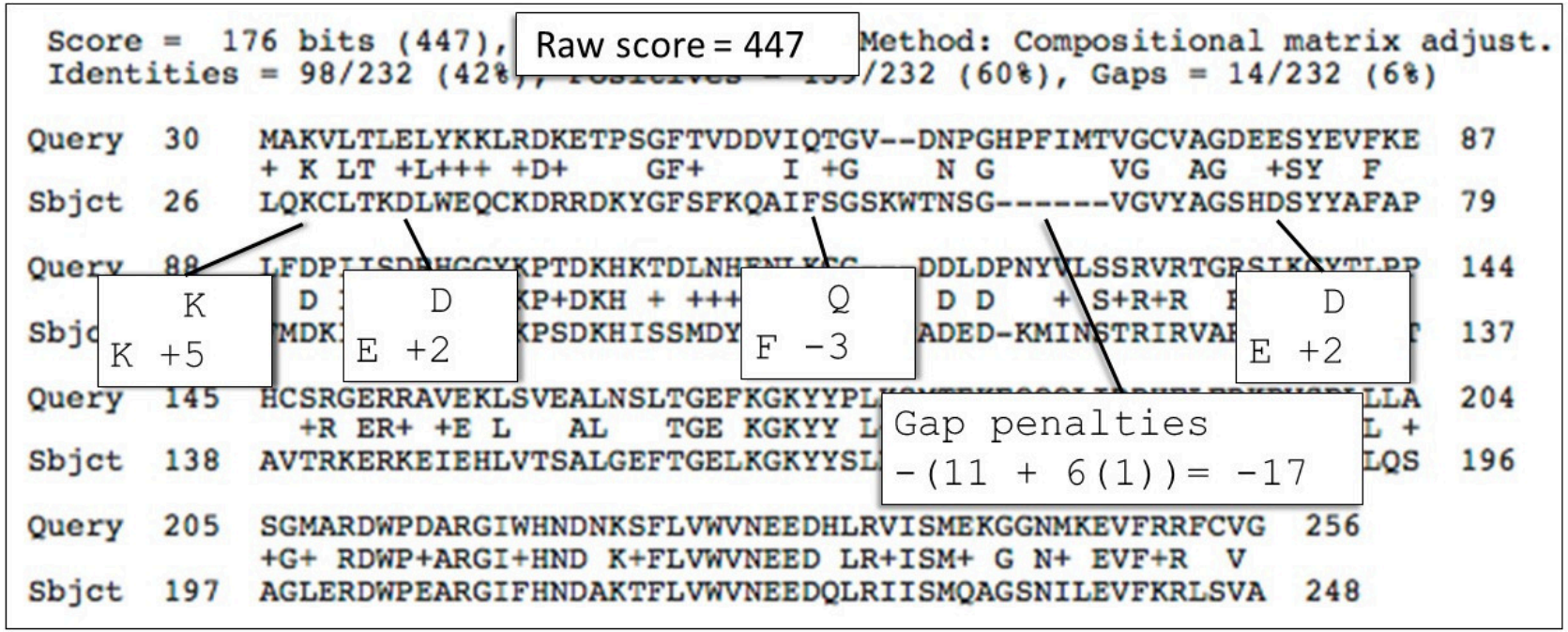
Advanced Scoring Methods: Matrices like PAM and BLOSUM offer a more nuanced approach, considering amino acid properties, substitution frequencies, and evolutionary distances.

Improved Alignment Accuracy: By incorporating these factors, advanced scoring methods enhance alignment accuracy and facilitate more meaningful analyses of protein evolution and function.

Algorithm parameters: Matrix

1. **Scoring in Protein Alignment:**
    1. When aligning two proteins, specific **scores** are assigned to residues based on their **matches or mismatches**.
    2. These **scores are critical** for evaluating the **quality** and **significance** of the alignment.
2. **Derivation of Scores for Matches and Mismatches:**
    1. The scores for matches and mismatches in protein alignment are derived **from evolutionary models**.
    2. Margaret **Dayhoff's** work, particularly from 1966 and 1978, provided insights into the rules governing evolutionary changes in proteins.
3. **Dayhoff Model:**
    1. Margaret Dayhoff proposed a comprehensive model outlining the rules of evolutionary change in proteins.
    2. The Dayhoff model, described in **seven steps**, serves as the foundation for developing a quantitative scoring system for pairwise alignments.
4. **Quantitative Scoring System:**
    1. Dayhoff's model forms the basis of a **quantitative** scoring system applicable to aligning any proteins, **irrespective** of their evolutionary relationship.
5. **BLOSUM Matrices:**
    1. Steven **Henikoff** and Jorja G. Henikoff introduced the **BLOSUM** (Blocks Substitution Matrix) matrices.
    2. These matrices are based on observed **substitutions in conserved regions** of aligned protein sequences.





Dayhoff Model Step 1 – Finding Accepted Point Mutations

DAYHOFF'S PROTEIN SUPERFAMILIES

(34 protein "superfamilies" grouped into 71 phylogenetic trees)

Approach: cataloging proteins and comparing (**alignment**) sequences of closely related proteins across various families.

rate of mutation acceptance

PROTEIN	PAMS PER 100 MILLION YEARS
Immunoglobulin (Ig) kappa chain C region	37
Kappa casein	33
Epidermal growth factor	26
Serum albumin	19
Hemoglobin alpha chain	12
Myoglobin	8.9
Nerve growth factor	8.5
Trypsin	5.9
Insulin	4.4
Cytochrome c	2.2
Glutamate dehydrogenase	0.9
Histone H3	0.14
Histone H4	0.10

	A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly	H His	I Ile	L Leu	K Lys	M Met	F Phe	P Pro	S Ser	T Thr	W Trp	Y Tyr	V Val
A																				
R	30																			
N	109	17																		
D	154	0	532																	
C	33	10	0	0																
Q	93	120	50	76	0															
E	266	0	94	831	0	422														
G	579	10	156	162	10	30	112													
H	21	103	226	43	10	243	23	10												
I	66	30	36	13	17	8	35	0	3											
L	95	17	37	0	y	75	15	17	40	253										
K	57	477	322	85	0	147	104	60	23	43	39									
M	29	17	0	0	0	20	7	7	0	57	207	90								
F	20	7	7	0	0	0	0	17	20	90	167	0	17							
P	345	67	27	10	10	93	40	49	50	7	43	43	4	7						
S	772	137	432	98	117	47	86	450	26	20	32	168	20	40	269					
T	590	20	169	57	10	37	31	50	14	129	52	200	28	10	73	696				
W	0	27	3	0	0	0	0	0	3	0	13	0	0	10	0	17	0			
Y	20	3	36	0	30	0	10	0	40	13	23	10	0	260	0	22	23	6		
V	365	20	13	17	33	27	37	97	30	661	303	17	77	10	50	43	186	0	17	
	A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly	H His	I Ile	L Leu	K Lys	M Met	F Phe	P Pro	S Ser	T Thr	W Trp	Y Tyr	V Val

Results: Numbers of accepted point mutations, multiplied by 10, in 1572 cases of amino acid substitutions from closely related protein sequences.

Dayhoff Model Step 2 (of 7): Frequency of Amino Acids

TABLE 3.1 Normalized frequencies of amino acid. These values sum to 1. If the 20 amino acids were equally represented in proteins, these values would all be 0.05 (i.e., 5%); instead, amino acids vary in their frequency of occurrence.

Gly	0.089	Arg	0.041
Ala	0.087	Asn	0.040
Leu	0.085	Phe	0.040
Lys	0.081	Gln	0.038
Ser	0.070	Ile	0.037
Val	0.065	His	0.034
Thr	0.058	Cys	0.033
Pro	0.051	Tyr	0.030
Glu	0.050	Met	0.015
Asp	0.047	Trp	0.010

Dayhoff Model Step 3 (of 7): Relative Mutability of Amino Acids

TABLE 3.2 Relative mutabilities of amino acids. The value of alanine is arbitrarily set to 100.

Asn	134	His	66
Ser	120	Arg	65
Asp	106	Lys	56
Glu	102	Pro	56
Ala	100	Gly	49
Thr	97	Tyr	41
Ile	96	Phe	41
Met	94	Leu	40
Gln	93	Cys	20
Val	74	Trp	18

$$\textbf{relative mutability, Ser} = \frac{\text{number of times Ser was observed to mutate (mi)}}{\text{overall frequency of occurrence of Ser(fi)}} = \frac{nx}{1!} + \frac{n(n-1)x^2}{2!} + \cdots$$

To calculate **relative mutability**, they divided the number of times each amino acid was observed to mutate (mi) by the overall frequency of occurrence of that amino acid (fi).

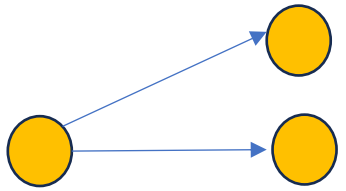
Why are some amino acids more mutable than others?

- The less mutable residues probably **have important structural or functional roles** in proteins, such that the consequence of replacing them with any other residue could be **harmful to the organism**.
- the most mutable amino acids – asparagine, serine, aspartic acid, and glutamic acid – have functions in proteins that are easily assumed by other residues. The most common substitutions are glutamic acid for aspartic acid (both are acidic), serine for alanine, serine for threonine (both are hydroxylated), and isoleucine for valine (both are hydrophobic and of a similar size).

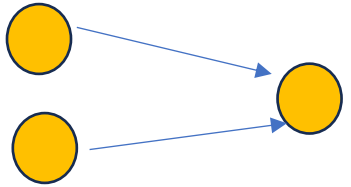
Dayhoff Model Step 4: Mutation Probability Matrix for 1 PAM

Evolution of Proteins Over Time

- **Divergent Evolution:**
  - Single ancestral protein diverges into multiple descendant proteins.
  - Result of different selective pressures or environmental conditions.
  - Accumulation of mutations leads to structural and functional changes.
  - Descendant proteins retain similarities due to common ancestry.



- **Convergent Evolution:**
  - Unrelated proteins independently evolve similar traits or functions.
  - Response to similar selective pressures or environmental challenges.
  - Independent acquisition of similar protein structures or functions.
  - Provides adaptive advantages in diverse evolutionary lineages.



Mutation probability matrix

- Dayhoff constructed a **mutation probability matrix, *M*** based on accepted mutations (st. 3) and amino acid occurrence probabilities (st. 2).
- Each element M_{ij} shows the likelihood of an original amino acid *j* being replaced by amino acid *i* over **one PAM**, representing **1% amino acid divergence**.

PAM1 mutation probability formula

- Equation for non-diagonal elements:

$$M_{ij} = \frac{\lambda m_j A_{ij}}{\sum_i A_{ij}}$$

- where *Aij* represents accepted mutations, *λ* is a constant, and *mj* is the mutability of amino acid *j*.
- Equation for diagonal elements:

$$M_{jj} = 1 - \lambda m_j$$

represents the probability of the original amino acid *j* **remaining unchanged**.

- Importance:
  - Provides insight into **likely amino acid substitutions**.
  - Forms the basis for scoring systems in **sequence alignment**.
- Assumption:
  - Accepted amino acid mutations are **undirected**.
  - In PAM1 matrix, ancestral residue **likely similar** to observed residues due to **close protein relationship**.

		Original amino acid																			
		A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly	H His	I Ile	L Leu	K Lys	M Met	F Phe	P Pro	S Ser	T Thr	W Trp	Y Tyr	V Val
Replacement amino acid	A	98.7	0.0	0.1	0.1	0.0	0.1	0.2	0.2	0.0	0.1	0.0	0.0	0.1	0.0	0.2	0.4	0.3	0.0	0.0	0.2
	R	0.0	99.1	0.0	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.0	0.2	0.0	0.0	0.0	0.1	0.0	0.1	0.0	0.0
	N	0.0	0.0	98.2	0.4	0.0	0.0	0.1	0.1	0.2	0.0	0.0	0.1	0.0	0.0	0.0	0.2	0.1	0.0	0.0	0.0
	D	0.1	0.0	0.4	98.6	0.0	0.1	0.5	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0
	C	0.0	0.0	0.0	0.0	99.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0
	Q	0.0	0.1	0.0	0.1	0.0	98.8	0.3	0.0	0.2	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0
	E	0.1	0.0	0.1	0.6	0.0	0.4	98.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	G	0.2	0.0	0.1	0.1	0.0	0.0	0.1	99.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.1
	H	0.0	0.1	0.2	0.0	0.0	0.2	0.0	0.0	99.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	I	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	98.7	0.1	0.0	0.2	0.1	0.0	0.0	0.1	0.0	0.0	0.3
	L	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.2	99.5	0.0	0.5	0.1	0.0	0.0	0.0	0.0	0.0	0.2
	K	0.0	0.4	0.3	0.1	0.0	0.1	0.1	0.0	0.0	0.0	0.0	99.3	0.2	0.0	0.0	0.1	0.1	0.0	0.0	0.0
	M	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.0	98.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	F	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.0	99.5	0.0	0.0	0.0	0.0	0.3	0.0
	P	0.1	0.1	0.0	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	99.3	0.1	0.0	0.0	0.0	0.0
	S	0.3	0.1	0.3	0.1	0.1	0.0	0.1	0.2	0.0	0.0	0.0	0.1	0.0	0.0	0.2	98.4	0.4	0.1	0.0	0.0
	T	0.2	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.1	0.1	0.0	0.1	0.3	98.7	0.0	0.0	0.1
	W	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	99.8	0.0	0.0
	Y	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	99.5	0.0
	V	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.6	0.1	0.0	0.2	0.0	0.0	0.0	0.1	0.0	0.0	99.0

FIGURE 3.9 The PAM1 mutation probability matrix. The original amino acid *j* is arranged in columns (across the top), while the replacement amino acid *i* is arranged in rows. Dayhoff et al. multiplied values by 10,000 (offering added precision) while here we multiply by 100 so that, for example, the first cell's value of 98.7 corresponds to 98.7% occurrence of ala remaining ala over this evolutionary interval.

- **Features of the matrix:**
  - Diagonal: Highest scores; indicates probability of amino acid remaining unchanged.
  - Column values sum to 100%.
  - Example: 98.7% probability of alanine remaining unchanged; 0.3% chance of being replaced by serine.

Accepted Point Mutation

- **PAM:** represents a **substitution** of one amino acid by another, **accepted** by **natural selection**.
- **Global alignment**
- **PAM1**: probability of amino acid substitutions equivalent to one mutation per 100 amino acids.
- The PAM1 matrix is based upon the alignment of **closely related** protein sequences, having an average of **1% change**.

Dayhoff Model Step 5 (of 7): PAM250 and Other PAM Matrices

Other PAMs

- - **PAM1** matrix based on closely related protein sequences with an average of **1% change**.
- Derivation of PAM Matrices:
- - **PAM100 or PAM250** reflect substitutions in **distantly related proteins**.
- - **Multiplying the PAM1 matrix by itself** produces other PAM matrices (up to hundreds of times).

- - PAM0: Unit diagonal matrix with no amino acid changes.
- - PAM∞: Equal likelihood of any amino acid, resembling background frequencies.
- - **PAM250**: Used in BLAST searches for proteins sharing about **20% amino acid identity**.

original amino acid

replacement amino acid

PAM0	A	R	N	D	C	Q	E	G
A	100	0	0	0	0	0	0	0
R	0	100	0	0	0	0	0	0
N	0	0	100	0	0	0	0	0
D	0	0	0	100	0	0	0	0
C	0	0	0	0	100	0	0	0
Q	0	0	0	0	0	100	0	0
E	0	0	0	0	0	0	100	0
G	0	0	0	0	0	0	0	100

original amino acid

PAM∞	A	R	N	D	C	Q	E	G
A	8.7	8.7	8.7	8.7	8.7	8.7	8.7	8.7
R	4.1	4.1	4.1	4.1	4.1	4.1	4.1	4.1
N	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0
D	4.7	4.7	4.7	4.7	4.7	4.7	4.7	4.7
C	3.3	3.3	3.3	3.3	3.3	3.3	3.3	3.3
Q	3.8	3.8	3.8	3.8	3.8	3.8	3.8	3.8
E	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0
G	8.9	8.9	8.9	8.9	8.9	8.9	8.9	8.9

Dayhoff Model Step 6 (of 7): From a Mutation Probability Matrix to a Relatedness Odds Matrix

Taking account the chance

- For the elements Mij of any given mutation probability matrix, what is the probability that amino acid j will change to i in a **homologous sequence**?

$$R_{ij} = \frac{M_{ij}}{f_i}.$$

- the normalized frequency fi is the probability of amino acid residue i occurring in the second sequence by chance.
- Values of *Rij*:
- 1 means substitution occurs as expected by chance.
- >1: Substitution occurs more often than expected.
- <1: Substitution not favored.

Dayhoff Model Step 7 (of 7): Log-Odds Scoring Matrix

Log-Odds Scoring Matrix

- **Log-Odds Formulation:**
  - $s = 10 \times \log_{10}\left(\frac{M_{ij}}{f_i}\right)$
  - M_{ij}: Observed frequency of substitution for each pair of amino acids.
  - f_i: Background frequency of replacement amino acid i.
- **Interpretation of Scores:**
  - **Positive** Score: Substitution **more frequent** than expected by chance.
  - **Negative** Score: Substitution **less frequent** than expected by chance.
  - Zero Score: Neutral.
- **Example:**
  - Substitution from cysteine to leucine:
    - $M_{ij} = 0.02$ (PAM250 matrix), $f_i = 0.085$ (normalized frequency of leucine).
    - $s = 10 \times \log_{10}\left(\frac{0.02}{0.085}\right) = -6.3$ (cysteine, leucine).

A	2																		
R	-2	6																	
N	0	0	2																
D	0	-1	2	4															
C	-2	-4	-4	-5	12														
Q	0	1	1	2	-5	4													
E	0	-1	1	3	-5	2	4												
G	1	-3	0	1	-3	-1	0	5											
H	-1	2	2	1	-3	3	1	-2	6										
I	-1	-2	-2	-2	-2	-2	-2	-3	-2	5									
L	-2	-3	-3	-4	-6	-2	-3	-4	-2	-2	6								
K	-1	3	1	0	-5	1	0	-2	0	-2	-3	5							
M	-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6						
F	-3	-4	-3	-6	-4	-5	-5	-5	-2	1	2	-5	0	9					
P	1	0	0	-1	-3	0	-1	0	0	-2	-3	-1	-2	-5	6				
S	1	0	1	0	0	-1	0	1	-1	-1	-3	0	-2	-3	1	2			
T	1	-1	0	0	-2	-1	0	0	-1	0	-2	0	-1	-3	0	1	3		
W	-6	2	-4	-7	-8	-5	-7	-7	-3	-5	-2	-3	-4	0	-6	-2	-5	17	
Y	-3	-4	-2	-4	0	-4	-4	-5	0	-1	-1	-4	-2	7	-5	-3	-3	0	10
V	0	-2	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	-2	4
	A	**R**	**N**	**D**	**C**	**Q**	**E**	**G**	**H**	**I**	**L**	**K**	**M**	**F**	**P**	**S**	**T**	**W**	**Y**	**V**

Log-odds matrix for PAM250. High PAM values (e.g., PAM250) are useful for aligning very divergent sequences. A variety of algorithms for pairwise alignment, multiple sequence alignment, and database searching (e.g., BLAST) allow you to select an assortment of PAM matrices such as PAM250, PAM70, and PAM30.

- **Comparison Across PAM Matrices:**
  - PAM250 vs. PAM10:
    - PAM10 favors higher scores for identical residue pairs.
    - Greater penalties for mismatches in PAM10 compared to PAM250.
    - Negative scores for certain substitutions in PAM10 scored positively in PAM250.

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	7																			
R	-10	9																		
N	-7	-9	9																	
D	-6	-17	-1	8																
C	-10	-11	-17	-21	10															
Q	-7	-4	-7	-6	-20	9														
E	-5	-15	-5	0	-20	-1	8													
G	-4	-13	-6	-6	-13	-10	-7	7												
H	-11	-4	-2	-7	-10	-2	-9	-13	10											
I	-8	-8	-8	-11	-9	-11	-8	-17	-13	9										
L	-9	-12	-10	-19	-21	-8	-13	-14	-9	-4	7									
K	-10	-2	-4	-8	-20	-6	-7	-10	-10	-9	-11	7								
M	-8	-7	-15	-17	-20	-7	-10	-12	-17	-3	-2	-4	12							
F	-12	-12	-12	-21	-19	-19	-20	-12	-9	-5	-5	-20	-7	9						
P	-4	-7	-9	-12	-11	-6	-9	-10	-7	-12	-10	-10	-11	-13	8					
S	-3	-6	-2	-7	-6	-8	-7	-4	-9	-10	-12	-7	-8	-9	-4	7				
T	-3	-10	-5	-8	-11	-9	-9	-10	-11	-5	-10	-6	-7	-12	-7	-2	8			
W	-2	-5	-11	-21	-22	-19	-23	-21	-10	-20	-9	-18	-19	-7	-20	-8	-19	13		
Y	-11	-14	-7	-17	-7	-18	-11	-20	-6	-9	-10	-12	-17	-1	-20	-10	-9	-8	10	
V	-5	-11	-12	-11	-9	-10	-10	-9	-9	-1	-5	-13	-4	-12	-9	-10	-6	-22	-10	8

Log-odds matrix for PAM10. Low PAM values such as this are useful for aligning very closely related sequences. Compare this with the PAM250 matrix (Fig. In previous slide) and note that there are larger positive scores for identical matches in this PAM10 matrix and larger penalties for mismatches.

Practical Usefulness of PAM Matrices in Pairwise Alignment

•Demonstration of Usefulness:

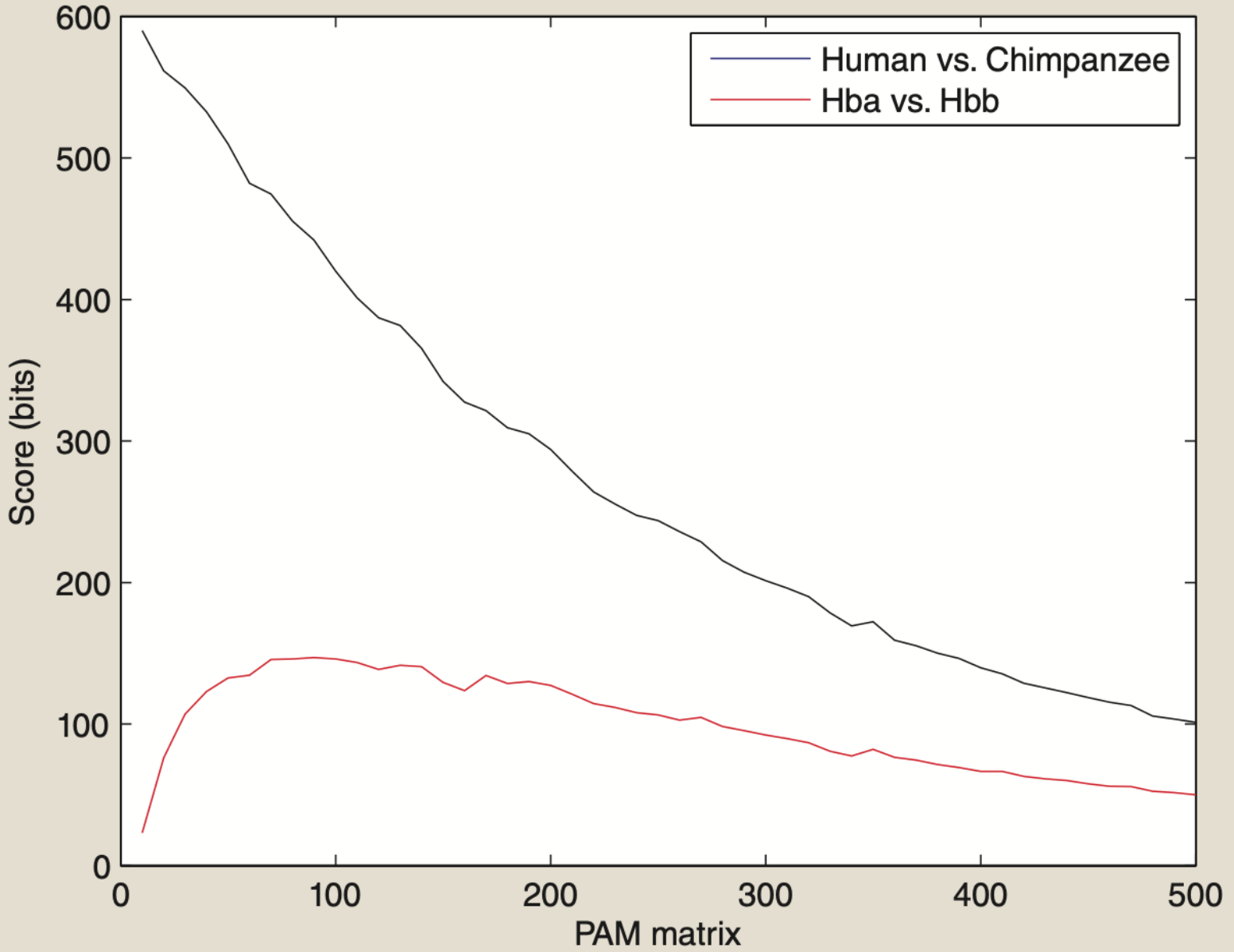
- Perform global pairwise alignments of both closely related and distantly related proteins.

•Example 1: Closely Related Proteins

- Human beta globin (NP_000509.1) vs. chimpanzee beta globin (XP_508242.1) with 100% amino acid identity.
- Bit scores decrease linearly: ~590 bits (PAM10) to ~200 bits (PAM250) to ~100 bits (PAM500).
- No mismatches or gaps.
- Higher relative entropy justifies high bit scores with low PAM matrices (e.g., PAM10).

•Example 2: Relatively Divergent Proteins

- Human beta globin vs. alpha globin (NP_000549.1).
- PAM70 matrix yields highest score.
- Lower PAM matrices (e.g., PAM10 to PAM60) produce lower bit scores due to 42% amino acid identity and assigned negative scores for mismatches.



Conclusion:

•Sensitivity of scoring matrices varies with sequence relatedness.

•Different matrices may be needed for different pairs of sequences.

Accepted Point Mutation: updated

- **PAM:** represents a **substitution** of one amino acid by another, **accepted** by **natural selection**.
- **Global alignment**
- **PAM1**: probability of amino acid substitutions equivalent to one mutation per 100 amino acids.
- The PAM1 matrix is based upon the alignment of **closely related** protein sequences, having an average of **1% change**.
- **PAM250** is a substitution matrix derived from observed mutations over a **longer evolutionary interval**, roughly equivalent to **250 mutations per 100 amino acids**.
- **Interpretation**: **Higher scores** in the matrix indicate **more probable substitutions**, reflecting the cumulative effect of evolutionary changes over a greater timescale.

Important Alternative to PAM: BLOSUM Scoring Matrices

Introduction to BLOSUM

- Blocks substitution matrix (BLOSUM) is widely used alongside PAM matrices.
- Developed by Henikoff (1992) using the **BLOCKS database**.
- BLOCKS contained over 500 groups of **local** multiple alignments (blocks) of **distantly related** protein sequences.
- Focus on **conserved regions (blocks)** of proteins that are distantly related.

- **BLOSUM Scoring Scheme:**
  - Employs a log-odds ratio using the base 2 logarithm:
    - $S = 2 \times \log \left(\frac{q_{ij}}{p_i p_j} \right)$ (Equation 3.6)
  - Similar to Equation 3.4 format used in PAM matrices.
- **General Log-Odds Formulation:**
  - Karlin and Altschul (1990) and Altschul (1991) demonstrated that substitution matrices can be described in a log-odds form:
    - $S = \ln \left(\frac{q_{ij}}{\lambda p_i p_j} \right)$ (Equation 3.7)
  - S_{ij}: Score of amino acid i aligning with j.
  - q_{ij}: Positive target frequencies summing to 1.
  - λ: Positive parameter providing a scale for the matrix.
  - λ encountered again when describing statistical measures of BLAST results.

BLOSUM62 Matrix and its Utility

- **Default Scoring Matrix in BLAST Programs:**
  - **BLOSUM62** matrix is the default scoring matrix in BLAST protein search programs at NCBI.
  - It merges all proteins in an alignment with 62% or greater amino acid identity into one sequence.
  - Example: In an aligned block of globin orthologs with 62%, 80%, and 95% identity, they are grouped as one sequence.
  - Substitution frequencies are weighted more by blocks with less than 62% identity.
  - Useful for scoring proteins with less than 62% identity.

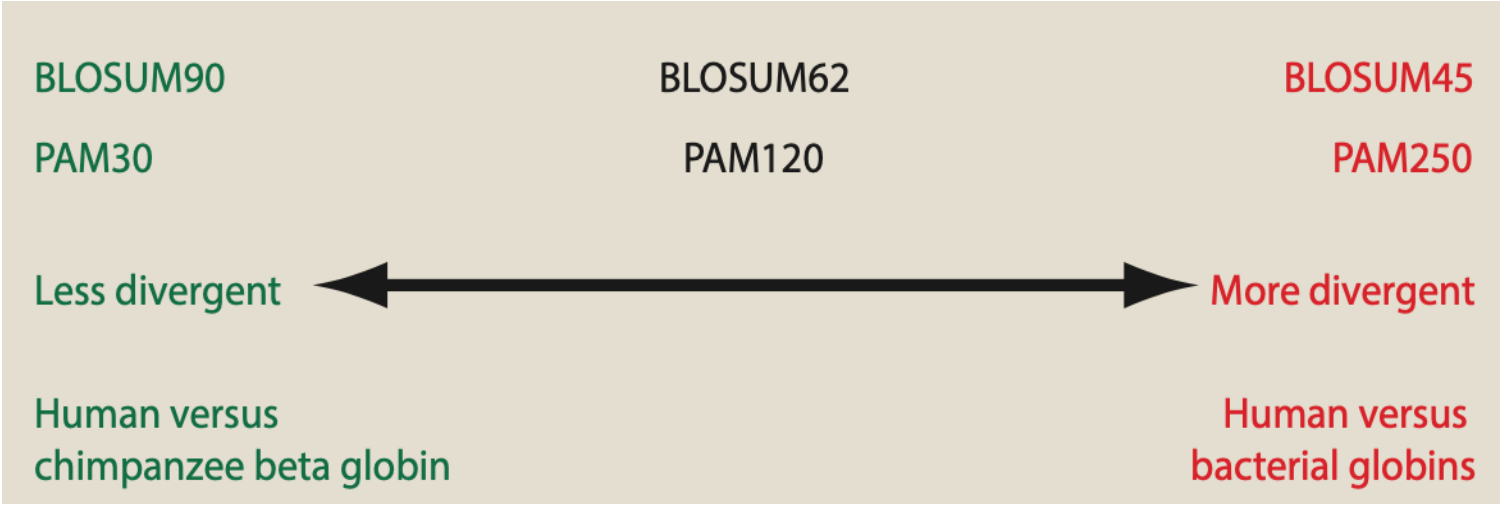
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4																			
R	-1	5																		
N	-2	0	6																	
D	-2	-2	1	6																
C	0	-3	-3	-3	9															
Q	-1	1	0	0	-3	5														
E	-1	0	0	2	-4	2	5													
G	0	-2	0	-1	-3	-2	-2	6												
H	-2	0	1	-1	-3	0	0	-2	8											
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
K	-1	2	0	-1	-1	1	1	-2	-1	-3	-2	5								
M	-1	-2	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

BLOSUM and PAM Performance

- **Performance Evaluation:**
  - Henikoff (1992) conducted tests to assess BLOSUM and PAM matrices in BLAST searches.
  - Results showed **BLOSUM62** outperformed BLOSUM60, BLOSUM70, and PAM matrices significantly in **identifying various proteins**.
  - BLOSUM matrices proved particularly effective for detecting **weakly scoring alignments**.
- **Commonly Used BLOSUM Matrices:**
  - Besides BLOSUM62, **BLOSUM50 and BLOSUM90** are commonly employed in BLAST searches.
  - BLOSUM50 is recommended for alignments with around 50% identity, with FASTA using it as a default.

BLOSUM and PAM Relationships

- **Relationships between PAM and BLOSUM:**
  - **Matrix selection** in pairwise sequence alignment or database searches depends on the suspected **degree of identity** between query and matches.
- **Basis of PAM and BLOSUM:**
  - PAM matrices rely on data from alignments of **closely related protein** families, assuming substitution probabilities for highly related proteins can be extrapolated to distantly related ones.
  - BLOSUM matrices are derived from empirical observations of **distantly related protein** alignments.



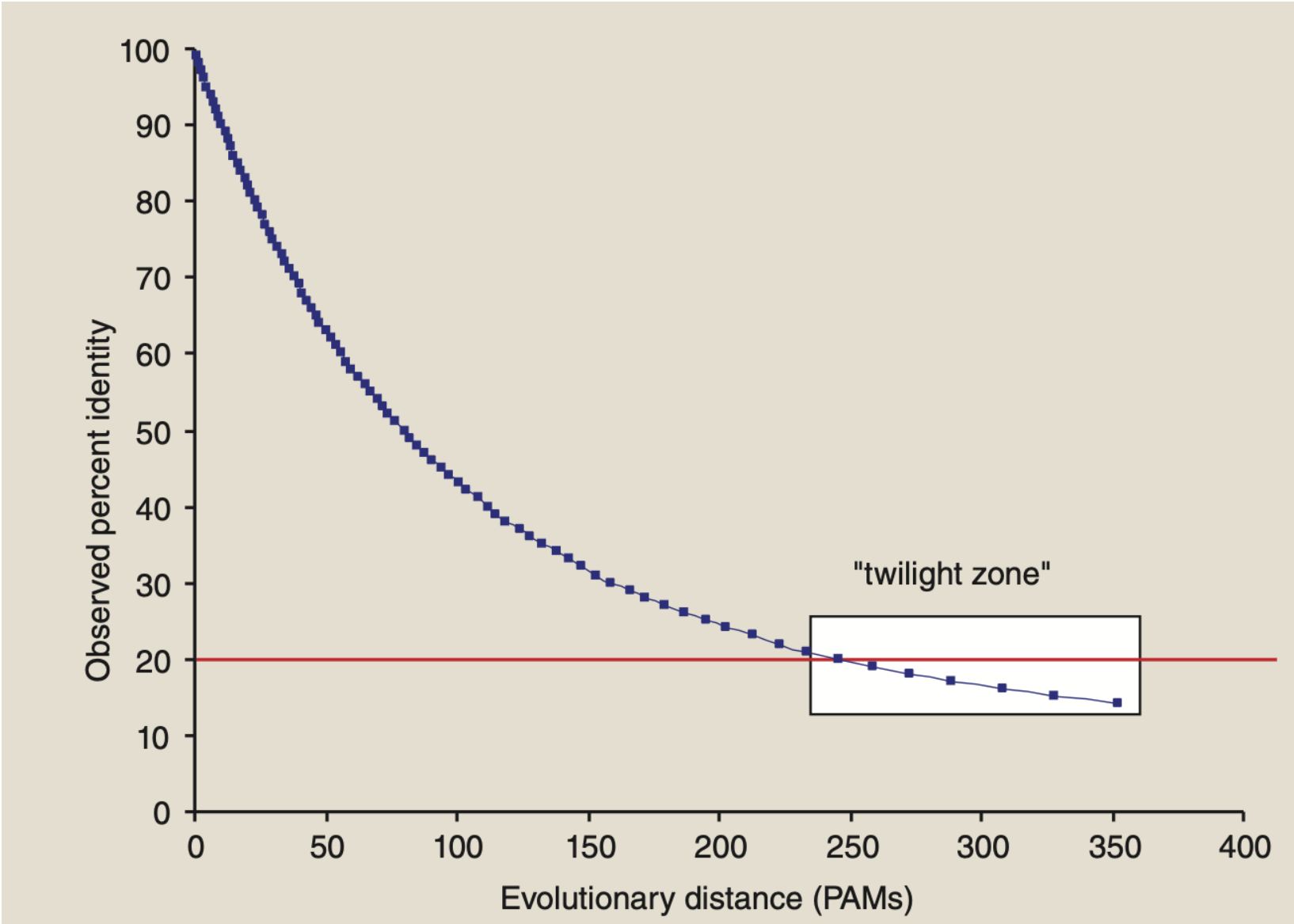
High-value BLOSUM matrices and low- value PAM matrices are best suited to study well-conserved proteins such as mouse and rat beta globin. BLOSUM matrices with low numbers (e.g., BLOSUM45) or high PAM numbers are best suited to detect distantly related proteins. Remember that in a BLOSUM45 matrix all members of a protein family with greater than 45% amino acid identity are grouped together, allowing the matrix to focus on proteins with less than 45% identity.

Recommended substitution matrices and gap costs based on query lengths:

Query Length	Substitution Matrix	Gap Costs
<35	PAM-30	(9,1)
35-50	PAM-70	(10,1)
50-85	BLOSUM-80	(10,1)
85	BLOSUM-62	(10,1)

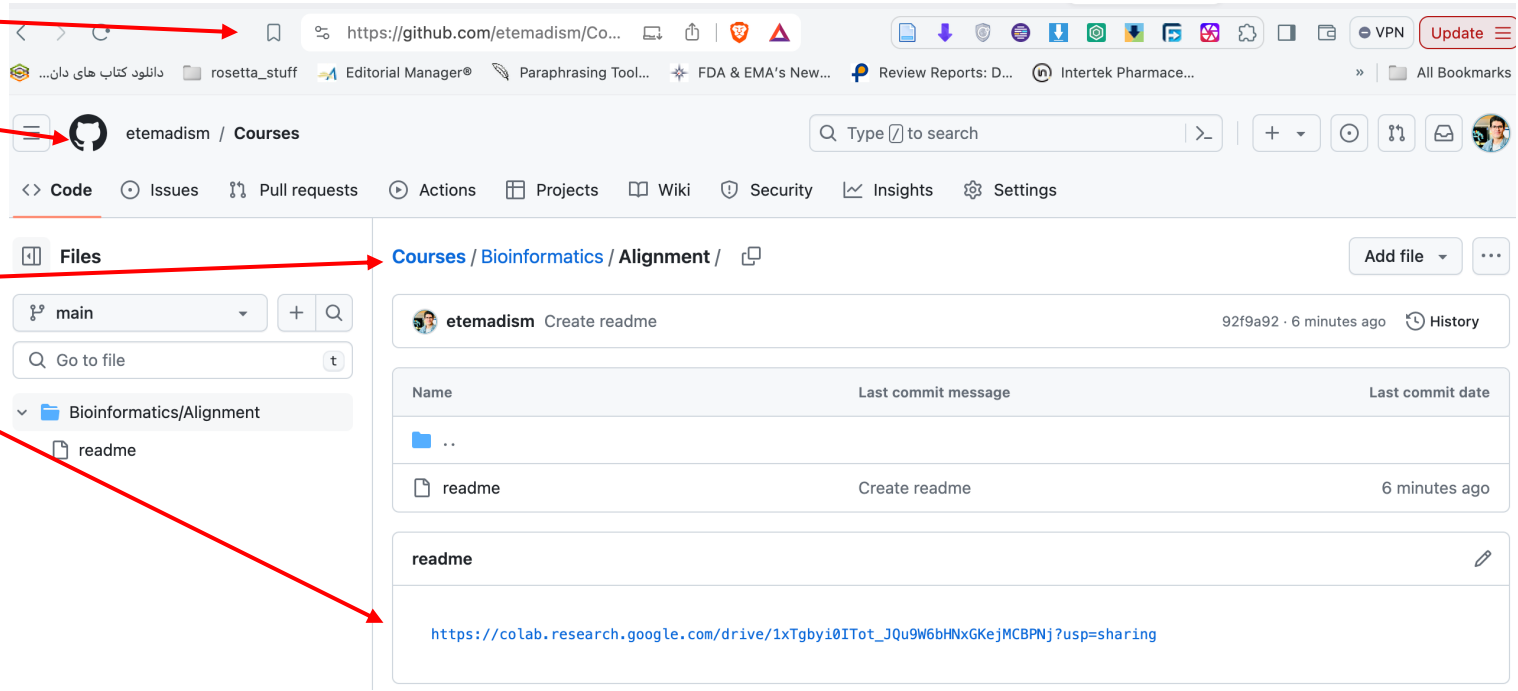
Twilight zone

Two randomly diverging protein sequences change in a negatively exponential fashion. This plot shows the observed number of amino acid identities per 100 residues of two sequences (y axis) versus the number of changes that must have occurred (the evolutionary distance in PAM units). The twilight zone (Doolittle, 1987) refers to the evolutionary distance corresponding to about **20% identity between** two proteins. Proteins with this degree of amino acid sequence identity **may be homologous**, **but such homology is difficult to detect**.



Run pairwise alignment in command line mode

- Go to github.
- Search for Username Etemadism
- Go to Folder Bioinformatics > Alignment
- Open readme link. It will direct you to google colab.



Types of Alignment Algorithms

Global Alignment:

Creates end-to-end alignments of sequences, considering the entire length.

Useful for comparing sequences with similar lengths.

Local Alignment:

Identifies the most similar regions within sequences, allowing for mismatches and gaps.

Suitable for finding conserved domains or motifs within larger sequences.

Pairwise Alignment:

Compares two sequences to identify regions of similarity.

Helps in understanding the relationship between individual sequences.

Multiple Alignment:

Aligns three or more sequences of similar length simultaneously.

Provides insights into evolutionary relationships and conserved regions among multiple sequences.

ALIGNMENT ALGORITHMS: GLOBAL AND LOCAL

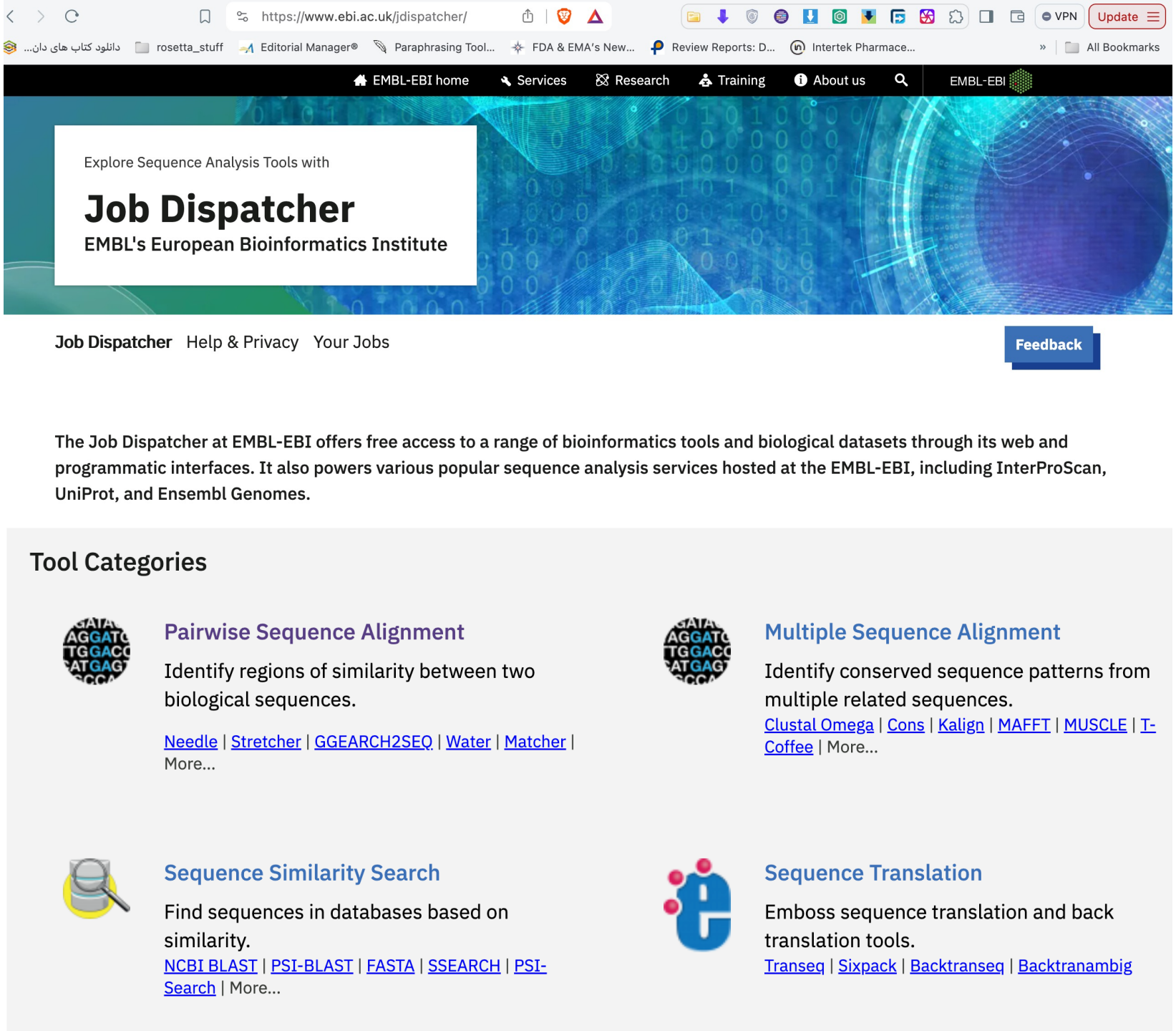
- **Global Alignment:**
  - **Needleman and Wunsch** (1970) introduced the concept of global alignment.
  - Global alignment encompasses the entire sequence of each protein or DNA molecule.
- **Local Alignment:**
  - **Smith and Waterman** (1981) developed the local alignment approach.
  - Local alignment focuses on identifying **regions of greatest similarity** between two sequences.
- **Database Search Algorithms:**
  - Most database search algorithms, including **BLAST**, utilize **local** alignments.

Global Sequence Alignment: Algorithm of Needleman and Wunsch

- **Introduction:**
- Needleman and Wunsch (1970) developed one of the earliest and most important algorithms for aligning two protein sequences.
- This algorithm facilitates optimal alignment, considering gaps, and is computationally efficient.
- Step 1: Setting Up a Matrix
- Step 2: Scoring the Matrix
- Step 3: Identifying the Optimal Alignment

Go to jdispatcher in EBI:

Choose PSA category



1- Global

- EMBOSS Needle
- EMBOSS Stretcher
- GGEARCH2SEQ

2- Local

- EMBOSS Water
- EMBOSS Matcher
- LALIGN
- SSEARCH2SEQ

3- Genomic

- GeneWise

Feedback

Pairwise Sequence Alignment

Pairwise Sequence Alignment is used to identify regions of similarity that may indicate functional, structural and/or evolutionary relationships between two biological sequences (protein or nucleic acid). By contrast, Multiple Sequence Alignment (MSA) is the alignment of three or more biological sequences of similar length. From the output of MSA applications, homology can be inferred and the evolutionary relationship between the sequences studied.

Global Alignment

Global alignment tools create an end-to-end alignment of the sequences to be aligned.

EMBOSS Needle

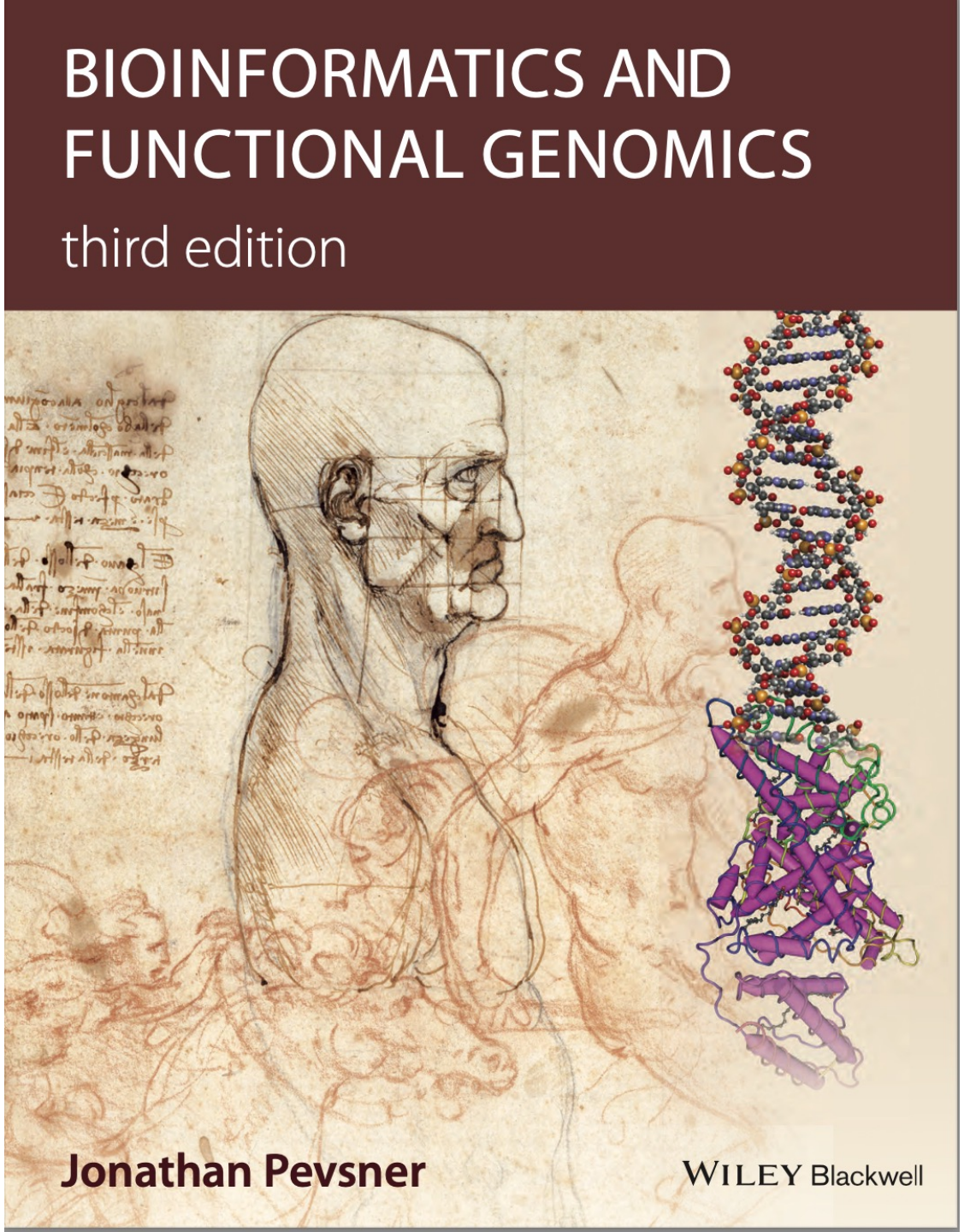
EMBOSS Needle creates an optimal global alignment of two sequences using the Needleman-Wunsch algorithm.

```
Launch EMBOSS Needle
```

EMBOSS Stretcher

EMBOSS Stretcher uses a modification of the Needleman-Wunsch algorithm that allows larger sequences to be globally aligned.

Reference


BIOINFORMATICS AND
FUNCTIONAL GENOMICS
third edition
Jonathan Pevsner
WILEY Blackwell