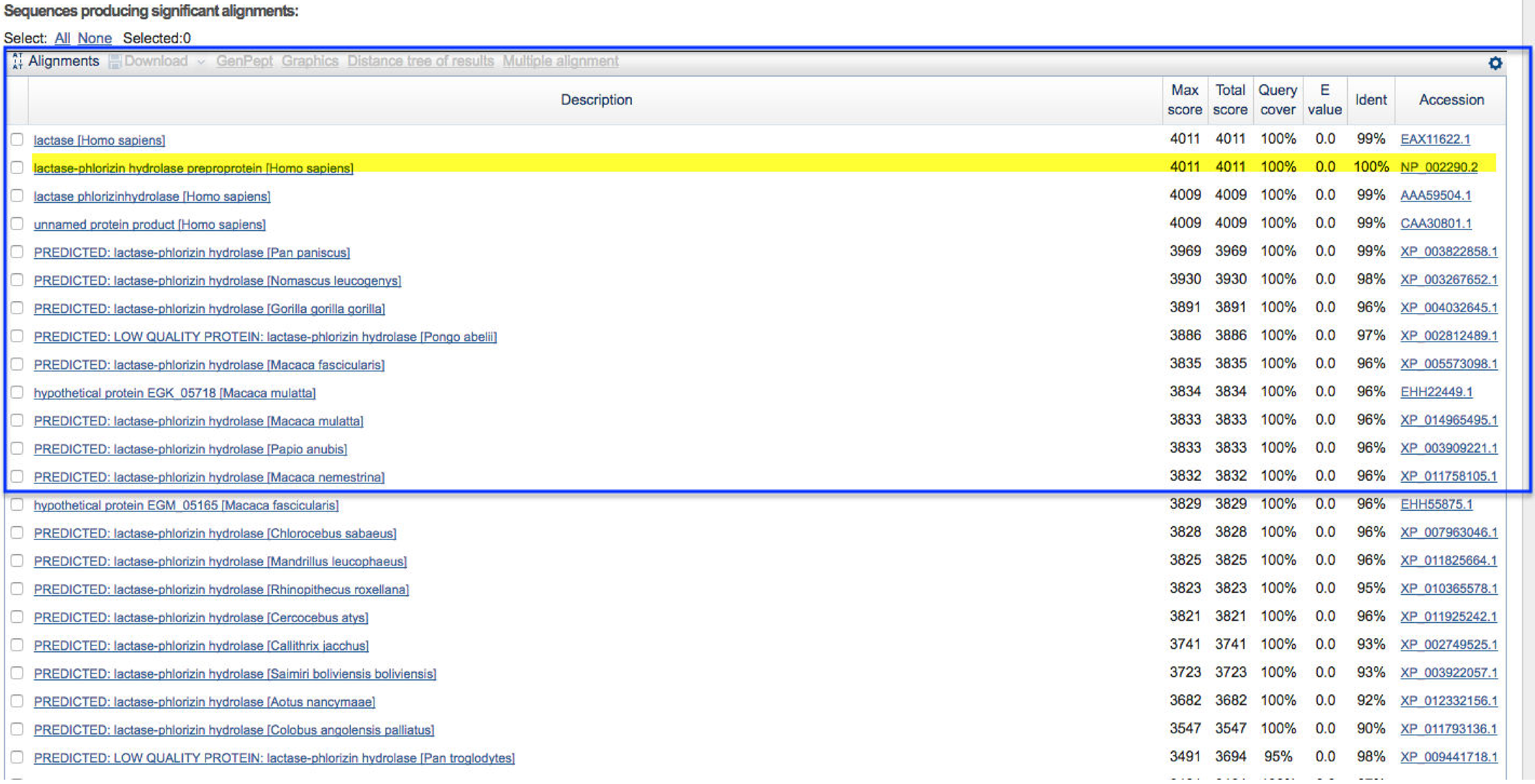
Statistical Significance of Sequence Alignment

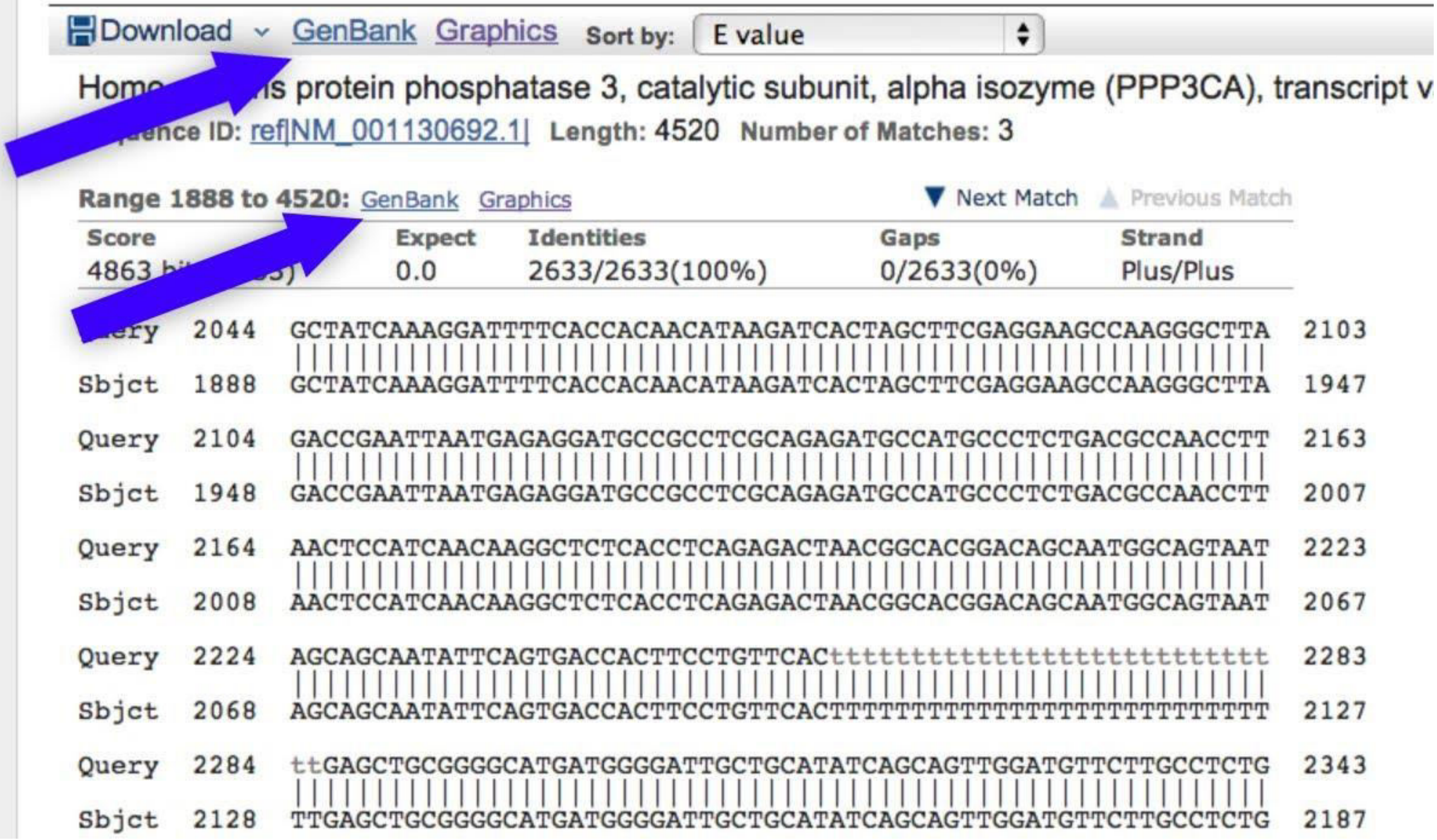
Course code: MIM 5531
Course Title: Bioinformatics and Biostatistics
Course Teacher: Jennifer Michellin Kiruba N



Sequences producing significant alignments:

Select: All None Selected:0

	Description	Max score	Total score	Query cover	E value	Ident	Accession
☐	lactase [Homo sapiens]	4011	4011	100%	0.0	99%	EAX11622.1
☐	lactase-phlorizin hydrolase preproprotein [Homo sapiens]	4011	4011	100%	0.0	100%	NP_002290.2
☐	lactase phlorizinhydrolase [Homo sapiens]	4009	4009	100%	0.0	99%	AAA59504.1
☐	unnamed protein product [Homo sapiens]	4009	4009	100%	0.0	99%	CAA30801.1
☐	PREDICTED: lactase-phlorizin hydrolase [Pan paniscus]	3969	3969	100%	0.0	99%	XP_003822858.1
☐	PREDICTED: lactase-phlorizin hydrolase [Nomascus leucogenys]	3930	3930	100%	0.0	98%	XP_003267652.1
☐	PREDICTED: lactase-phlorizin hydrolase [Gorilla gorilla gorilla]	3891	3891	100%	0.0	96%	XP_004032645.1
☐	PREDICTED: LOW QUALITY PROTEIN: lactase-phlorizin hydrolase [Pongo abelii]	3886	3886	100%	0.0	97%	XP_002812489.1
☐	PREDICTED: lactase-phlorizin hydrolase [Macaca fascicularis]	3835	3835	100%	0.0	96%	XP_005573098.1
☐	hypothetical protein EGK_05718 [Macaca mulatta]	3834	3834	100%	0.0	96%	EHH22449.1
☐	PREDICTED: lactase-phlorizin hydrolase [Macaca mulatta]	3833	3833	100%	0.0	96%	XP_014965495.1
☐	PREDICTED: lactase-phlorizin hydrolase [Papio anubis]	3833	3833	100%	0.0	96%	XP_003909221.1
☐	PREDICTED: lactase-phlorizin hydrolase [Macaca nemestrina]	3832	3832	100%	0.0	96%	XP_011758105.1
☐	hypothetical protein EGM_05165 [Macaca fascicularis]	3829	3829	100%	0.0	96%	EHH55875.1
☐	PREDICTED: lactase-phlorizin hydrolase [Chlorocebus sabaeus]	3828	3828	100%	0.0	96%	XP_007963046.1
☐	PREDICTED: lactase-phlorizin hydrolase [Mandrillus leucophaeus]	3825	3825	100%	0.0	96%	XP_011825664.1
☐	PREDICTED: lactase-phlorizin hydrolase [Rhinopithecus roxellana]	3823	3823	100%	0.0	95%	XP_010365578.1
☐	PREDICTED: lactase-phlorizin hydrolase [Cercocebus atys]	3821	3821	100%	0.0	96%	XP_011925242.1
☐	PREDICTED: lactase-phlorizin hydrolase [Callithrix jacchus]	3741	3741	100%	0.0	93%	XP_002749525.1
☐	PREDICTED: lactase-phlorizin hydrolase [Saimiri boliviensis boliviensis]	3723	3723	100%	0.0	93%	XP_003922057.1
☐	PREDICTED: lactase-phlorizin hydrolase [Aotus nancymaae]	3682	3682	100%	0.0	92%	XP_012332156.1
☐	PREDICTED: lactase-phlorizin hydrolase [Colobus angolensis palliatus]	3547	3547	100%	0.0	90%	XP_011793136.1
☐	PREDICTED: LOW QUALITY PROTEIN: lactase-phlorizin hydrolase [Pan troglodytes]	3491	3694	95%	0.0	98%	XP_009441718.1



Home… is protein phosphatase 3, catalytic subunit, alpha isozyme (PPP3CA), transcript v

…ence ID: ref|NM_001130692.1| Length: 4520 Number of Matches: 3

Range 1888 to 4520: GenBank Graphics

Score	Expect	Identities	Gaps	Strand
4863 bi… …3)	0.0	2633/2633(100%)	0/2633(0%)	Plus/Plus

```
…ery   2044  GCTATCAAAGGATTTTCACCACAACATAAGATCACTAGCTTCGAGGAAGCCAAGGGCTTA  2103
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1888  GCTATCAAAGGATTTTCACCACAACATAAGATCACTAGCTTCGAGGAAGCCAAGGGCTTA  1947

Query  2104  GACCGAATTAATGAGAGGATGCCGCCTCGCAGAGATGCCATGCCCTCTGACGCCAACCTT  2163
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1948  GACCGAATTAATGAGAGGATGCCGCCTCGCAGAGATGCCATGCCCTCTGACGCCAACCTT  2007

Query  2164  AACTCCATCAACAAGGCTCTCACCTCAGAGACTAACGGCACGGACAGCAATGGCAGTAAT  2223
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2008  AACTCCATCAACAAGGCTCTCACCTCAGAGACTAACGGCACGGACAGCAATGGCAGTAAT  2067

Query  2224  AGCAGCAATATTCAGTGACCACTTCCTGTTCACtttttttttttttttttttttttttt  2283
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2068  AGCAGCAATATTCAGTGACCACTTCCTGTTCACTTTTTTTTTTTTTTTTTTTTTTTTTTT  2127

Query  2284  ttGAGCTGCGGGGCATGATGGGATTGCTGCATATCAGCAGTTGGATGTTCTTGCCTCTG  2343
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2128  TTGAGCTGCGGGGCATGATGGGATTGCTGCATATCAGCAGTTGGATGTTCTTGCCTCTG  2187
```

- The calculation of alignment scores involves addition of the match/mismatch values from the matrix for every nucleotide base or amino acid residue involved in the alignment to obtain a gross alignment score.
- Then the total gap penalty is calculated.
- The total gap penalty value is subtracted from the gross alignment score value to obtain the final alignment score.

- The statistical significance of the raw score, S, of an alignment is assessed to determine whether the observed alignment is specific or could be the result of random chance.
- This is done by creating many random sequences of the same length from one of the two aligned sequences by shuffling the sequence and running the alignment again.
- Typically this reshuffling and realignment process is repeated 200 times or more.
- Each alignment using these random sequences produces an alignment score (s).
- These scores (s1. . .sn) are plotted to generate a distribution pattern, a threshold of significance is set, and the original score (S) is compared against this distribution.
- If the S is located at one end of the distribution (extreme value distribution) that means that the alignment is not likely to be produced by random chance.

Score, Bit-score, P-value, E-value

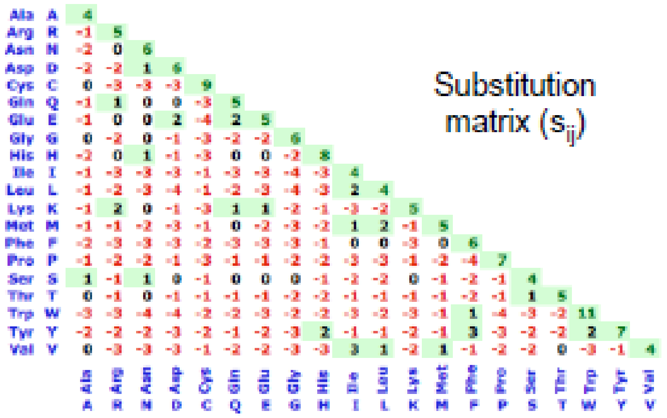
Score: A number used to assess the biological relevance of a finding.

In the context of sequence alignments, a score is a numerical value that describes the overall quality of an alignment. Higher numbers correspond to higher similarity. The score scale depends on the scoring system used (substitution matrix, gap penalty).

$$S = \sum_{i=1}^{L} s_{r_{1,i} r_{2,i}}$$

Example:

```
R   L   A   S   V   -   E   T   D   M   W   T   P   L   T   L   R   Q   H
.   |   .   |   :       :   |   .   :           .   |   .   .   |
T   L   T   S   L   A   Q   T   T   L   -   -   K   A   H   L   G   T   H
-1  +4  +0  +4  +1  -4  +2  +5  -1  +2  -4  -1  -1  -1  -2  +4  -2  -1  +8  = 12
```



Ala	A	4																			
Arg	R	-1	5																		
Asn	N	-2	0	6																	
Asp	D	-2	-2	1	6																
Cys	C	0	-3	-3	-3	9															
Gln	Q	-1	1	0	0	-3	5														
Glu	E	-1	0	0	2	-4	2	5													
Gly	G	0	-2	0	-1	-3	-2	-2	6												
His	H	-2	0	1	-1	-3	0	0	-2	8											
Ile	I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4
		Ala A	Arg R	Asn N	Asp D	Cys C	Gln Q	Glu E	Gly G	His H	Ile I	Leu L	Lys K	Met M	Phe F	Pro P	Ser S	Thr T	Trp W	Tyr Y	Val V

Substitution matrix (s_{ij})

gap penalty (s_{i-})

gap opening	-4
gap extension	-1
end gap	0

Z-Score

- In the statistical sense, Z is the distance between S and the mean of scores obtained using randomized sequences.
- The Z-score is calculated by repeating the reshuffling and realignment process, as described above, and noting the raw score (s) of each alignment using the randomized sequences ($s_1 . . . s_n$).
- The mean (x) and the standard deviation (σ) of $s_1 . . . s_n$ are calculated and from these the Z-score of the target alignment can be determined.
- The calculation of the Z-score assumes that the alignment of the shuffled random sequences shows a normal distribution.

interpretation of the Z-score is as follows:

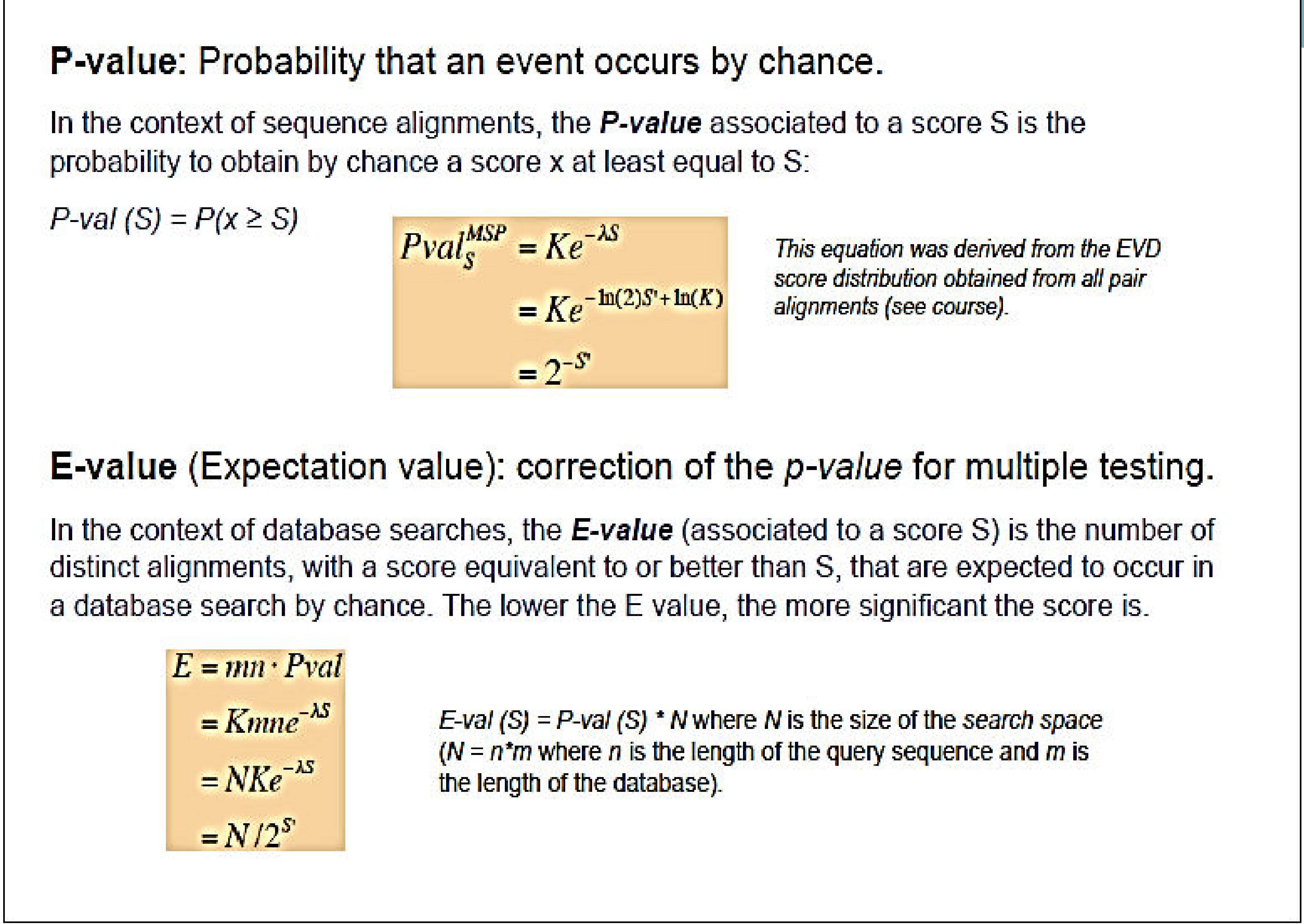
- Z>20: two sequences are definitely homologous (Family)
- Z between 10 and 20: two sequences most likely homologous (Family/Superfamily)
- Z between 6 and 8: two sequences are less likely to be homologous
- Z<6: not significant.

P-Value

- The P-value of an alignment represents the probability of obtaining a score$\geq S$ by chance.
- For example, if the P-value is 10^5, it means that the probability of obtaining an alignment with a score$\geq S$ is 1 out of 10^5.
- Thus, different alignments can be compared based on their P-values.
- The P-value ranges from 0 to 1; the closer it is to 0, the better is the alignment.

E-Value

- The E-value is the expectation value that indicates the number of alignments with a score≥S that one can expect to find by chance in a database of size N.
- Hence, the E-value is dependent on the database size and the query length.
- The closer the E-value to 0, the better is the alignment.
- The E-value is the most widely used measure for estimating the quality of sequence alignment—that is, the extent of sequence similarity.
- The typical threshold for the E-value when judging homology, particularly using BLAST, is E≤1e-5, and the lower the value, the better it is.
- lowering the default value makes the search more stringent and fewer chance matches are reported.



P-value: Probability that an event occurs by chance.

In the context of sequence alignments, the ***P-value*** associated to a score S is the probability to obtain by chance a score x at least equal to S:

P-val (S) = P(x ≥ S)

$$
\begin{aligned}
Pval_S^{MSP} &= Ke^{-\lambda S} \\
&= Ke^{-\ln(2)S' + \ln(K)} \\
&= 2^{-S'}
\end{aligned}
$$

This equation was derived from the EVD score distribution obtained from all pair alignments (see course).

E-value (Expectation value): correction of the *p-value* for multiple testing.

In the context of database searches, the ***E-value*** (associated to a score S) is the number of distinct alignments, with a score 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 is.

$$
\begin{aligned}
E &= mn \cdot Pval \\
&= Kmne^{-\lambda S} \\
&= NKe^{-\lambda S} \\
&= N/2^{S'}
\end{aligned}
$$

*E-val (S) = P-val (S) * N* where *N* is the size of the *search space* (*N = n*m* where *n* is the length of the query sequence and *m* is the length of the database).

Bit Score

- The bit score (S_0) is a normalized raw score expressed in bits; it is an estimate of the search space one has to search through—that is, the number of sequence pairs one has to score—before one can come across a raw alignment score≥S, by chance.
- It should be emphasized that the bit score is dependent on sequence length, and short sequences may not produce high bit scores despite very high identity.
- To summarize the utility of the statistical estimates of sequence alignment in simple terms, the better the alignment (e.g. homologous sequences),
- the lower the P- and E-values,
- and the higher the Z- and bit scores.

Bit-score: A log-scaled version of a score.

In the context of sequence alignments (BLAST), the **bit-score S'** is a normalized score expressed in *bits* that lets you estimate the magnitude of the *search space* you would have to look through before you would expect to find an score as good as or better than this one by chance. Althshul proposes to following definition:

$$S' = \frac{\lambda S - \ln(K)}{\ln(2)}$$

S is the raw score. Parameters λ and K depend on the substitution matrix and on the gap penalties (Altchul).

Ex: If the bit-score is 30, you would have to score, on average, about 2^{30} = 1 billion independent segment pairs to find a score this score by chance. Each additional bit doubles the size of the search space.

The bit-scores is thus a rescaled version of the raw alignment score that is *independent of the size of the search space*.

The **size of the search space** is proportional to the product of the query sequence length (*n*) * the sum of the lengths of the sequences in the database (*m*): ***N=n*m***. The size of the search space is then obtained by multiplying N by a coefficient *K* (Altschul).

Ex: When searching protein databases with protein queries, *K* is about 0.13. Thus, for a protein of length n=235 aa which is searched against a database of size m=12 496 420 aa, the size of the search space is equal to 0.13 * 235 * 12 496 420 = about 0.38 billion. In this case, a bit score of 30 (which corresponds to a space of 2^{30} = 1 billion) may have occurred by chance alone.

- BLAST E-Value Cut-Off
- For nucleic-acid-based search, the suggested threshold (minimum significant hit) for the E-value is ≤1e-6 and a sequence identity of ≥70%.
- For protein-based search, the suggested threshold for the E-value ≤ 1e-4, with a sequence identity of ≥ 35%g. However, typically for protein-based homology search, the threshold used is E ≤ 1e-5, and the lower it is, the better. For example, an E-value of 1e-25 will indicate a clear homology.

- To summarize the utility of the statistical estimates of sequence alignment in simple terms, the better the alignment (e.g. homologous sequences),
- the lower the P- and E-values,
- and the higher the Z- and bit scores.