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

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).

Gap penalty	Alignment	Identity / Similarity	Gaps	Score
0	1 GTC-ATGCTA-GTCGT---GG---GTAGCATTTA-GCT-ATG-TGGG-GT 38 1 -TCGATGCT-GGTCG-CAAGGCAAGTAG---TTATG-TCATGCT---AG- 39	27/50 (54.0%)	23/50	S=135
5	1 GTC-ATGCTAGTCG--TGGGTAGCATTTA-GCT-ATG-TGGGGT 38 1 -TCGATGCTGGTCGCAAGGCAAGTAGTTATG-TCATGCTAG--- 39	26/44 (59.1%)	11/44	S=67
10	1 ----------GTCATGCTAGTCGTGGGTAGC 21 1 TCGATGCTGGTCGCAAGGCAAGTAGTTATGTCATGCTAG---------- 39 22 ATTTAGCTATGTGGGGT 38 39 ---------- 39	10/67 (14.9%)	57/67	S=50

Observations: If the gap penalty is too large, gaps are avoided and the sequences can not be properly aligned. If the gap penalty is too low, gaps are inserted everywhere to prevent mismatches. This does not produce any informative alignement. The "best" alignment is obtained for an intermediary gap penalty.

Remark: The scores of these different alignments can not be compared (neither used to select the best alignment) because their scale depends on the gap penalty.

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

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.

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

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)

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

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.

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

*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).

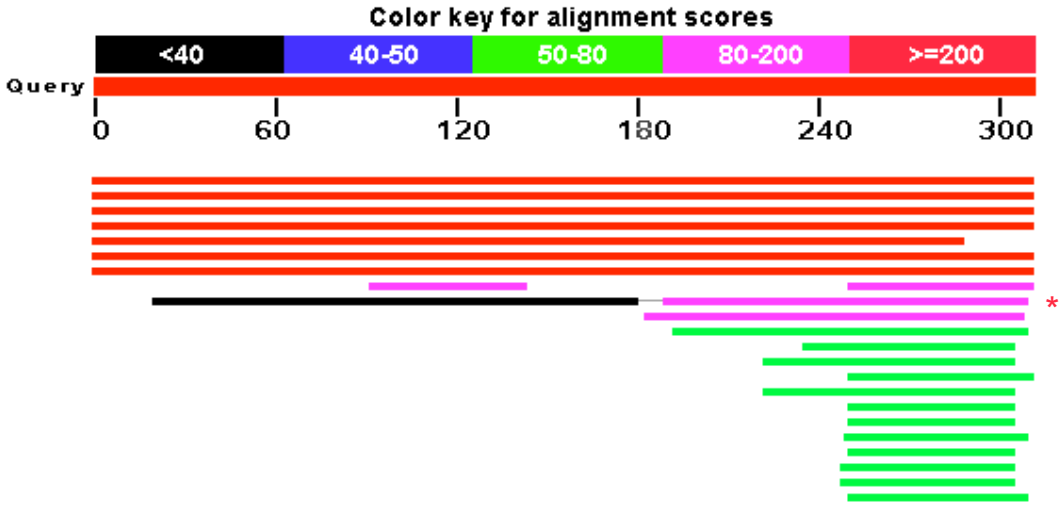
Score, Bit-score, P-value, E-value: example

Example: BLAST - Pho4p (*S. cerevisiae*)

```
>gi|259146228|emb|CAY79487.1| Pho4p [Saccharomyces cerevisiae EC1118]
MGRTTSEGIHGFVDDLEPKSSILDKVGDFITVNTKRHDGREDFNEQNDELNSQEHHNSSENGNENENEQD
SLALDDLDRAFELVEGMDMDWMMPSHAHHSPATTATIKPRLLYSPLIHTQSAVPVTISPNLVATATSTTS
ANKVTKNKSNSSPYLNKRRGKPGPDSATSLFELPDSVIPTPKPKPKPKQYPKVILPSNSTRRISPVTAKT
SSSAEGVVVASESPVIAPHGSSHSRSLSKRRSSGALVDDDKRESHKHAEQARRNRLAVALHELASLIPAE
WKQQNVSAAPSKATTVEAACRYIRHLQQNVST
```

Query (input) sequence
(Pho4p from *S. cerevisiae*)

BLAST (default parameters) ↓



Explanation of Output of a BLAST Search:
http://www.ncbi.nlm.nih.gov/staff/tao/URLAPI/new_view.html

Results (output) of BLAST

- The top segment displays the color key and the query based scale.
- The colored bars represent the actual HSPs. The position of each bar indicates the region of the query the HSP covers.
- The thin line (see *) indicates that the two HSPs are from the same sequence.
- Small vertical lines (not obtained here) indicate breaks, i.e., segments which are not connected in the actual alignment.

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

Example: BLAST - Pho4p (*S. cerevisiae*)

Results (output) of BLAST

Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E value
CAY79487.1	Pho4p [Saccharomyces cerevisiae EC1118]	640	640	100%	0.0
EDV09876.1	myc-family transcription factor [Saccharomyces cerevisiae RM1	637	637	100%	0.0
NP_116692.1	Basic helix-loop-helix (bHLH) transcription factor of the myc-fam	637	637	100%	0.0
EEU04180.1	Pho4p [Saccharomyces cerevisiae JAY291]	629	629	100%	2e-178
CAA27345.1	unnamed protein product [Saccharomyces cerevisiae]	584	584	92%	4e-165
CAA36809.1	unnamed protein product [Saccharomyces cerevisiae]	562	562	100%	2e-158
CAA36810.1	unnamed protein product [Saccharomyces cerevisiae]	561	561	100%	3e-158
1A0A_A	Chain A, Phosphate System Positive Regulatory Protein Pho4DN	126	126	19%	4e-27
EDZ72385.1	YFR034Cp-like protein [Saccharomyces cerevisiae AWRI1631]	102	102	16%	4e-20
XP_002553686.1	KLTH0E04664p [Lachancea thermotolerans] >emb|CAR23249.1|	84.0	120 *	90%	2e-14
NP_983973.2	ADL123Cp [Ashbya gossypii ATCC 10895] >gb|AAS51797.2| AD	83.6	83.6	40%	3e-14
XP_002489917.1	hypothetical protein [Pichia pastoris GS115] >emb|CAY67636.1|	72.4	72.4	37%	6e-11
XP_445634.1	unnamed protein product [Candida glabrata] >emb|CAG58545.1	68.6	68.6	22%	1e-09

Max score = highest alignment score (bit-score) between the query sequence and the database sequence segment .

Total score = sum of alignment scores of all segments from the same database sequence that match the quary sequence (calculated over all segments). This score is different from the max score if several parts of the database sequence match different parts of the query sequence (see " * " in the example).

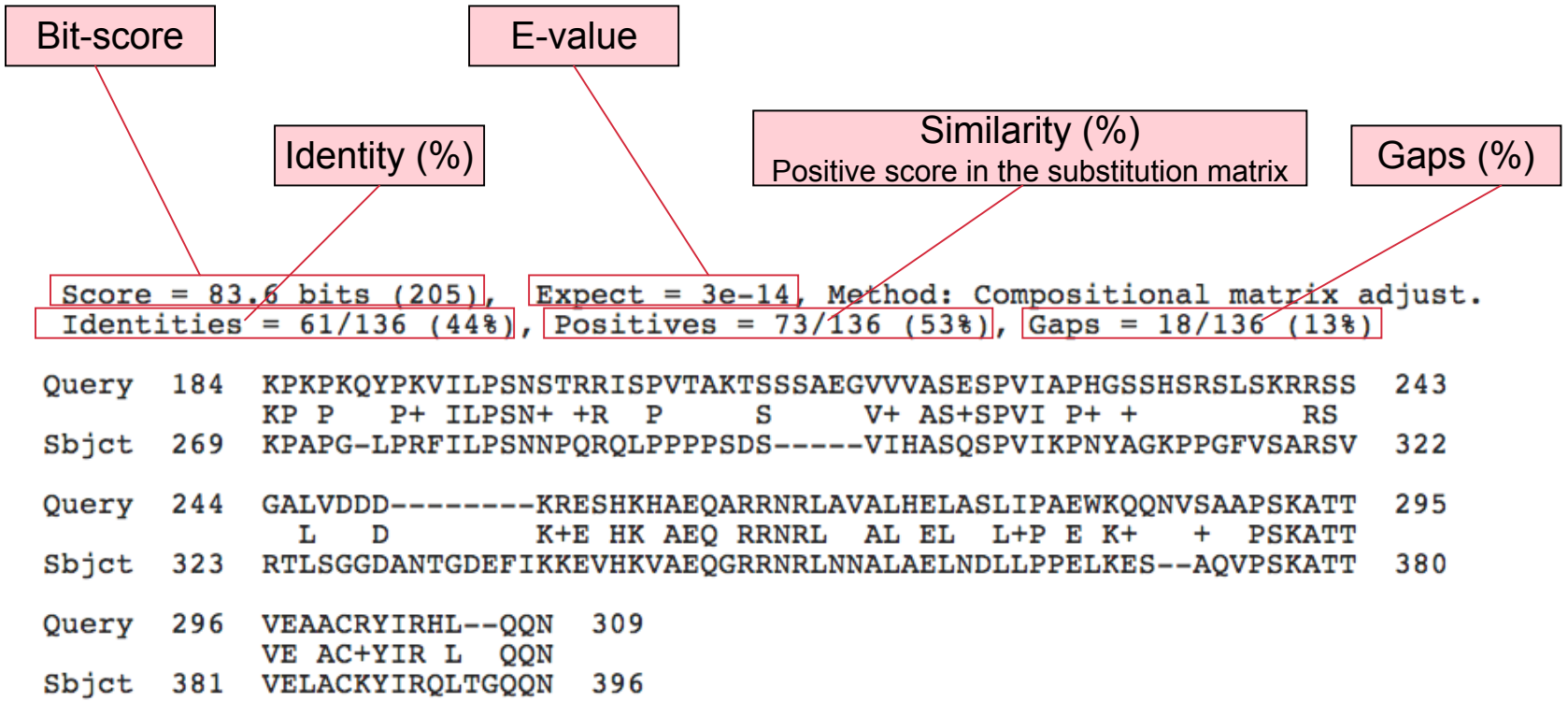
Query coverage = percent of the query length that is included in the aligned segments. This coverage is calculated over all segments (cf. total score).

E-value = number of alignments expected by chance with a particular score or better. The expect value is the default sorting metric and normally gives the same sorting order as Max score.

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

Example: BLAST - Pho4p (*S. cerevisiae*)

Results (output) of BLAST



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

Example: FASTA - PL6 human vs PL6 mouse

row score z-score bit-score E-value Identity Similarity Overlap

```
>>sp|Q9WUH1|TM115_MOUSE Transmembrane protein 115 OS=Mus musculus GN=Tmem115  (350 aa)
 s-w opt: 2163  Z-score: 1995.8  bits: 377.8  E(): 2.3e-105
Smith-Waterman score: 2163;  94.8% identity  (98.9% similar) in  348 aa overlap (1-348:1-348)
Entrez Lookup  Re-search database  General re-search
               10        20        30        40        50        60        70        80
sp|Q12 MQRALPGARQHLGAILASASVVVKALCAAVLFLYLLSFAVDTGCLAVTPGYLFPPNFWIWTLATHGLMEQHVWDVAISLT
       :::::::::::::::::::::::::::::::::::::.:::::::::::::::::::::::::::::::::::::::::::.
sp|Q9W MQRALPGARQHLGAILASASVVVKALCAVVLFLYLLSFAVDTGCLAVTPGYLFPPNFWIWTLATHGLMEQHVWDVAISLA
               10        20        30        40        50        60        70        80

               90       100       110       120       130       140       150       160
sp|Q12 TVVVAGRLLEPLWGALELLIFFSVVNVSVGLLGAFAYLLTYMASFNLVYLFTVRIHGALGFLGGVLVALKQTMGDCVVLR
       ::::::::::::::::::::::::::::::::::::::.:::::::::::::::::::.:::::::::::::::::::::::
sp|Q9W TVVVAGRLLEPLWGALELLIFFSVVNVSVGLLGALAYLLTYMASFNLVYLFTIRIHGALGFLGGVLVALKQTMGDCVVLR
               90       100       110       120       130       140       150       160

              170       180       190       200       210       220       230       240
sp|Q12 VPQVRVSVMPMLLLALLLLLRLATLLQSPALASYGFGLLSSWVYLRFYQRHSRGRGDMADHFAFATFFPEILQPVVGLLA
       ::::::::.:::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
sp|Q9W VPQVRVSVVPMLLLALLLLLRLATLLQSPALASYGFGLLSSWVYLRFYQRHSRGRGDMADHFAFATFFPEILQPVVGLLA
              170       180       190       200       210       220       230       240

              250       260       270       280       290       300       310       320
sp|Q12 NLVHSLLVKVKICQKTVKRYDVGAPSSITISLPGTDPQDAERRRQLALKALNERLKRVEDQSIWPSMDDDEEESGAKVDS
       ::::.:::::::::::::::::::::::::::::::::::::::::::::::::::: :::::::::.:::.::
sp|Q9W NLVHGLLVKVKICQKTVKRYDVGAPSSITISLPGTDPQDAERRRQLALKALNERLKRVEDQSAWPSMDDDEEEAGAKTDS
              250       260       270       280       290       300       310       320

              330       340       350
sp|Q12 PLPSDKAPTPPGKGAAPESSLITFEAAPPTL
       ::: ..: ::::: ..:::::::.:.::
sp|Q9W PLPLEEASTPPGKVTVPESSLITLETAPLL
              330       340       350
```

NB: The alignment statistics are here computed by shuffling the second sequence many times (here 200).

Z-score

$$Z - score = \frac{S - mean(S)}{st.\ dev.(S)}$$

The program FASTA was run at http://wrpmg5c.bioch.virginia.edu/fasta_www2/fasta_www.cgi?rm=shuffle