BLAST/FASTA flavors

The BLAST and FASTA package contain programs for different purposes:

BLAST program	**FASTA program**	**query type**	**database type**
`blastn`	`fasta`	nucleotide	nucleotide
`blastp`	`fasta`	protein	protein
`blastx`	`fastx`	nucleotide (translated in 6 frames)	protein
`tblastn`	`tfasta`	protein	nucleotide (translated in 6 frames)
`tblastx`		nucleotide (both translated	nucleotide in 6 frames)

Further FASTA programs: `fasty`, `tfasty` (translation with frameshifts), `ssearch` (Smith-Waterman).

BLAT - BLAST-Like Alignment Tool (Kent, 2002)

- To reduce search time BLAT requires only one single, but longer exact match (instead on several consecutive ones), to be candidate for the more rigorous search.
- This speeds up the candidate search, but makes the search less sensitive...
- that means it might miss (more) relevant hits.
- Nevertheless, BLAT works well if the database sequences and the query are closely related.
- However, this is the scenario BLAT was developed for (mapping reads against a closely related reference).

Criteria to compare search methods (I)

When doing a database search, the following can happen

- we find a sequence which is indeed related to the query (=true positive, TP)
- we discard a sequence which is indeed not related to the query (=true negative, TN)
- we discard a sequence which is actually related to the query (=false negative, FN)
- we find a sequence which is actually not related to the query (=false positive, FP)

Criteria to compare search methods (II)

Criteria:

- sensitivity $\frac{TP}{(TP+FN)}$: The proportion of those correctly found among all those which should have been found.
- specificity $\frac{TN}{(TN+FP)}$: The proportion of those correctly dicarded among all those which should have been discarded.
- positive predictive value $\frac{TP}{(TP+FP)}$: The proportion of those correctly found among all found. (a.k.a. precision)
- negative predictive value $\frac{TN}{(TN+FN)}$: The proportion of those correctly discarded among all discarded.

Criteria to compare search methods (III)

Confusion matrix

		Correct classification	
		related sequences	unrelated sequences
Search result	retrieved (found) sequences	**TP** (true positives)	**FP** (false positives, type I error)
	discarded sequences	**FN** (false negatives, type II error)	**TN** (true negatives)

$$Specificity = \frac{TN}{TN + FP} \qquad Sensitivity = \frac{TP}{TP + FN}$$

A comparison of BLAST, FASTA, Smith-Waterman (Shpaer et al., 1996)

Shpaer et al. (1996) did a comparison of BLAST, FASTA, and Smith-Waterman. They found that

- Smith-Waterman (SW) dynamic programming method and the optimized version of FASTA are significantly better able to distinguish true similarities from statistical noise than is the popular database search tool BLAST.
- FASTA performs worse than Smith-Waterman, but much better than BLAST.
- On the other hand, Smith-Waterman takes much longer.

- Despite its good performance, Smith-Waterman is by far the slowest.
- The reason that many people use a software like BLAST, does not make it better.
- Note: These points results do not reflect improvements done to the softwares since 1996.

Search Methods Based on Multiple Alignments

BLAST, FASTA, Smith-Waterman, BLAT etc. search with a single query sequence in a Database of sequences.

If we have multiple sequence alignments available (e.g. of a sequence family or of conserved regions), we can use it to extract information to search in databases.

Database searching: A Typical workflow

With a new sequence

- search SwissProt/UniProt, EMBL-ENA/GenBank/DDBJ databases for similar sequences.
- collect similar sequences from fast heuristic searches
- use more optimal methods to sort out false positives, if necessary
- use multiple alignments methods to produce an MSA
- maybe use the multiple alignments to train other tools (e.g. HMMs) to find more distantly related sequences.
- extend the MSA with newly found sequences
- do further Analyses with the data: e.g. phylogenetic analyses

Searching for Patterns with PSSM

Searching for specific pattern can be accomplished using position specific scoring matrices (PSSM).

... S I I C D N C N ...

	1	2	3	4	5	6
A	–3	0	–2	–1	0	–3
C	–4	1	11	–5	4	11
D	–4	–3	–6	6	–3	–6
E	.	.	.	.	.	.
F	.	.	.	.	.	.
G	.	.	.	.	.	.
H	.	.	.	.	.	.
I	2	2	–3	–5	1	–3
K	.	.	.	.	.	.
L	.	.	.	.	.	.
M	.	.	.	.	.	.
N	0	–3	–5	4	1	–5
Q	.	.	.	.	.	.
P	.	.	.	.	.	.
R	.	.	.	.	.	.
S	0	3	–3	–2	–1	–3
T	.	.	.	.	.	.
V	.	.	.	.	.	.
W	.	.	.	.	.	.
Y	6	–3	–4	–4	0	–4
	+0	+2	–3	–5	–3	–5 =–14

... S I I C D N C N ...

	1	2	3	4	5	6
A	–3	0	–2	–1	0	–3
C	–4	1	11	–5	4	11
D	–4	–3	–6	6	–3	–6
E	.	.	.	.	.	.
F	.	.	.	.	.	.
G	.	.	.	.	.	.
H	.	.	.	.	.	.
I	2	2	–3	–5	1	–3
K	.	.	.	.	.	.
L	.	.	.	.	.	.
M	.	.	.	.	.	.
N	0	–3	–5	4	1	–5
Q	.	.	.	.	.	.
P	.	.	.	.	.	.
R	.	.	.	.	.	.
S	0	3	–3	–2	–1	–3
T	.	.	.	.	.	.
V	.	.	.	.	.	.
W	.	.	.	.	.	.
Y	6	–3	–4	–4	0	–4
	+2	+2	+11	+6	+1	+11 =+33

... S I I C D N C N ...

	1	2	3	4	5	6
A	–3	0	–2	–1	0	–3
C	–4	1	11	–5	4	11
D	–4	–3	–6	6	–3	–6
E	.	.	.	.	.	.
F	.	.	.	.	.	.
G	.	.	.	.	.	.
H	.	.	.	.	.	.
I	2	2	–3	–5	1	–3
K	.	.	.	.	.	.
L	.	.	.	.	.	.
M	.	.	.	.	.	.
N	0	–3	–5	4	1	–5
Q	.	.	.	.	.	.
P	.	.	.	.	.	.
R	.	.	.	.	.	.
S	0	3	–3	–2	–1	–3
T	.	.	.	.	.	.
V	.	.	.	.	.	.
W	.	.	.	.	.	.
Y	6	–3	–4	–4	0	–4
	+2	+1	–6	+4	+4	–5 =0

PSSM - construction

1. Given a set of aligned sequences containing the wanted pattern,
2. Compute the relative frequencies $f_{i,c}$ of each amino acid i for each column c
3. Compute the overall frequencies p_i for each amino acid i from all counts
4. Construct the log-odds entries for each position and amino acid of the PSSM using $\mathrm{PSSM}_{i,c} = \frac{1}{\lambda} \log_b \frac{f_{i,c}}{p_i}$.
5. (this works the same for nucleotides)
   Note: PSSMs are also called *position-specific weight matrices (PWM)*

Specialized BLAST developments: Psi-BLAST

Position-specific iterated BLAST does repeated searches using PSSMs to search for distantly related proteins.

1. First search: ordinary BLAST
2. The matches found are combined in an multiple sequence alignment.
3. Generate a consensus sequence and a PSSM from the alignment.
4. From now on: search with the consensus and the PSSM in the DB.
5. Iterate steps 2-4, adding more and more sequences.

This gives better results than blastp, if the first blastp is able to return a reasonable starting set of sequences and if we are searching for distantly related proteins in the DB.

Specialized BLAST developments: Phi-BLAST

Pattern-hit initiated BLAST searches for 'regular expressions' of motifs in a protein database.

- Often we have certain patterns that have to occur in a sequence but we cannot write them down as one sequence,
- then Patterns can help.
- If the Patterns we are searching for must contain a Tryptophane and then after 9-11 residues a Phenylalanine, Valine, or Tyrosine followed by an Alanine
- we can code it as: `W-x(9-11)-[FVY]-A`
- and feed it to Phi-BLAST.

Psi-BLAST and Phi-BLAST

- Psi-BLAST searches for quantitative sequence motifs
- Phi-BLAST searches for qualitative sequence motifs

DB Search Algorithms/Methods

Ordinary Search Strategies:

- Smith-Waterman (1981)
- Baeza-Yates and Perleberg (1992)
- Chang and Lawler (1994)
- Myers (1994)
- FASTA (Pearson and Lipman, 1988)
- BLAST (Altschul et al., 1990)
- Gapped-BLAST/BLAST2 (Altschul et al., 1997)
- BLAT (Kent, 2002)
- ...

Specialized Approaches:

- PSSMs
- HMMs
- Psi-BLAST (Altschul et al., 1997)
- Phi-BLAST (Zhang et al., 1998)
- ...

Introduction to Probability and Statistics of Sequence Alignments

Motivation

We have found that the score of the local alignment between two sequences is S.

- **Question:** What is the 'significance' of this score?
- Differently stated, what is the probability P that the alignment of two random sequences has a score at least equal to S?
- P is the P-value, and is considered a measure of statistical significance.
- If P is small, the initial alignment is significant.

Hypothesis testing

If one wants to decide if the score S of an alignment is significantly larger than expected one needs to test a hypothesis.

1. One typically starts with the Null-hypothesis H_0:
   The sequence pair is not homologous, i.e. the score is expected by chance.
2. Find the alignment or the optimal sub-alignment.
3. Compute the probability distribution under H_0
4. Determine the significance level α you want to work on
5. Determine the actual score S
6. Compute the probability to obtain a score at least as large as S assuming H_0.

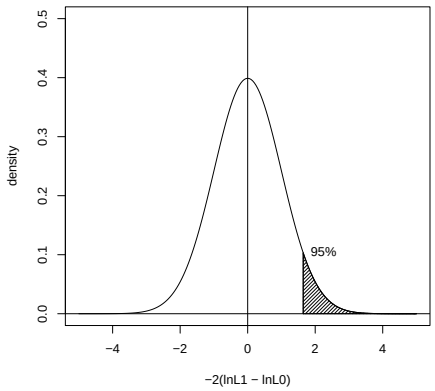
A simple application

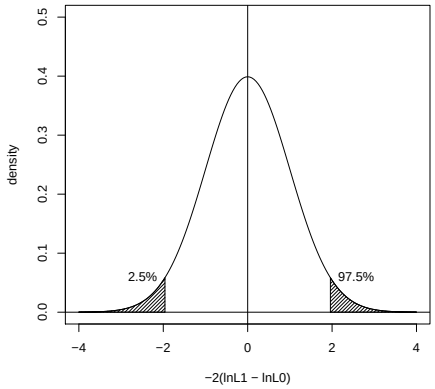
Tossing a coin n times and counting the number of heads H
Null hypothesis H_0: the coin is fair, i.e. the expected number of head $|H|$ and tails $|T|$ are equal or $E(|H| - |T|) = 0$.

The Null hypothesis is typically tested against an alternative hypothesis H_A that depends on the question one wants to answer.

For example: H_A the number of Heads is larger than the number of Tails.

Usual Null-Hypotheses:





First the **Null hypothesis** (H_0) has to be stated, for example:

- **top:** The observed value x is not significantly larger than expected under the Null distribution.
- **bottom:** The observed value x is not significantly different from what is expected under the Null distribution – i.e. the expected value $E(x) = 0$.

If the observed value falls into the white area, the Null hypothesis cannot be rejected.
If it falls into the grey area, this is interpreted as support for the alternative hypothesis by rejecting the Null hypothesis.

Local alignment

What is a local alignment ?

> *'A local alignment without gaps consists simply of a pair of equal length segments, one from each of the two sequences [. . .] whose scores can not be improved by extension or trimming. These are called high-scoring segment pairs or HSPs.'*
> *(NCBI online BLAST tutorial)*

Significance of alignment scores I

$s1$ and $s2$ are two random sequences of length m and n, respectively, then Karlin and Altschul (1990), Altschul et al (1997) showed, that

- The number of pair-wise subalignments with score larger than S, follows approximately a Poisson-distribution with expectation

$$E(S) = Kmne^{-\mu S}$$

where the constants K and μ depend on the scoring matrix.

Significance of alignment scores II

Database query:
Let m be the length of the query sequence s_1 and N be the size of the database D, then the expected number of segment pairs with score larger S equals

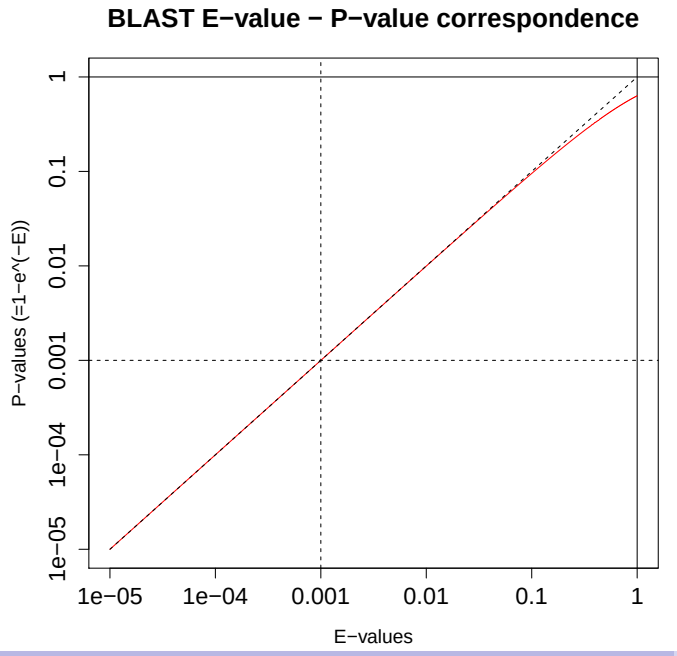
$$E(S) = KmNe^{-\mu S}$$

this is the so-called E-value.

Significance of alignment scores III

The probability of a score of at least S in a database of random sequences can be calculated as follows:

$$P(S \geq x) = 1 - e^{(-KmNe^{-\mu x})} = 1 - e^{-E(x)}$$



Thus, BLAST E-values can be treated similar to P-values if they are at least < 0.001.

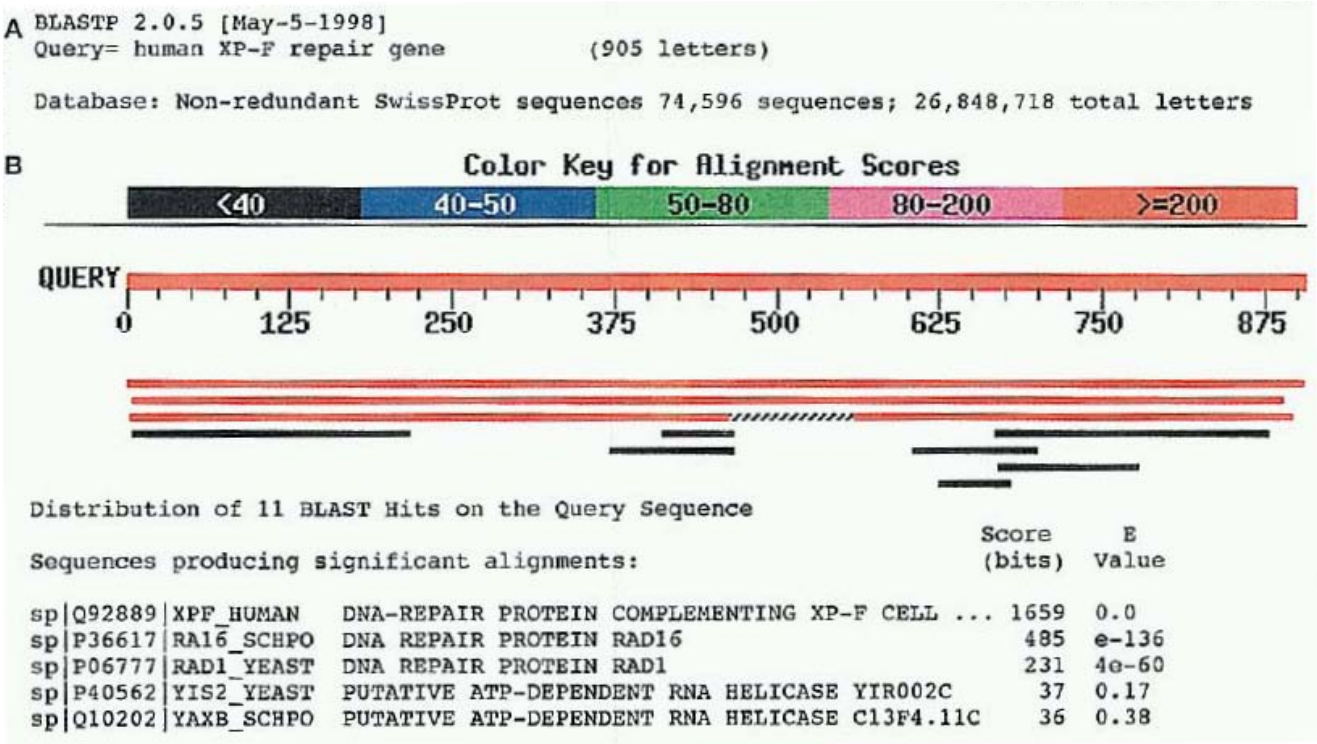
Bit Score of a sequence alignment

Raw scores have little meaning without knowledge of the scoring scheme used for the alignment (or of the parameters K and μ).
Scores can be normalized according to:

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

S is the bit score of the alignment.
Then, the E-value can be simplified as follows: $E = mn2^{-S'}$

BLAST output



Mutually exclusive, independent and dependent events

- If two events are **mutually exclusive**, but alternatives, their probabilities are summed up.
- Example: what is the probability to get either ⚄ or ⚅ when rolling dice.
- If two events are **independent** from each other their probabilities are multiplied.
- Example: from a bag with 3 red and 2 blue marbles (●●●●●), you are drawing one marbel, put it back and draw again (chances do not change).
- Probability of drawing twice a blue marble

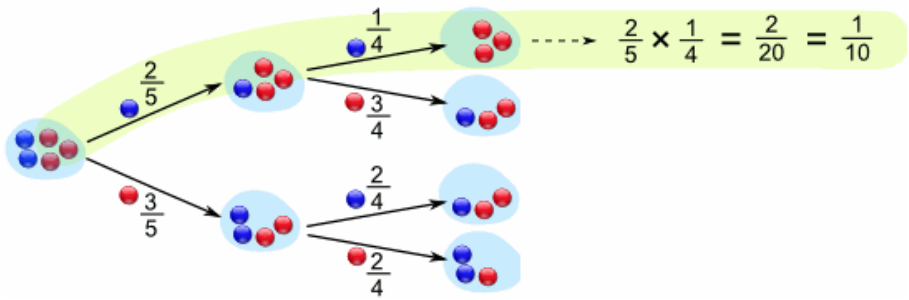
$$P(\bullet\bullet) = P(\bullet) \cdot P(\bullet) = \frac{2}{5} \cdot \frac{2}{5} = \frac{4}{25} = 0.16$$

However, if the probability of one event depends on another, that does not work.

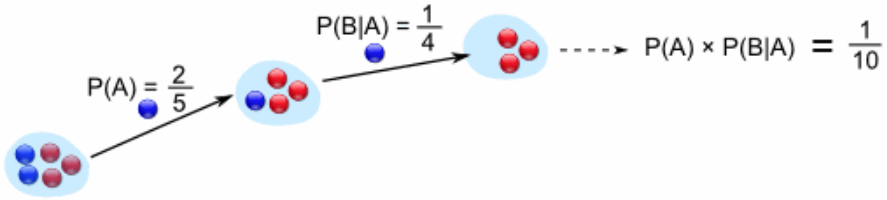
- Example: what is $P(\bullet\bullet)$ if we do not put the marbel back after drawing...?

Conditional probabilities

the conditional probabilities can be listed in a tree diagram:



Conditional probabilities



- such conditional probabilities are denoted as $P(B|A)$ which stands for: *probability of event B given ('|') event A*
- the total probability P(A,B) of events A and B is

$$P(A, B) = P(A) \cdot P(B|A)$$

- or for the marbels in the case that $A = \bullet$ and $B = \bullet$:

$$P(\bullet^{1st}, \bullet^{2nd}) = P(\bullet^{1st}) \cdot P(\bullet^{2nd}|\bullet^{1st})$$