(Aspekte der Thermodynamik in der Strukturbiologie)

Einführung in die Bioinformatik

Wintersemester 2012/13

Peter Güntert

BLAST Algorithm

Significance of Alignments

Outline

- BLAST algorithm
- Significance of Alignments

BLAST: background on sequence alignment

There are two main approaches to sequence alignment:

- Global alignment (Needleman & Wunsch 1970) using dynamic programming to find optimal alignments between two sequences. (Although the alignments are optimal, the search is not exhaustive.) Gaps are permitted in the alignments, and the total lengths of both sequences are aligned (hence "global").
- Local sequence alignment (Smith & Waterman, 1980). The alignment may contain just a portion of either sequence, and is appropriate for finding matched domains between sequences.

- BLAST is a heuristic approximation to local alignment. It examines only part of the search space.

How a BLAST search works

"The central idea of the BLAST algorithm is to confine attention to segment pairs that contain a word pair of length w with a score of at least T."

Altschul et al. (1990)

How the original BLAST algorithm works: Phase 1

- Compile a list of word pairs (w = 3) above threshold T
- Example: for a human RBP query
  …**FSGTWYA**… (query word is in red)

  A list of words (w = 3) is:

  FSG SGT GTW TWY WYA
  YSG TGT ATW SWY WFA
  FTG SVT GSW TWF WYS

Phase 1: compile a list of words (*w* = 3)

neighborhood
word hits
> threshold
(T=11)

neighborhood
word hits
< below threshold

GTW	6,5,11	22
GSW	6,1,11	18
ATW	0,5,11	16
NTW	0,5,11	16
GTY	6,5,2	13
GNW		10
GAW		9

Pairwise alignment scores are determined using a scoring matrix such as Blosum62

	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

BLAST: Word size

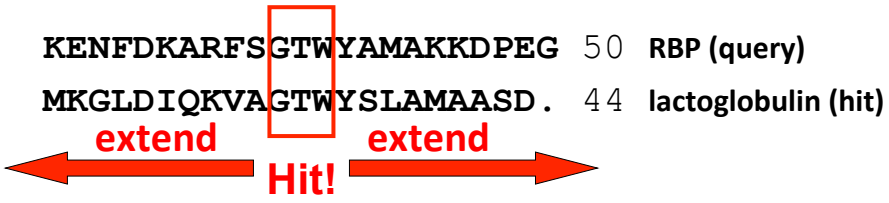
- For proteins, the word size is normally *w* = 3.
- For nucleic acids, the word size is typically 7, 11, or 15 (EXACT match). Changing word size is like changing threshold of proteins. *w* = 15 gives fewer matches and is faster than *w* = 11 or *w* = 7.
- For megablast, the word size is 28 and can be adjusted to 64. What will this do? Megablast is very fast for finding closely related DNA sequences!

How a BLAST search works: Phase 2

- Scan the database for entries that match the compiled list.
- This is fast and relatively easy.

How a BLAST search works: Phase 3

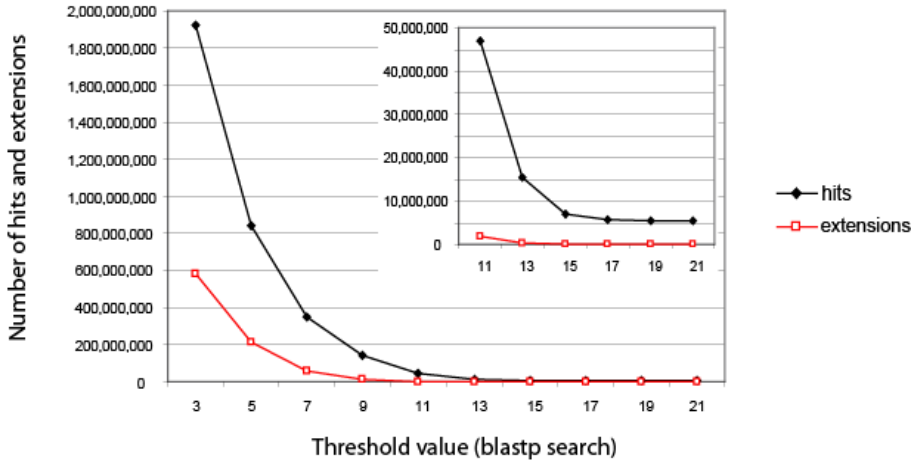
- Phase 3: when you find a hit (i.e. a match between a "word" and a database entry), extend the hit in either direction.
- Keep track of the score (use a scoring matrix).
- Stop when the score drops below some threshold.



How a BLAST search works: Phase 3

- In the original (1990) implementation of BLAST, hits were extended in either direction.
- In a 1997 refinement of BLAST, two independent hits are required. The hits must occur in close proximity to each other. With this modification, only one seventh as many extensions occur, greatly speeding the time required for a search.

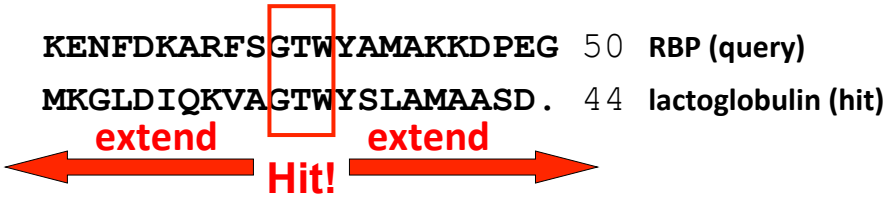
Number of hits/extensions



Human β globin

How a BLAST search works: Phase 3

- Phase 3: when you manage to find a hit (i.e. a match between a "word" and a database entry), extend the hit in either direction.
- Keep track of the score (use a scoring matrix).
- Stop when the score drops below some cutoff.



Alignment scores for random sequences

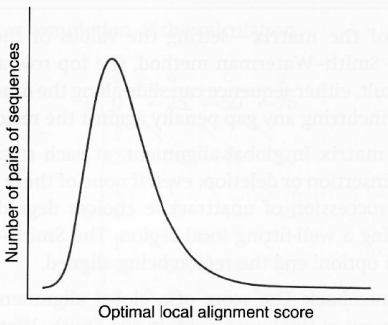


Fig. 5.5 Optimal local alignment scores for pairs of random amino acid sequences of the same length follow an extreme value distribution. For any score x, the probability of observing a score $\geq x$ is:

$$P(\text{Score} \geq x) = 1 - \exp(-Ke^{-\lambda x}),$$

where K and λ are parameters related to the position of the maximum and the width of the distribution. Note the long tail at the right. This means that a score several standard deviations above the mean has a higher probability of arising by chance (i.e. it is *less* significant) than if the scores followed a normal distribution.

How to play with matches but not get burned

- Pairwise alignments and database searches often show tenuous but tantalizing sequence similarities.
- **How can we decide whether we are seeing a true relationship?**
- Statistics cannot answer biological questions directly, but can tell us the likelihood that a similarity as good as the one observed would appear, just by chance, among unrelated sequences.
- To do this, we want to compare our result with alignments of the same sequences to a large population. This 'control' population should be similar in general features to our aligned sequences, but should contain few sequences related to them.
- Only if the observed match stands out from the population can we regard it as significant.

Control populations of sequences

- To what population of sequences should we compare our alignment?
  - For pairwise alignments, we can pick one of the two sequences, make many scrambled copies of it using a random-number generator, and align each permuted copy to the second sequence.
  - For probing a database, the entire database provides a comparison population.
- Alignments of our sequence to each member of the control population generates a large set of scores.
- How does the score of our original alignment rate?

Z-scores

- If randomized sequences score as well as the original one, the alignment is unlikely to be significant.
- We can measure the mean and standard deviation of the scores of the alignments of randomized sequences, and ask whether the score for the original sequences is unusually high.
- The *Z*-score reflects the extent to which the original result is an outlier from the population:

$$\text{Z-score} = \frac{\text{score} - \text{mean}}{\text{standard deviation}}$$

Statistical parameters to evaluate the significance of alignments

The *Z*-score is a measure of how unusual our original match is, in terms of the mean and standard deviation of the population scores. If the original alignment has score S,

$$Z\text{-score of S} = \frac{\text{S} - \text{mean}}{\text{standard deviation}}$$

A *Z*-score of 0 means that the observed similarity is no better than the average of the control population, and might well have arisen by chance. The higher the *Z*-score, the greater the probability that the observed alignment has not arisen simply by chance. Experience suggests that *Z*-scores ≥ 5 are significant.

Statistical parameters to evaluate the significance of alignments

- Many programs report P = the probability that a random alignment will give a better score than the given alignment.
- The relationship between Z and P depends on the distribution of the scores from the control population, which do not follow the normal distribution.

A rough guide to interpreting *P*-values:

$P \leq 10^{-100}$	exact match
P in range $10^{-100} - 10^{-50}$	sequences very nearly identical, e.g. alleles or SNPs
P in range $10^{-50} - 10^{-10}$	closely related sequences, homology certain
P in range $10^{-5} - 10^{-1}$	usually, distant relatives
$P > 10^{-1}$	match probably insignificant

Statistical parameters to evaluate the significance of alignments

For database searches, some programs (including PSI-BLAST) report *E*-values. The *E*-value of an alignment is the expected number of sequences that give the same *Z*-score or better if the database is probed with a random sequence. *E* is found by multiplying the value of *P* by the size of the database probed. Note that *E* but not *P* depends on the size of the database. Values of *P* are between 0 and 1.0. Values of *E* are between 0 and the number of sequences in the database searched.

A rough guide to interpreting *E*-values:

$E \leq 0.02$	sequences probably homologous
E between 0.02 and 1	homology unproven but can't be ruled out
$E > 1$	you'd have to expect this good a match just by chance

Statistics are a useful guide, but not a substitute for thinking carefully about the results, and further analysis of ones that look promising!

Rules of thumb based on sequence identity

- Many 'rules of thumb' are expressed in terms of percentage identical residues in the optimal alignment.
- Two proteins with > 45% identical residues in their optimal alignment have very similar structures, and are very likely to have a common or at least a similar function.
- Two proteins with > 25% identical residues are likely to have a similar general folding pattern. On the other hand, observation of a lower degree of sequence similarity cannot rule out homology.
- R. F. Doolittle defined the region of 18-25% sequence identity as the 'twilight zone' in which the suggestion of homology is tantalizing but dangerous.
- Below the twilight zone is a region where pairwise sequence alignments tell very little.

Rules of thumb based on sequence identity

- Lack of significant sequence similarity does not preclude genuine homology.
- Although the twilight zone is a treacherous region, we are not entirely helpless.
- In deciding whether there is a genuine relationship, the 'texture' of the alignment is important: Are the similar residues isolated and scattered throughout the sequence, or are there 'icebergs', local regions of high similarity, which may correspond to a shared active site?
- We may need to rely on other information, about shared ligands or function. Of course if the structures are known, we could examine them directly.

Examples of low sequence similarity

- Sperm whale myoglobin and lupin leghaemoglobin have 15% identical residues in optimal alignment. This is even below the twilight zone. But we also know that both molecules have similar three-dimensional structures, both contain a haem group and both bind oxygen. They are indeed distantly related homologues.
- The sequences of the N- and C- terminal halves of rhodanese have 11% identical residues in optimal alignment. If these appeared in independent proteins, one could not conclude from the sequences alone that they were related. However, their appearance in the same protein suggests that they arose via gene duplication and divergence. The striking similarity of their structures confirms their relationship.

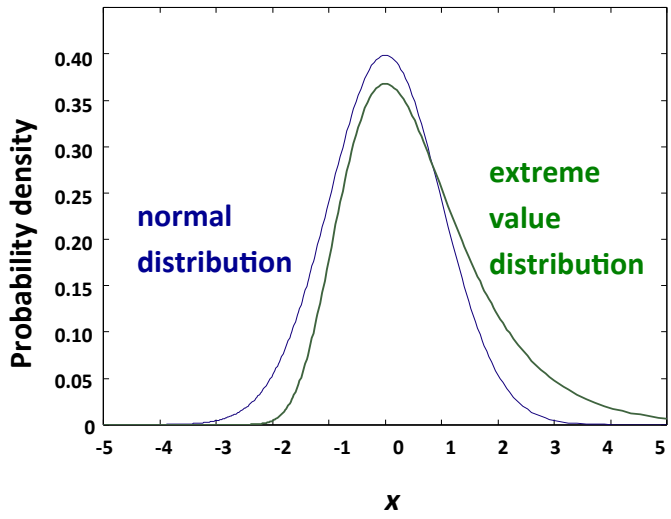
Examples of low sequence similarity

- As a cautionary note, consider the proteinases chymotrypsin and subtilisin. They have 12% identical residues in optimal alignment. These enzymes have a common function, and a common catalytic triad. However, they have dissimilar folding patterns, and are not related. Their common function and mechanism is an example of convergent evolution.
- This case serves as a warning against special pleading for relationships between proteins with dissimilar sequences on the basis of similarities of function and mechanism.

How to interpret a BLAST search: expect value

- It is important to assess the statistical significance of search results.
- For global alignments, the statistics are poorly understood.
- For local alignments (including BLAST search results), the statistics are well understood. The scores follow an extreme value distribution (EVD) rather than a normal distribution.

Probability density function of the extreme value distribution (characteristic value u = 0 and decay constant λ = 1)



How to interpret a BLAST search: expect value

- The expect value *E* is the number of alignments with scores greater than or equal to score *S* that are expected to occur by chance in a database search.
- An *E* value is related to a probability value *p*.
- The key equation describing an *E* value is:

$$E = Kmn\, e^{-\lambda S}$$

$E = Kmn\, e^{-\lambda S}$

This equation is derived from a description of the extreme value distribution.

S = the score

E = the expect value = the number of high-scoring segment pairs (HSPs) expected to occur with a score of at least S

m, n = the length of two sequences

λ, K = Karlin Altschul statistics

Some properties of the equation $E = Kmn\, e^{-\lambda S}$

- The value of E decreases exponentially with increasing S (higher S values correspond to better alignments). Very high scores correspond to very low E values.
- The E value for aligning a pair of random sequences must be negative! Otherwise, long random alignments would acquire great scores.
- Parameter K describes the search space (database).
- For E = 1, one match with a similar score is expected to occur by chance. For a very much larger or smaller database, you would expect E to vary accordingly.

From raw scores to bit scores

- There are two kinds of scores:
  - raw scores (calculated from a substitution matrix)
  - bit scores (normalized scores)
- Bit scores are comparable between different searches because they are normalized to account for the use of different scoring matrices and different database sizes.
- Bit score: $S' = (\lambda S - \ln K) / \ln 2$
- The E value corresponding to a given bit score is: $E = mn\, 2^{-S'}$
- Bit scores allow you to compare results between different database searches, even using different scoring matrices.

How to interpret BLAST: *E* values and *p* values

- The expect value E is the number of alignments with scores greater than or equal to score S that are expected to occur by chance in a database search.
- The p value is the probability of a chance alignment with the score in question or better.

$$p = 1 - e^{-E}$$

How to interpret BLAST: *E* values and *p* values

- Very small *E* values are very similar to *p* values.
- *E* values of about 1 to 10 are far easier to interpret than corresponding *p* values.

E	*p*
10	0.99995460
5	0.99326205
2	0.86466472
1	0.63212056
0.1	0.09516258 (about 0.1)
0.05	0.04877058 (about 0.05)
0.001	0.00099950 (about 0.001)
0.0001	0.0001000

BLAST web server

http://blast.ncbi.nlm.nih.gov

Unterlagen zur Vorlesung

http://www.bpc.uni-frankfurt.de/guentert/wiki/index.php/Teaching