BLAST and FASTA Heuristics in pairwise sequence alignment

(based on materials of Christoph Dieterich
Department of Evolutionary Biology
Max Planck Institute for Developmental Biology)

Heuristics for large-scale database searching

- Pairwise alignment is used to detect homologies between different protein or DNA sequences, e.g. as global or local alignments.
- Problem solved using dynamic programming in *O(nm)* time and *O(n)* space.
- This is **too slow** for searching current databases.
- In practice algorithms are used that run much faster, at the expense of possibly missing some significant hits due to the heuristics employed.
- Such algorithms are usually *seed-and-extend* approaches in which first small exact matches are found, which are then extended to obtain long inexact ones.

Preprocessing

- Preprocessing should save time for subsequent searches, but the databases are changing – they are split into fixed and a dynamic part. Fixed part is preprocessed and the results of the preprocessing is stored in appropriate structures, e.g. hash tables.
- Information about substrings of length n can be stored in a hash table. For an alphabet Σ there are $|\Sigma|^n$ different substrings of length n.
- We will describe two methods
  - BLAST
  - FASTA

BLAST (1)

- BLAST, the Basic Local Alignment Search Tool (Altschul et al., 1990), is an alignment heuristic that determines “local alignments” between a query and a database. It is based on Smith-Waterman algorithm (local alignment).
- BLAST consists of two components:
  1. a search algorithm and
  2. a computation of the statistical significance of solutions.

BLAST (2)

- Let *q* be the query and *d* the database. A *segment* is simply a substring *s* of *q* or *d*.
- A segment-pair (*s, t*) (or hit) consists of two segments, one in *q* and one *d*, of the same length.

  Example:

  V A L L A R

  P A M M A R
- We think of *s* and *t* as being aligned without gaps and score this alignment using a substitution score matrix, e.g. BLOSUM or PAM in the case of protein sequences.
- The alignment score for (*s, t*) is denoted by $\sigma(s, t)$.

BLAST (3)

- A *locally maximal segment pair (LMSP)* is any segment pair (s,t) whose score cannot be improved by shortening or extending the segment pair.
- A *maximum segment pair (MSP)* is any segment pair (s, t) of maximal alignment score $\sigma(s, t)$.
- Given a cut-off score S, a segment pair (s, t) is called a *high-scoring segment pair (HSP)*, if it is locally maximal and $\sigma(s, t) \geq S$.
- Finally, a *word* is simply a short substring of fixed length w.

BLAST (4) – goal

- **Goal**: Find all HSPs for a given cut-off score.

- Three parameters:
  - a word size *w*,
  - a word similarity threshold *T* and
  - a minimum cut-off score *S*.
- We are looking for a segment pair with a score of at least *S* that contains at least one word pair of length *w* with score at least *T*.

BLAST (5) – Preprocessing

1. For the query *q*, generate all subwords of length *w*.
2. Generate a list of all *w*-mers of length *w* over the alphabet Σ that have similarity > *T* to some subword in the query sequence *q*.

Example: For the query sequence RQCSAGW the list of words of length *w* = 2 with a score *T* > 8 using the BLOSUM62 matrix are:

word	2-mer with score > 8
RQ	RQ
QC	QC, RC, EC, NC, DC, KC, MC, SC
CS	CS,CA,CN,CD,CQ,CE,CG,CK,CT
SA	-
AG	AG
GW	GW,AW,RW,NW,DW,QW,EW,HW,KW,PW,SW,TW,WW

BLAST (6) – Searching

- **Localization** of the hits: The database sequence *d* is scanned for all hits *t* of *w*-mer *s* in the list, and the positions of the hits are saved.
- **Detection** of hits: First all pairs of hits are searched that have a distance of at most *A* (think of them lying on the same diagonal in the matrix of the SW-algorithm).
- **Extension** to HSPs: Each such seed (s, t) is extended in both directions until its score $\sigma(s, t)$ cannot be enlarged (LMSP). Then all best extensions are reported that have score $\geq S$, these are the HSPs.

- In practice, $w = 3$ and $A = 40$ for proteins.
- Originally the extension did not include gaps, a newer BLAST2 algorithm allows insertion of gaps.

BLAST (7) – Searching

- The list L of all words of length w that have similarity $> T$ to some word in the query sequence q can be produced in $O(|L|)$ time.
- These are placed in a “keyword tree” and then, for each word in the tree, all exact locations of the word in the database d are detected in time linear to the length of d.
- As an alternative to storing the words in a tree, a finite-state machine can be used.

BLAST (8) : Extension

- As BLAST does not allow indels at that stage, hit extension is very fast.
- Use of seeds of length *w* and the termination of extensions with fading scores (score drop-off threshold *X*) are both steps that speed up the algorithm, but also imply that BLAST is not guaranteed to find all HSPs (after all it is a heuristic).

- Recent improvements (BLAST 2.0):
  - Two word hits must be found within a window of A residues.
  - Explicit treatment of gaps.
  - Position-specific iterative BLAST (PSI-BLAST).

BLAST for DNA

For DNA sequences, BLAST operates as follows:

- The list of all words of length *w* in the query sequence *q* is generated. In practice, *w = 12* for DNA.
- The database *d* is scanned for all hits of words in this list.
- Blast uses a two-bit encoding for DNA. This saves space and also search time, as four bases are encoded per byte.
- Note that the "*T*" parameter dictates the speed and sensitivity of the search.

Different versions of BLAST

BLASTN: compares a DNA query sequence to a DNA sequence database;	$q_{\text{DNA}} \leftrightarrow s_{\text{DNA}}$
BLASTP: compares a protein query sequence to a protein sequence database;	$q_{\text{prot}} \leftrightarrow s_{\text{prot}}$
TBLASTN: compares a protein query sequence to a DNA sequence database (6 frames translation);	$q_{\text{prot}} \leftrightarrow \begin{matrix} s_{t_1(DNA)} & s_{t_1^c(DNA)} \\ s_{t_2(DNA)} & s_{t_2^c(DNA)} \\ s_{t_3(DNA)} & s_{t_3^c(DNA)} \end{matrix}$
BLASTX: compares a DNA query sequence (6 frames translation) to a protein sequence database.	$\begin{matrix} q_{t_1(DNA)} & q_{t_1^c(DNA)} \\ q_{t_2(DNA)} & q_{t_2^c(DNA)} \\ q_{t_3(DNA)} & q_{t_3^c(DNA)} \end{matrix} \leftrightarrow s_{\text{prot}}$
TBLASTX: compares a DNA query sequence (6 frames translation) to a DNA sequence database (6 frames translation).	$\begin{matrix} q_{t_1(DNA)} & q_{t_1^c(DNA)} \\ q_{t_2(DNA)} & q_{t_2^c(DNA)} \\ q_{t_3(DNA)} & q_{t_3^c(DNA)} \end{matrix} \leftrightarrow \begin{matrix} s_{t_1(DNA)} & s_{t_1^c(DNA)} \\ s_{t_2(DNA)} & s_{t_2^c(DNA)} \\ s_{t_3(DNA)} & s_{t_3^c(DNA)} \end{matrix}$

BLAST – statistical analysis

Problem:

Given an HSP (s, t) with score $\sigma(s, t)$. How significant is this match (i.e., local alignment)?

Steps:

1. The null hypothesis H_0 states that the two sequences (s, t) are ***not*** homologous. Then the alternative hypothesis states that the two sequences ***are homologous***.
2. Choose an experiment to find the pair (s, t): use BLAST to detect HSPs.
3. Compute the probability of the result under the hypothesis H_0, $P(\text{Score} \geq \sigma(s, t) \mid H_0)$ by generating a probability distribution with random sequences.
4. Fix a rejection level for H_0.
5. Perform the experiment, compute the probability of achieving the result or higher and compare with the rejection level.

BLAST – statistical analysis

Poisson distribution

- The Karlin and Altschul theory for local alignments (without gaps) is based on Poisson and extreme value distributions. The details of that theory are beyond the scope of this lecture, but basics are sketched in the following.
- The Poisson distribution with parameter v is given by

$$P(X = x) = \frac{v^x}{x!} e^{-v} \quad x = 0,1,2,\ldots$$

- Note that v is the expected value as well as the variance. From the equation we follow that the probability that a variable X will have a value at least x is

$$P(X \geq x) = 1 - \sum_{i=0}^{x-1} \frac{v^i}{i!} e^{-v}$$

- **Problem**
  - Given an HSP (s, t) with score $\sigma(s, t)$. How significant is this match (i.e., local alignment)?

BLAST – statistical analysis

Statistical significance of an HSP

- Given the scoring matrix $S(a, b)$, the expected score for aligning a random pair of amino acid is required to be negative:

$$E = \sum_{a,b \in \Sigma} p_a p_b S(a,b) < 0$$

  otherwise long random alignments would have large score.
- The sum of a large number of independent identically distributed (i.i.d) random variables tends to a normal distribution. The maximum of a large number of i.i.d. random variables tends to an extreme value distribution as we will see below.
- HSP scores are characterized by two parameters, K and λ. The parameters K and λ depend on the background probabilities of the symbols and on the employed scoring matrix. λ is the unique value for y that satisfies the equation

$$\sum_{a,b \in \Sigma} p_a p_b e^{S(a,b)y} = 1$$

  K and λ are scaling-factors for the search space and for the scoring scheme, respectively.

BLAST – statistical analysis

Statistical significance of an HSP

- The number of random HSPs (s, t) with $\sigma(s, t) \geq S$ can be described by a Poisson distribution with parameter $v = Kmne^{-S}$. The number of HSPs with score $\geq S$ that we expect to see due to chance is then the parameter v, also called the E-value:

$$E(\text{HSPs with score} \geq S) = Kmne^{-\lambda S} = mnKe^{-\lambda S}$$

m – the length of the query string
n – the length of the database

- Hence, the probability of finding exactly x HSPs with a score $\geq S$ is given by

$$P(X = x) = e^{-E}\frac{E^x}{x!}$$

where E is the E-value for S (i.e. v).

- The probability of finding at least one HSP "by chance" is

$$P(S) = 1 - P(X = 0) = 1 - e^{-E}$$

BLAST – statistical analysis

bit score

- We would like to "hide" the parameters K and λ to make it easier to compare results from different BLAST searches.
- For a given HSP (s, t) we transform the raw score S into a bit-score:

$$Ke^{-\lambda S} = 2^{-S'}$$

$$S' = -\log_2\left(Ke^{-\lambda S}\right) = -\frac{\ln\left(e^{\ln K - \lambda S}\right)}{\ln 2} = \frac{\lambda S - \ln K}{\ln 2}$$

- Such bit-scores can be compared between different BLAST searches, as the parameters of the given scoring systems are subsumed in them.

$$E = mn2^{-S'}$$

BLAST – statistical analysis

E-value

- To determine the significance of a given bit-score *S'* the only additional value required is the size of the search space. Since $S = (S' \ln 2 + \ln K)/\lambda$, we can express the *E*-value in terms of the bit-score as follows:

$$E = Kmne^{-\lambda S} = Kmne^{-(S'\ln 2 + \ln K)} = mn\,2^{-S'}$$

- For each match an expectation values *E* is computed (it is printed in scientific notation – an exponent only) the smaller the number, i.e. the closer it is to 0, the more significant the match is. Expectation values show us how often we could expect that particular alignment match to occur merely by chance alone in a search of that size database.
- This is the P-value associated with the score S. For example, if one expects to find three HSPs with score >= S, the probability of finding at least one is 0.95. The BLAST programs report E-value rather than P-values because it is easier to understand the difference between, for example, E-value of 5 and 10 than P-values of 0.993 and 0.99995. However, when E < 0.01, P-values and E-value are nearly identical.

FASTA

- FASTA (pronounced fast-ay) is a heuristic for finding significant matches between a query string *q* and a database string *d*. It is the older of the two heuristics introduced in the lecture.
- FASTA's general strategy is to find the most significant diagonals in the dot-plot or dynamic programming matrix.
- The algorithm consists of four phases:

  Phase 1: Hashing,

  Phase 2: 1st scoring,

  Phase 3: 2nd scoring,

  Phase 4: alignment.

FASTA Phase 1: hashing

- The first step of the algorithm is to determine all exact matches of length k (word-size) between the two sequences, called hot-spots.
- A hot-spot is given by (i, j), where i and j are the locations (i.e., start positions) of an exact match of length k in the query and database sequence respectively.
- Any such hot-spot (i, j) lies on the diagonal $(i - j)$ of the dot-plot or dynamic programming matrix. Using this scheme, the main diagonal has number 0 ($i = j$), whereas diagonals above the main one have positive numbers ($i < j$), the ones below negative ($i > j$).
- A diagonal run is a set of hot-spots that lie in a consecutive sequence on the same diagonal. It corresponds to a gapless local alignment.
- A score is assigned to each diagonal run. This is done by giving a positive score to each match (using e.g. the PAM250 match score matrix in the case of proteins) and a negative score for gaps in the run, the latter scores decrease with increasing length of the gap between hot-spots.
- The algorithm then locates the ten best diagonal runs.

FASTA Phase 2+3: scoring

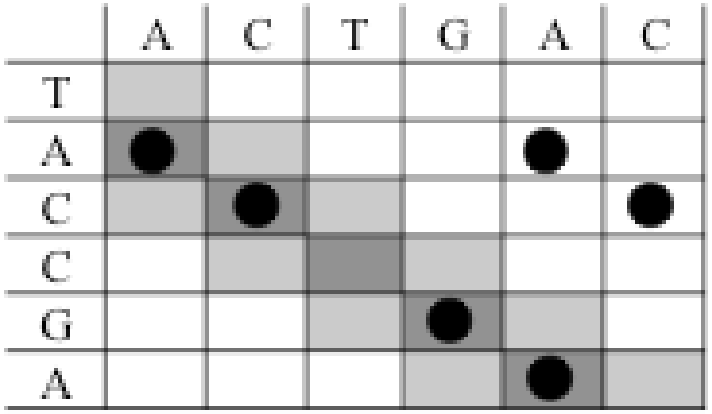
- Each of the ten diagonal runs with highest score are further processed. Within each of these scores an optimal local alignment is computed using the match score substitution matrix. These alignments are called initial regions.
- The score of the best sub-alignment found in this phase is reported as **init1**.
- The next step is to combine high scoring sub-alignments into a single larger alignment, allowing the introduction of gaps into the alignment. The score of this alignment is reported as **initn.**

FASTA Phase 4: alignment

- Finally, a banded Smith-Waterman dynamic program is used to produce an optimal local alignment along the best matched regions. The center of the band is determined by the region with the score init1, and the band has width 8. The score of the resulting alignment is reported as **opt**.
- In this way, FASTA determines a highest scoring region, not all high scoring alignments between two sequences. Hence, FASTA may miss instances of repeats or multiple domains shared by two proteins.
- After all sequences of the databases have thus been searched a statistical significance similar to the BLAST statistics is computed and reported.

FASTA example

- Two sequences ACTGAC and TACCGA: The hot spots for $k = 2$ are marked as pairs of black bullets, a diagonal run is shaded in dark grey. An optimal sub-alignment in this case coincides with the diagonal run. The light grey shaded band of width 3 around the sub-alignment denotes the area in which the optimal local alignment is searched.



Comparing BLAST and FASTA

- BLAST: individual seeds are found and then extended without indels.
- FASTA: individual seeds contained in the same diagonal are merged and the resulting segments are then connected using a banded Smith-Waterman alignment.