5 Blast

This lecture is based on the following, which are all recommended reading:

- R. Merkl, S. Waack: Bioinformatik Interaktiv. Chapter 11.4-11.7

5.1 Introduction

Pairwise alignment is used to detect homologies between different protein or DNA sequences, usually as global or local alignments.

Solving this by dynamic programming takes quadratic time, that is, time proportional to the product of the lengths of the two sequences being compared.

This is too slow for searching current databases. Hence, in practice heuristics are used that run much faster, at the expense of possibly missing some significant hits.

Such heuristics are usually based on a seed-and-extend approach in which first small exact matches are found, which are then extended to obtain long inexact ones.

5.2 Overview

BLAST, the Basic Local Alignment Search Tool (Altschul et al., 1990), is perhaps the most widely used bioinformatics tool ever written. It is an alignment heuristic that determines “local alignments” between a query and a database. It uses an approximation of the Smith-Waterman algorithm.

One important aspect of BLAST is that it provides statistics to distinguish random from real similarities.

BLAST consists of two components: a search algorithm and computation of the statistical significance of solutions.

BLAST starts with the localization of substrings (so-called segment pairs or hits) in two sequences that have a certain similarity score. The hits are the starting point for deriving HSPs, locally optimal pairs that contain the hit. Extending to the left or right of an HSP would lead to a lower score.

BLAST searches for alignments local alignments between a query and a database.
BLAST consists of two components:

1. A heuristic for finding local alignments, and
2. a computation of the statistical significance of solutions to distinguish random from real similarities.

The heuristic for finding local alignments operates as follows:

1. First search for pairs substrings (so-called segment pairs or hits) in two sequences that have a certain similarity score.
2. The hits are considered “seeds”.
3. Hits are extended to HSPs - High-Scoring Segment Pairs until the “score” of the extended hit drops by a certain amount.

5.3 BLAST terminology

**Definition 5.3.1** Let $q$ be the query and $d$ the database over an alphabet $\Sigma$. A segment is a substring $s$ of $q$ or $d$. Let $S_{\Sigma}$ be a scoring matrix. A segment-pair $(s, t)$ (or hit) consists of two segments, one in $q$ and one in $d$, of the same length. The score $\sigma(s, t)$ of the segment pair is the alignment score for $(s, t)$ given $S_{\Sigma}$.

Example:

$$\begin{align*}
\text{V A L L A R} \\
\text{P A M M A R} \\
\sigma(s, t) \text{ (using BLOSUM62)} &= 15
\end{align*}$$

**Definition 5.3.2** 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 cutoff 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 $w$, and a word pair is a pair of short substrings of fixed length $w$.

Given the cut-off score $S$, the goal of BLAST is to compute all HSPs.

5.4 Preprocessing

Given three parameters, i.e. a word size $w$, a word similarity threshold $T$ and a minimum cut-off score $S$, the goal is to find 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 operates as follows:

**Preprocessing:** Of the query sequence $q$ first all words of length $w$ are generated. Then a list of all $w$-mers of length $w$ over the alphabet $\Sigma$ that have similarity $> T$ to some word in the query sequence $q$ is generated.

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

The search algorithm consists then of three steps:

1. **Localization of the hits**: The database sequence $d$ is scanned for all hits $t$ of $w$-mers $s$ in the list, and the position of the hit is saved.

2. **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).

3. **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. Originally the extension did not include gaps, the modern BLAST2 algorithm allows insertion of gaps.

In practice, $w = 3$ and $A = 40$ for proteins, and $w = 12$ for DNA sequences.

5.5 The BLAST family

There are a number of different variants of the BLAST program:

- **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; $q_{\text{protein}} \leftrightarrow s_{\text{protein}}$

- **TBLASTN**: compares a protein query sequence to a DNA sequence database (6 frames translation);
  $q_{\text{protein}} \leftrightarrow s_{t1}(\text{DNA})$, $q_{\text{protein}} \leftrightarrow s_{t2}(\text{DNA})$, $q_{\text{protein}} \leftrightarrow s_{t3}(\text{DNA})$, $q_{\text{protein}} \leftrightarrow s_{t1}(\text{DNA})$, $q_{\text{protein}} \leftrightarrow s_{t2}(\text{DNA})$, $q_{\text{protein}} \leftrightarrow s_{t3}(\text{DNA})$

- **BLASTX**: compares a DNA query sequence (6 frames translation) to a protein sequence database;
  $q_{t1}(\text{DNA}) \leftrightarrow s_{\text{protein}}$, $q_{t2}(\text{DNA}) \leftrightarrow s_{\text{protein}}$, $q_{t3}(\text{DNA}) \leftrightarrow s_{\text{protein}}$, $q_{t1}(\text{DNA}) \leftrightarrow s_{\text{protein}}$, $q_{t2}(\text{DNA}) \leftrightarrow s_{\text{protein}}$, $q_{t3}(\text{DNA}) \leftrightarrow s_{\text{protein}}$

- **TBLASTX**: compares a DNA query sequence (6 frames translation) to a DNA sequence database (6 frames translation); $q_{t1}(\text{DNA}) \leftrightarrow s_{t1}(\text{DNA})$, $\cdots$, $q_{t5}(\text{DNA}) \leftrightarrow s_{t5}(\text{DNA})$

5.6 Statistical analysis

When a local alignment has been computed, next one needs to assess its statistical significance. This is done by the general approach of hypothesis testing.
5.6.1 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.

Definition 5.6.1 The Poisson distribution with parameter $v$ is given by

$$P(X = x) = \frac{v^x}{x!} e^{-v}$$  \hspace{1cm} (5.1)

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

$$P(X \geq x) = 1 - \sum_{i=0}^{x-1} \frac{v^i}{i!} e^{-v}$$  \hspace{1cm} (5.2)

5.6.2 Statistical significance of an HSP

Given an HSP $(s,t)$ with score $\sigma(s,t)$.

How significant is this match (i.e., local alignment)? To analyze how high a score is likely to arise by chance, a model of random sequences over the alphabet $\Sigma$ is needed.

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

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

Were this not the case, long alignments would tend to have high score independently of whether the segments aligned were related, and the statistical theory would break down.

5.6.3 Statistical significance

Assume that the length $m$ and $n$ of the query and database respectively are sufficiently large.

The number of random HSPs $(s,t)$ with $\sigma(s,t) \geq S$ can be described by a Poisson distribution with parameter $v = Kmne^{-\lambda S}$. 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:

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

The parameters $K$ and $\lambda$ depend on the background probabilities of the symbols and on the employed scoring matrix. $\lambda$ is the unique value for $y$ that satisfies the equation

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

$K$ and $\lambda$ are scaling-factors for the search space and for the scoring scheme, respectively.

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 $S$.

The probability of finding at least one HSP “by chance” is

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

Thus we see that the probability distribution of the scores follows an *extreme value distribution*. BLAST reports $E$-values rather than “$P$-values” as it is easier to interpret the difference between $E$-values than to interpret the difference between $P$-values.

### 5.6.4 BLAST as a web service

[BLAST home page](http://www.ncbi.nlm.nih.gov/BLAST/)
5.6.5 BLAST run example

5.6.6 BLAST run example
5.6.7 BLAST run output

BLAST run output

5.6.8 BLAST run output
5.7 Remarks

- Another well-known heuristic for pairwise local alignments is FASTA (and its variants), developed by Lipman and Pearson \( ^1 \)

---