Genomanalyse

Vorlesung Genomanalyse vom 9.11.2010

Mario Stanke
Institut für Mathematik und Informatik
Universität Greifswald
1 Biological Databases
   DBs in General
   Sequence DBs

2 Homology Search
   Search in General
   Heuristic homology search
   Motivation
   BLAST
References

- “Bioinformatik Interaktiv”, Rainer Merkl und Stephan Waack, Wiley, 2009
Large amounts of biological or biochemical data are usually stored in databases, frequently publicly available.

*Nucleic Acids Research* 2011:
1330 databases covering various aspects of molecular and cell biology

### nonexhaustive list of DB categories

- nucleotide sequence databases (*GenBank*, European Nucleotide Archive, TRANSFAC)
- RNA sequence databases (HIV sequence database (LANL), Ribosomal Database Project (RDP))
- protein sequence databases (*UniProt*, *RefSeq*, SwissProt, Blocks, InterPro, *Pfam*, CyMoBase, COG)
- structure databases (*Protein Data Bank* (PDB), SCOP)
- genomics databases (z.B. *UCSC Genome Browser*, FlyBase, TAIR, Wormbase)
- metabolic and signaling pathways (KEGG)
- human genes and diseases (e.g. *dbSNP*)
- microarray data and other gene expression databases
- proteomics resources (e.g. PeptideAtlas)

Another important database: *Pubmed* (literature from the life sciences)
Biological Databases

Large amounts of biological or biochemical data are usually stored in databases, frequently publicly available.  
*Nucleic Acids Research* 2011: 1330 databases covering various aspects of molecular and cell biology

**nonexhaustive list of DB categories**

- nucleotide sequence databases (*GenBank*, European Nucleotide Archive, TRANSFAC)
- RNA sequence databases (HIV sequence database (LANL), Ribosomal Database Project (RDP))
- protein sequence databases (*UniProt*, *RefSeq*, SwissProt, Blocks, InterPro, *Pfam*, CyMoBase, COG)
- structure databases (*Protein Data Bank* (PDB), SCOP)
- genomics databases (z.B. *UCSC Genome Browser*, FlyBase, TAIR, Wormbase)
- metabolic and signaling pathways (KEGG)
- human genes and diseases (e.g. *dbSNP*)
- microarray data and other gene expression databases
- proteomics resources (e.g. PeptideAtlas)

Another important database: **Pubmed** (literature from the life sciences)
Example: Protein Data Bank

PDB entry

| HEADER | HYDROLASE                          | 29-AUG-01 1JVA          |
| TITLE  | CRYSTAL STRUCTURE OF THE VMA1-DERIVED ENDONUCLEASE BEARING |                      |
| TITLE  | 2 THE N AND C EXTEIN PROPEPTIDES  |                      |
| ...    |                                   |                      |
| ATOM   | 1  N  VAL A 282                   | 53.188 37.546 63.656 | 1.00 30.60 N |
| ATOM   | 2  CA VAL A 282                   | 51.994 38.322 63.214 | 1.00 29.81 C |
| ATOM   | 3  C  VAL A 282                   | 52.214 38.902 61.809 | 1.00 29.27 C |
| ATOM   | 4  O  VAL A 282                   | 53.043 39.793 61.611 | 1.00 29.35 O |
| ATOM   | 5  CB VAL A 282                   | 51.693 39.476 64.196 | 1.00 30.89 C |
| ATOM   | 6  CG1 VAL A 282                  | 52.886 40.414 64.283 | 1.00 32.85 C |
| ATOM   | 7  CG2 VAL A 282                  | 50.461 40.242 63.749 | 1.00 31.38 C |
| ATOM   | 8  N  GLY A 283                   | 51.467 38.386 60.838 | 1.00 26.91 N |
| ATOM   | 9  CA GLY A 283                   | 51.596 38.879 59.483 | 1.00 25.13 C |
| ATOM   | 10 C  GLY A 283                   | 51.224 40.346 59.406 | 1.00 24.20 C |
| ATOM   | 11 O  GLY A 283                   | 50.311 40.802 60.098 | 1.00 22.08 O |
| ...    |                                   |                      |

3-dimensional Cartesian coordinates \((x, y, z)\) of each atom are given in Angstrom units (Å).
Example: Protein Data Bank

<table>
<thead>
<tr>
<th>PDB entry</th>
</tr>
</thead>
<tbody>
<tr>
<td>HEADER    HYDROLASE 29-AUG-01 1JVA</td>
</tr>
<tr>
<td>TITLE     CRYSTAL STRUCTURE OF THE VMA1-DERIVED ENDONUCLEASE BEARING</td>
</tr>
<tr>
<td>TITLE     2 THE N AND C EXTEN PROPEPTIDES</td>
</tr>
<tr>
<td>ATOM 1 N VAL A 282 53.188 37.546 63.656 1.00 30.60 N</td>
</tr>
<tr>
<td>ATOM 2 CA VAL A 282 51.994 38.322 63.214 1.00 29.81 C</td>
</tr>
<tr>
<td>ATOM 3 C VAL A 282 52.214 38.902 61.809 1.00 29.27 C</td>
</tr>
<tr>
<td>ATOM 4 O VAL A 282 53.043 39.793 61.611 1.00 29.35 O</td>
</tr>
<tr>
<td>ATOM 5 CB VAL A 282 51.693 39.476 64.196 1.00 30.89 C</td>
</tr>
<tr>
<td>ATOM 6 CG1 VAL A 282 52.886 40.414 64.283 1.00 32.85 C</td>
</tr>
<tr>
<td>ATOM 7 CG2 VAL A 282 50.461 40.242 63.749 1.00 31.38 C</td>
</tr>
<tr>
<td>ATOM 8 N GLY A 283 51.467 38.386 60.838 1.00 26.91 N</td>
</tr>
<tr>
<td>ATOM 9 CA GLY A 283 51.596 38.879 59.483 1.00 25.13 C</td>
</tr>
<tr>
<td>ATOM 10 C GLY A 283 51.224 40.346 59.406 1.00 24.20 C</td>
</tr>
<tr>
<td>ATOM 11 O GLY A 283 50.311 40.802 60.098 1.00 22.08 O</td>
</tr>
</tbody>
</table>

3-dimensional Cartesian coordinates \((x, y, z)\) of each atom are given in Angstrom units (Å).
Example: FASTA

FASTA file format

- plain text
- header: “>” followed by name of sequence, optional additional info
- sequence (nucleotide or protein) using line breaks
- multiple sequences are concatenated: header1, sequence1, header2, sequence2, ...

Example (fasta file, suffix .fa)

>gi|129270072|ref|NM_001034993.2| Danio rerio Indian hedgehog homolog a (ihha), mRNA
TAAATAATGGACTTGCCGCTCCTCCTCAATTTGGGTATTTCTCGAGAGACCAGCGTTTCGCGTTTTA
TGGGCAACTGAACATTTGGATACAGCTTTCTCGGCGGCTCTCCATGACTGGCTTTGCGTAAAGC
GAATTTTACCAGGACGGATGGCAGGATAAGGCTAATAAAGAAGTAGAGGTTTTTGTCCTCCCTCCCTCCATT
TACCCCGTGGGGGGAAGAAAGGATCTCCTGCGGGGGGGATAGTTTCTCGAGAGGCTGACACCCCGTAGCTCGG
CCAGCGGAGAATCGAGGGCAAGATCACCAGGAACTCAGAGCGCTTTAAAGAGCTCACGCCGAACTACAAC
CCGGACATCATCTCTTTTCCACTTGTGCGGATCGACTCATGACGCAGCGCTGTAAAGATA
...
Example: GenBank

Genbank file format

- header with meta info (name, publications, authors, organism, ...)
- features (e.g. genes) and their sequence coordinates
- sequence and coordinates (human readable)

Example (genbank file, suffix .gb)

<table>
<thead>
<tr>
<th>LOCUS</th>
<th>NM_001034993</th>
<th>1616 bp</th>
<th>mRNA</th>
<th>linear</th>
<th>VRT 04-NOV-2010</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEFINITION</td>
<td>Danio rerio Indian hedgehog homolog a (ihha), mRNA.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SOURCE</td>
<td>Danio rerio (zebrafish)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>REFERENCE</td>
<td>1 (bases 1 to 1616)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TITLE</td>
<td>UDP xylose synthase 1 is required for morphogenesis and histogenesis of the craniofacial skeleton</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FEATURES</td>
<td>Location/Qualifiers</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>source</td>
<td>1..1616</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gene</td>
<td>1..1616</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>/gene=&quot;ihha&quot;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CDS</td>
<td>318..1559</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ORIGIN</td>
<td>1 taaataaatg gacctgccccg ctcgctctca tttgggtgat ttctcggaga gaccagcgtt</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>61 tcgcgtttta tcgggactga acatgggtga taactggggtc tttgctccgg cggctctcca</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>121 tgactggtct tgctgtaagc gaattttcag cacggacgat ggcaacggata aggcttaataa</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>181 agaagtagag gtttttgtct ccccatcatt tacccggctg gggggaagaa ggatctcgcg</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>241 gggggcgagttt ttcagtgaa cggcgagcgtg tgcccaccgta ccactcggca tcggagtcg</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
//
## Sequence Databases

<table>
<thead>
<tr>
<th>Database Name</th>
<th>Details</th>
</tr>
</thead>
</table>
  - ESTs (single-pass cDNA sequences, or Expressed Sequence Tags)  
    67,122,842 sequence entries (Nov 2010)  
  - Sequence Read Archive (SRA)  
    raw sequencing data from the "next" generation of sequencing platforms  
    0.8 Pb (=800,000 Gb) (Nov 2010) |
| **EMBL (European Molecular Biology Laboratory)**    | - European Nucleotide Archive (ENA)  
  - EMBL-Bank  
    (Release 105, 27-AUG-2010 contains 195,241,608 sequences comprising  
    292,078,866,691 nucleotides) |
| **secondary databases of genome projects**          | - GOLD (Genomes OnLine Database)  
  - diArk |


### Availability of Data

#### How to retrieve/query the data

- **web interface/search:**
  - keyword search
  - bulk retrieval
  - **homology search**

- download of DB dump (FTP), often flat file format

#### Accession number

- database identifier
- used in literature

**Example:** NM_001034993
Issues with Biological Databases

Problems of DBs

- frequently unreliable because of errors:
  - incorrectly annotated genes (structural)
  - incorrectly annotated genes (functional)
  - missing genes
  - sequencing errors, misassemblies, misalignments

- propagation of errors:
  a falsely annotated gene may lead to errors in other genes (through similarity searching)

- redundancy
  - often very many almost identical database hits
  - effort do clean up redundancies: *nr* subset of proteins
Problem Definition

Homology Search

Given a database $D$ of sequences and a query sequence $S$, find all sequences in $D$ that are homologs of $S$.

Remarks

- relevant for both protein and nucleotide sequences
- homology is usually (but not necessarily) decided on the basis of sequence similarity as measured by a local pairwise alignment
- retrieving the actual alignment is often a required part of the solution
- want to solve homology search problem for many sequences $S_1, S_2, \ldots$ and the same database $D$
DB homology search can be considered to be a task of binary classification: Each item \( T \) in \( D \) is classified/predicted as either homologue (positive) or non-homologue (negative).

**Definition (true positive)**

A true positive (TP) is an item that is positive and predicted as positive.

**Definition (false positive)**

A false positive (FP) is an item that is negative but that is predicted as positive.

**Definition (false negative)**

A false negative (FN) is an item that is positive but predicted as negative.

**Definition (true negative)**

A true negative (TN) is an item that is negative and predicted as negative.
### Sensitivity, Specificity

#### Definition (Sensitivity)

The **sensitivity** of a prediction method is defined as

\[
Sn = \frac{\#TP}{\#TP + \#FN}.
\]

\#TP + \#FN is the number of *actual* positives.

#### Definition (Specificity)

The **specificity** of a prediction method is defined as

\[
Sp = \frac{\#TP}{\#TP + \#FP}.
\]

\#TP + \#FN is the number of *predicted* positives. 

(Warning: Sometimes variants with a different meaning are used)
**Sensitivity, Specificity**

**Definition (Sensitivity)**

The sensitivity of a prediction method is defined as

\[
Sn = \frac{\#TP}{\#TP + \#FN}.
\]

\(\#TP + \#FN\) is the number of actual positives.

**Definition (Specificity)**

The specificity of a prediction method is defined as

\[
Sp = \frac{\#TP}{\#TP + \#FP}.
\]

\(\#TP + \#FN\) is the number of predicted positives.

(Warning: Sometimes variants with a different meaning are used)
### Sensitivity, Specificity

**Example**

<table>
<thead>
<tr>
<th>truth</th>
<th>prediction</th>
<th>number</th>
</tr>
</thead>
<tbody>
<tr>
<td>negative</td>
<td>negative</td>
<td>1000</td>
</tr>
<tr>
<td>negative</td>
<td>positive</td>
<td>20</td>
</tr>
<tr>
<td>positive</td>
<td>negative</td>
<td>10</td>
</tr>
<tr>
<td>positive</td>
<td>positive</td>
<td>30</td>
</tr>
</tbody>
</table>

Sensitivity = $Sn = \frac{30}{40} = 70\%$

Specificity = $Sp = \frac{30}{50} = 60\%$
Assessing Significance of Hits

Example

LPPGVKFTITSTMIQLLNLKGMFRGASGDDANQHLMNFVAICKSQEIMPPTEQLVIRPHIVPLL---PTFHGMESENPYVHIKEFEEVCNTFRE

Motivation

What degree of similarity do we require to assume that similarity means homology?

Observations:

- want to use alignment score as criterion: the higher the score the more likely is the homology hypothesis
- significance must depend on database size: the larger the database the more likely is some random hit of a given score
- decision must also depend on query (sequence composition)
Assessing Significance of Hits

Example

LPPGVKFTITSTMIQLLNLKGMFRGASGDDANQHLMNFVAICKSQEI
MPPETQLVIRPHIVPLL---PTFHGMEMSENPVYVHIKEFEEVCNTFRE

Motivation

What degree of similarity do we require to assume that similarity means homology?

Observations:

- want to use alignment score as criterion: the higher the score the more likely is the homology hypothesis
- significance must depend on database size: the larger the database the more likely is some random hit of a given score
- decision must also depend on query (sequence composition)
Assessing Significance of Hits

Example

LPPGVKFTITSTMIQLNLKGMFRGASGDDANQHLMNFVAICKSQEI
MPPTEQLVIRPHIVPPL---PTFHGMESENPYVHIKEFEEvCNTFRE

Motivation

What degree of similarity do we require to assume that similarity means homology?

Observations:

- want to use alignment score as criterion: the higher the score the more likely is the homology hypothesis
- significance must depend on database size: the larger the database the more likely is some random hit of a given score
- decision must also depend on query (sequence composition)
Assessing Significance of Hits

Example

LPPGVKFTITSTMIQLLNLKGMFRGASGDDANQHLMNFVAICKSQEIMPPTEQLVIRPHIVPLL---PTFHGMesenpyvhikehfeevcntfren

Motivation

What degree of similarity do we require to assume that similarity means homology?

Observations:

- want to use alignment score as criterion: the higher the score the more likely is the homology hypothesis
- significance must depend on database size: the larger the database the more likely is some random hit of a given score
- decision must also depend on query (sequence composition)
Assessing Significance of Hits

Let $S$ be the query in a database search against database $D$ and let $h$ be a hit with score $s$.

**Definition (E-value)**

The E-value of $h$ is the expected number of hits with score at least $s$ when performing the same search with a random query sequence $S'$ with the same length as $S$ and in a database $D'$ with random sequences of the same size as $D$.

**Definition (P-value)**

The P-value of $h$ is the probability of a hit with score at least $s$ when performing the same search with a random query sequence $S'$ with the same length as $S$ and in a database $D'$ with random sequences of the same size as $D$.

Note, that the E- and P-value depend on and require a random model of sequences.
Assessing Significance of Hits

Let $S$ be the query in a database search against database $D$ and let $h$ be a hit with score $s$.

**Definition (E-value)**

The E-value of $h$ is the expected number of hits with score at least $s$ when performing the same search with a random query sequence $S'$ with the same length as $S$ and in a database $D'$ with random sequences of the same size as $D$.

**Definition (P-value)**

The P-value of $h$ is the probability of a hit with score at least $s$ when performing the same search with a random query sequence $S'$ with the same length as $S$ and in a database $D'$ with random sequences of the same size as $D$.

Note, that the E- and P-value depend on and require a random model of sequences.
Assessing Significance of Hits

Let $S$ be the query in a database search against database $D$ and let $h$ be a hit with score $s$.

**Definition (E-value)**

The E-value of $h$ is the expected number of hits with score at least $s$ when performing the same search with a random query sequence $S'$ with the same length as $S$ and in a database $D'$ with random sequences of the same size as $D$.

**Definition (P-value)**

The P-value of $h$ is the probability of a hit with score at least $s$ when performing the same search with a random query sequence $S'$ with the same length as $S$ and in a database $D'$ with random sequences of the same size as $D$.

Note, that the E- and P-value depend on and require a random model of sequences.
Tradeoff Between Sensitivity, Specificity and Speed

Tradeoff between sensitivity and specificity

- usually, sensitivity and specificity can be traded off against each other: e.g. by increasing a threshold of minimal alignment scores, we increase specificity at the expense of lowering sensitivity.
- typical curve Sn-versus-Sp (*chalk board*)
- optimal choice depends on application (user has to assess individually the risks for the different error kinds (FP, FN))

Tradeoff between speed, sensitivity and specificity

- in complex search problems, often an increase of speed can be achieved at the cost of a lower sensitivity and/or specificity
- ideas: neglect improbable cases / interrupt search if not promising
- shift of curve Sn-versus-Sp (*chalk board*)
### Tradeoff between sensitivity and specificity

- usually, sensitivity and specificity can be traded off against each other: e.g. by increasing a **threshold** of minimal alignment scores, we increase specificity at the expense of lowering sensitivity.
- typical curve $S_n$-versus-$S_p$ (*chalk board*)
- optimal choice depends on application (user has to assess individually the risks for the different error kinds (FP, FN)

### Tradeoff between speed, sensitivity and specificity

- in complex search problems, often an increase of speed can be achieved at the cost of a lower sensitivity and/or specificity
- ideas: neglect improbable cases / interrupt search if not promising
- shift of curve $S_n$-versus-$S_p$ (*chalk board*)
Exact Alignments Are Too Slow

DP too slow

Database searching using the Smith-Waterman algorithm for finding the optimal local alignment between query $S$ and all database sequences $T$ is too slow: $\Theta(mn)$

$n = |S|$ query sequence length
$m = \text{total database size (sum of lengths of all sequences in db)}$

Heuristics for speedup

- preprocessing
  - either query $S$ or database $D$ is preprocessed, so that the actual search can be done faster
- use several stages for search
  - fast initial stage (speed vs sensitivity)
  - slower final stage (specificity)
if $S$ and $T$ do not look similar enough using a fast check: premature interruption of comparison, use next $T$
Exact Alignments Are Too Slow

**DP too slow**

Database searching using the Smith-Waterman algorithm for finding the optimal local alignment between query $S$ and all database sequences $T$ is too slow: $\Theta(mn)$

$n = |S|$ query sequence length

$m = \text{total database size (sum of lengths of all sequences in db)}$

**Heuristics for speedup**

- **preprocessing**
  
  either query $S$ or database $D$ is preprocessed, so that the actual search can be done faster

- **use several stages for search**
  
  fast initial stage (speed vs sensitivity)
  
  slower final stage (specificity)
  
  if $S$ and $T$ do not look similar enough using a fast check: premature interruption of comparison, use next $T$
### Aho-Corasick Algorithm

#### Definition (exact string matching)

String $P$ is said to occur at position $i$ of string $T$ if $P[j] = T[j + i - 1]$ for all $j = 1, \ldots, |P|$. $P$ is also called a pattern.

#### Definition (exact set matching problem)

Find all occurrences in string $T$ of any pattern in a set of patterns $\{P_1, \ldots, P_z\}$.

#### Aho-Corasick algorithm

There is an algorithm developed by Aho and Corasick, that solves the exact set matching problem in time $O(n + m + k)$, where $m = |T|$ is the length of $T$ and $n = |P_1| + \cdots + |P_z|$ is the total length of all patterns and $k$ is the number of occurrences of patterns (solution set size).
Aho-Corasick Algorithm

Definition (exact string matching)

String $P$ is said to occur at position $i$ of string $T$ if $P[j] = T[j + i - 1]$ for all $j = 1, \ldots, |P|$. $P$ is also called a pattern.

Definition (exact set matching problem)

Find all occurrences in string $T$ of any pattern in a set of patterns $\{P_1, \ldots, P_z\}$.

Aho-Corasick algorithm

There is an algorithm developed by Aho and Corasick, that solves the exact set matching problem in time $O(n + m + k)$, where $m = |T|$ is the length of $T$, $n = |P_1| + \cdots + |P_z|$ is the total length of all patterns and $k$ is the number of occurrences of patterns (solution set size).
Aho-Corasick Algorithm

**Definition (exact string matching)**

String $P$ is said to occur at position $i$ of string $T$ if $P[j] = T[j + i - 1]$ for all $j = 1, \ldots, |P|$. $P$ is also called a pattern.

**Definition (exact set matching problem)**

Find all occurrences in string $T$ of any pattern in a set of patterns $\{P_1, \ldots, P_z\}$.

**Aho-Corasick algorithm**

There is an algorithm developed by Aho and Corasick, that solves the exact set matching problem in time $O(n + m + k)$, where $m = |T|$ is the length of $T$, $n = |P_1| + \cdots + |P_z|$ is the total length of all patterns and $k$ is the number of occurrences of patterns (solution set size).
The BLAST Program

The Basic Local Alignment Search Tool

- released in 1990
- still the single most used Bioinformatics tool
- most highly cited paper published in the 1990s
  (~8000 citations of original paper in PubMed alone)
- contributions by: Lipman, Gish, Miller, Myers, Karlin, Altschul, Dembo
- written in C
The BLAST Heuristics

Main ideas for speedup

- If \( S \) and \( T \) are sufficiently similar then it is very likely that some high-scoring **gapless** alignments exists.

- High-scoring gapless alignments are very likely to contain high-scoring **gapless alignments of fixed size**.

- Use **exact string matching** because it can be done fast.
### The BLAST Heuristics

#### Outline of the BLAST-algorithm

1. remove low-complexity sequence regions from $S$
2. compile list of all $k$-mers $\{P_1, \ldots, P_z\}$ that score above a threshold $t$ when matched against $S$
3. find all occurrences of a $k$-mer in $\{P_1, \ldots, P_z\}$ in the database using an algorithm similar to the Aho-Corasick algorithm
4. extend the exact matches to high-scoring segment pairs (HSP)
5. list all of the HSPs in the database whose score is above another threshold
6. evaluate the significance of the HSP score
7. join two or more HSP regions into a longer alignment
8. use Smith-Waterman to locally align the query with each of the matched database sequences
9. report the matches whose E-value is below a threshold
BLAST Step 2

Compilation of high-scoring $k$-mers

Find all $k$-mers $P \in \Sigma^k$ such that a substring $U$ of $S$ of length $k$ exists, such that $\sum_{i=1}^{k} s(P[i], U[i]) \geq t$.

Here, $s$ is a scoring matrix (for nucleotides or amino acids).

Example

$S=WAILCYV$, $k = 3$, BLOSUM-62 scoring matrix, $t = 16$
BLAST Step 2

**Compilation of high-scoring \( k \)-mers**

Find all \( k \)-mers \( P \in \Sigma^k \) such that a substring \( U \) of \( S \) of length \( k \) exists, such that \( \sum_{i=1}^{k} s(P[i], U[i]) \geq t \). Here, \( s \) is a scoring matrix (for nucleotides or amino acids).

**Example**

\( S = \text{WAILCYV}, \ k = 3, \ \text{BLOSUM-62 scoring matrix}, \ t = 16 \)
**BLAST Step 4**

**Definition (HSP)**

Given two strings $S$ and $T$, a segment pair is a pair of equal length substrings of $S$ and $T$ aligned without gaps. A locally maximal segment pair is a segment pair whose score would not increase if it were extended or shortened on either side by 1. Given a threshold a high-scoring segment pair is a segment pair with score above the threshold.

**BLAST Step 4**

The HSPs constructed by BLAST are locally maximal high-scoring segment pairs.

**Example**

<table>
<thead>
<tr>
<th>unaligned</th>
<th>aligned</th>
<th>unaligned</th>
</tr>
</thead>
<tbody>
<tr>
<td>...DRPLLDKFL</td>
<td>FWAILCEV</td>
<td>KTNAVA...</td>
</tr>
<tr>
<td>...PCSDTAQNA</td>
<td>WWGIAYV</td>
<td>DDLKTH...</td>
</tr>
</tbody>
</table>
Joining of words on the “same diagonal”

Sometimes two or more matching $k$-mers have the same distance in both query and target. They may be joined to a single, larger HSP.

Example

*(chalk board)*
BLAST Step 6: Significance of the HSP score

Assumptions

- sequence characters are independent from each other
- sequence characters are identically distributed

Significance of the HSP score

For good hits and large $m$ and $n$, the E-value can be approximated with an extreme value distribution:

$$\text{E-value} \approx K m n e^{-\lambda s}$$

Here, $s$ is the score of the hit and the constants $\lambda > 0$ and $K > 0$ depend upon the substitution matrix and character frequencies.
BLAST Step 8

Alignment with gaps

- use HSP as “seed” of alignment, i.e. find a local alignment that includes the HSP as sub-alignment

- another heuristic: “\( \text{xdrop} \)"
  instead of filling in the DP table up to the sequence ends, stop evaluating DP cells beyond points where the value drops by more than \( \text{xdrop} \) below the currently maximal value

Example

\((\text{chalk board})\)
Alignment with gaps

- use HSP as “seed” of alignment, i.e. find a local alignment that includes the HSP as sub-alignment
- another heuristic: “xdrop” instead of filling in the DP table up to the sequence ends, stop evaluating DP cells beyond points where the value drops by more than xdrop below the currently maximal value

Example

(chalk board)
### Example (BLAST result between a tomato query protein and a rice target protein)

```plaintext
>gb|EEC66694.1| hypothetical protein OsI_33012 [Oryza sativa Indica Group]
Length=259

Score = 55.1 bits (131), Expect = 2e-05, Method: Compositional matrix adjust.
Identities = 38/145 (27%), Positives = 69/145 (48%), Gaps = 16/145 (11%)

<table>
<thead>
<tr>
<th>Query 124</th>
<th>HEYE--DFALCGIPFTWWDARDKDCIFKRIDRVVPNQDLQDMLGNLEIMHLRDRTGSDHA 181</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sbjct 9</td>
<td>HNYQLFDQFGKFGKSHTYDNKCREGRNNVKNVLRDLAADDNNWRNFFTQVTHLISPCSDHC 68</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Query 182</th>
<th>PLLLV--TGSLAQKFNSKLVRLFKFWTENDDFMEMVNT---QKNGARSVAPINEPTGEAI 236</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sbjct 69</td>
<td>PILKFDTESQSQHKTCKLRYEIFWREAAALQEVINSWEDSGKGKQNLGDIIKALKVGM- 127</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Query 237</th>
<th>RGIGRGRGVRAPAENAVPRE 261</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sbjct 128</td>
<td>--------RALHSWSKAKCINVGRE 144</td>
</tr>
</tbody>
</table>
```
**Six-Frame Translation**

**Six-frame translation**

There are 6 different ways how a nucleotide sequence can be translated to a protein sequence in one piece. The term **six-frame translation** refers to this (hypothetical) set of 6 protein sequences defined by a nucleotide sequence.

**Example (6-frame translation)**

<table>
<thead>
<tr>
<th>3-frame translation (forward)</th>
<th>3-frame translation (reverse)</th>
</tr>
</thead>
<tbody>
<tr>
<td>M P L K A A S R Y H S N S K K P</td>
<td>S A M S L R L C T D N S S S S F A</td>
</tr>
<tr>
<td>R C H * K P Q A G I I R T R K S</td>
<td>I G N F A A L L L Y * E F E F F F G</td>
</tr>
<tr>
<td>R D A I E S R K Q V S F E L E K A</td>
<td>R H W Q F G C A P I M R V R F L W</td>
</tr>
</tbody>
</table>
## Variants of BLAST

<table>
<thead>
<tr>
<th>Variant</th>
<th>Query</th>
<th>Database (&quot;target&quot;)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLASTN</td>
<td>nucleotide</td>
<td>nucleotide</td>
</tr>
<tr>
<td>BLASTP</td>
<td>protein</td>
<td>protein</td>
</tr>
<tr>
<td>BLASTX</td>
<td>nucleotide (6-frame tr.)</td>
<td>protein</td>
</tr>
<tr>
<td>TBLASTN</td>
<td>protein</td>
<td>nucleotide (6-frame tr.)</td>
</tr>
<tr>
<td>TBLASTX</td>
<td>nucleotide (6-frame tr.)</td>
<td>nucleotide (6-frame tr.)</td>
</tr>
</tbody>
</table>