Similarity Searches on Sequence Databases

Lorenza Bordoli
Swiss Institute of Bioinformatics
EMBnet Course, Geneva, February 2006

Outline

• Importance of Similarity

• Heuristic Sequence Alignment:
  – Principle
  – FASTA algorithm
  – BLAST algorithm

• Assessing the significance of sequence alignment
  – Raw score, normalized (bits) score, Extreme Value Distribution, P-value, E-Value

• BLAST:
  – Protein Sequences
  – DNA Sequences
  – Choosing the right Parameters

• Other members of the BLAST family
Importance of Similarity

- Same Sequence
- Same Origin
- Same 3D Fold
- Same Function

Similar sequences: probably have the same ancestor, share the same structure, and have a similar biological function.

Importance of Similarity

sequence DB

unknown function ?

Similarity Search

similar protein with known function

extrapolate

function
Importance of Similarity

**Rule-of-thumb:**
If your sequences are more than 100 amino acids long (or 100 nucleotides long) you can consider them as homologues if 25% of the aa are identical (70% of nucleotide for DNA). Below this value you enter the **twilight zone**.

Twilight zone = protein sequence similarity between ~0-20% identity:
is not statistically significant, i.e. could have arisen by chance.

**Beware:**
- E-value (*Expectation value*)
- Length of the segments similar between the two sequences
- The number of insertions/deletions

Alignment score

**Amino acid substitution matrices**
- Example: PAM250
- Most used: Blosum62

**Raw score of an alignment**

Score = 1 + 6 + 0 + 2 = 9
Insertions and deletions

Gap penalties

- Opening a gap penalizes an alignment score
- Each extension of a gap penalizes the alignment's score
- The gap opening penalty is in general higher than the gap extension penalties (simulating evolutionary behavior)

- The raw score of a gapped alignment is the sum of all amino acid substitutions from which we subtract the gap opening and extension penalties.

Alignment

Alignment types:

- **Global**: Alignment between the complete sequence A and the complete sequence B
- **Local**: Alignment between a sub-sequence of A and a sub-sequence of B

Computer implementation (Algorithms):

Dynamic programming (exact algorithm)

- **Global**: Needleman-Wunsch
- **Local**: Smith-Waterman
Heuristic Sequence Alignment

- With the Dynamic Programming algorithm, one obtains an alignment in a time that is proportional to the product of the lengths of the two sequences being compared. Therefore, when searching a whole database, the computation time grows linearly with the size of the database. With current databases, calculating a full Dynamic Programming alignment for each sequence of the database is too slow (unless implemented in a specialized parallel hardware).

- The number of searches that are presently performed on whole genomes creates a need for faster procedures.

⇒ Two methods that are at least 50-100 times faster than dynamic programming were developed: FASTA and BLAST

Heuristic Sequence Alignment: Principle

- **Dynamic Programming**: computational method that provides in mathematical sense the best alignment between two sequences, given a scoring system.

- **Heuristic Methods** (e.g., BLAST, FASTA) they prune the search space by using fast approximate methods to select the sequences of the database that are likely to be similar to the query and to locate the similarity region inside them.

  = > Restricting the alignment process:
  - Only to the selected sequences
  - Only to some portions of the sequences (search as small a fraction as possible of the cells in the dynamic programming matrix)
Heuristic Sequence Alignment: Principle

• These methods are heuristic; i.e., an empirical method of computer programming in which rules of thumb are used to find solutions.

• They almost always works to find related sequences in a database search but does not have the underlying guarantee of an optimal solution like the dynamic programming algorithm (But good ones often do).

• Advantage: This methods that are least 50-100 times faster than dynamic programming therefore better suited to search databases.

FASTA & BLAST: story

1985 : FASTP (D. Lipman and W. Pearson)
   Global gapped alignments

1988 : FASTA (W. Pearson and D. Lipman)
   Local gapped alignments

1990 : BLAST1
   (S. Altschul, W. Gish, W. Miller, E. Myers, and D. Lipman)
   Local ungapped alignments

**Gapped BLASTs :**

1996: WU–BLAST2 (W. Gish)

1997: NCBI–BLAST2 (and PSI–BLAST)
   (S. Altschul, T. Madden, A. Schaffer, J. Zhang, Z. Zhang, W. Miller and D. Lipman)
FASTA: Algorithm (4 steps)

Localyze the 10 best regions of similarity between the two seq. Each identity between two “word” is represented by a dot.

Each diagonal: ungapped alignment. The smaller the k, the sensitive the method but slower.

Find the best combination of the diagonals -> compute a score. Only those sequences with a score higher than a threshold will go to the fourth step.

DP applied around the best scoring diagonal.

BLAST: Algorithm

1. Blast algorithm: creating a list of similar words

A substitution matrix is used to compute the word scores.
BLAST: Algorithm

2. Blast algorithm: eliminating sequences without word hits

Database sequences

Search for exact matches

List of words matching the query with a score > T

List of sequences containing words similar to the query (hits)

BLAST: Algorithm

Third step:

For each word match (=hit), extend ungapped alignment in both directions. Stop when S decreases by more than X from the highest value reached by S.

Each match is then extended. The extension is stopped as soon as the score decreases more then X when compared with the highest value obtained during the extension process.

HSP = High Scoring Segment Pair

Reports all HSPs having score S above a threshold, or equivalently, having E-value below a threshold.
Each match is then extended. The extension is stopped as soon as the score decreases more than X when compared with the highest value obtained during the extension process.

3. Blast algorithm: extension of hits

Ungapped extension if:
- 2 "Hits" are on the same diagonal but at a distance less than A.

Extension using dynamic programming:
- limited to a restricted region.
BLAST: Algorithm

The «two-hits» requirement

First step: as with BLAST1, generate lists of words scoring more than T with words of the query.

Second step: generation of hits: identify all word matches in DB sequences

Third step: extension of hits: requires a second hit on the same diagonal at a distance of less than A.

Additional step: Gapped extension of the hits slower-> therefore: requirement of a second hits on the diagonal. (hits not joined by ungapped extensions could be part of the same gapped alignment)

Fourth step: gapped extension of HSPs having score above a threshold $S_g$

Assessing the significance of sequence alignment

• Scoring System:
  1. Scoring (Substitution) matrix (or match mismatch for DNA): In proteins some substitutions are more acceptable than others. Substitution matrices give a score for each substitution of one amino-acid by another (e.g. PAM, BLOSUM)
  2. Gap Penalties: simulate as closely as possible the evolutionary mechanisms involved in gap occurrence. Gap opening penalty: Counted each time a gap is opened in an alignment and Gap extension penalty: Counted for each extension of a gap in an alignment.

• Based on a given scoring system: you can calculate the raw score of the alignment
  Raw score = sum of the amino acid substitution scores (or match/mismatch) and gap penalties
Assessing the significance of sequence alignment

Caveats:

1. We need a normalised (bit) score to compare different alignments, based on different scoring systems, e.g. different substitution matrices.

2. A method to assess the statistical significance of the alignment is needed (is an alignment biological relevant?): E-value

Assessing the significance of sequence alignment

• How?

⇒ Evaluate the probability that a score between random or unrelated sequences will reach the score found between two real sequences of interest:

If that probability is very low, the alignment score between the real sequences is significant.

Frequency of aa occurring in nature

Ala 0.1
Val 0.3
Trp 0.01
...

Random sequence 1

Random sequence 2

Real sequence 1

Real sequence 2

If SCORE > SCORE => the alignment between the real sequences is significant
Karlin and Altschul observed that in the framework of local alignments without gaps: the distribution of random sequence alignment scores follow an EVD.

\[
Y = \lambda \exp[-\lambda(x - \mu) - e^{\lambda(x - \mu)}]
\]

\(\mu, \lambda\) : parameters depend on the length and composition of the sequences and on the scoring system

\[
P(S < x) = \exp[-e^{-\lambda(x - \mu)}]
\]
The Extreme Value Distribution

The Extreme Value Distribution

\[ Y = \lambda \exp[-\lambda(x - \mu)] - e^{-\lambda(x - \mu)} \]

\[ P(S < x) = \exp[-e^{-\lambda(x - \mu)}] \]

\[ P(S \geq x) = 1 - \exp[-e^{-\lambda(x - \mu)}] \]

\[ P-value = \text{the probability of obtaining a score equal or greater than } x \text{ by chance} \]
Assessing the significance of sequence alignment

- Local alignment without gaps:
  - Theoretical work: Karlin-Altschul statistics: \( \Rightarrow \) Extreme Value Distribution

- Local alignments with gaps:
  - Empirical studies: \( \Rightarrow \) Extreme Value Distribution

Statistics derived from the scores:

- **P-value**
  \( \Rightarrow \) Probability that an alignment with this score occurs by chance in a database of this size
  \( \Rightarrow \) The closer the P-value is towards 0, the better the alignment

- **E-value**
  \( \Rightarrow \) Number of matches with this score one can expect to find by chance in a database of size \( N \)
  \( \Rightarrow \) The closer the e-value is towards 0, the better the alignment

- Relationship between E-value and P-value:
  \( \Rightarrow \) In a database containing \( N \) sequences
  \[ E = P \times N \]
BLAST

Basic Local Alignment Search Tool

A Blast for each query

Different programs are available according to the type of query

<table>
<thead>
<tr>
<th>Program</th>
<th>Query</th>
<th>Database</th>
</tr>
</thead>
<tbody>
<tr>
<td>blastp</td>
<td>protein</td>
<td>protein</td>
</tr>
<tr>
<td>blastn</td>
<td>nucleotide</td>
<td>nucleotide</td>
</tr>
<tr>
<td>blastx</td>
<td>nucleotide</td>
<td>protein</td>
</tr>
<tr>
<td>tblastn</td>
<td>protein</td>
<td>nucleotide</td>
</tr>
<tr>
<td>tblastx</td>
<td>nucleotide</td>
<td>nucleotide</td>
</tr>
</tbody>
</table>
BLASTing protein sequences

blastp = Compares a protein sequence with a protein database

If you want to find something about the function of your protein, use blastp
to compare your protein with other proteins contained in the databases;
identify common regions between proteins, or collect related proteins (phylogenetic
analysis);

tblastn = Compares a protein sequence with a nucleotide database

If you want to discover new genes encoding proteins (from multiple organisms),
use tblastn to compare your protein with DNA sequences translated into their six
possible reading frames; map a protein to genomic DNA;

Three of the most popular blastp online services:

• NCBI (National Center for Biotechnology Information) server:

• ExPASy server:

• Swiss EMBnet server (European Molecular Biology network):
  http://www.ch.embnet.org/software/bBLAST.html (basic)
  http://www.ch.embnet.org/software/aBLAST.html (advanced)
BLASTing protein sequences: Swiss EMBnet blastp server

Basic BLAST

Usage: Choose the suitable BLAST program and database for your query sequence. Paste your sequence in one of the supported formats into the sequence field below and press the "Run BLAST" button. Don't forget your e-mail address, so that we can send you the results in case of traffic jam... Make sure that the format button (next to the sequence field) shows the correct format. See also our BLAST database description.

Please select the program:

Programs:
- blastp
- blastx
- blastn
- tblastn
- tblastx

Please select the database:

- DNA databases
- Protein databases
- Gapped alignment on/off
- BLAST filter on/off
- Graphic output on/off

Paste your sequence here:

(required for blast[...] programs)

DB:

- EMBL (without HTG and ESTs)
- HTG
- EST+HTG
- Genomes
- Other

E-mail address:

HTML

BLASTing protein sequences: Swiss EMBnet blastp server

Advanced BLAST

Usage: Choose the suitable BLAST program and database for your query sequence. Paste your sequence in one of the supported formats into the sequence field below and press the "Run BLAST" button. Don't forget your e-mail address, so that we can send you the results in case of traffic jam... Make sure that the format button (next to the sequence field) shows the correct format. See also our BLAST database description and the NCBI BLAST help.

Please select the program:

You can make multiple selections.

Please select the database(s):

DNA databases:
- Current release (74)
- Cumulative updates

Protein databases:
- Various
- Proteomes

Other:
- Rf Seq
- Human
- M. tuberculosis
- Yeast (S. cerevisiae)
- E. coli
- Arabidopsis

Other:
- Rf Seq
- Human
- M. tuberculosis
- Yeast (S. cerevisiae)
- E. coli
- Arabidopsis

EMBL
- HTG
- HTG_Biopart
- HTG_Ambiguities
- HTG_Biopart
- Human
- Mouse
- Fugu
- D. melanogaster
- C. elegans
- A. thaliana (Iron TIGF)
- Yeast (S. cerevisiae)

Other:
- Rf Seq
- Human
- M. tuberculosis
- Yeast (S. cerevisiae)
- E. coli
- Arabidopsis
BLASTing protein sequences: Swiss EMBnet blasp server

- Greater choice of databases to search
- Advanced Blast parameter modification

Understanding your BLAST output

1. **Graphic display:**
   shows you where your query is similar to other sequences

2. **Hit list:**
   the name of sequences similar to your query, ranked by similarity

3. **The alignment:**
   every alignment between your query and the reported hits

4. **The parameters:**
   a list of the various parameters used for the search
Understanding your BLAST output: 1. Graphic display

The display can help you see that some matches do not extend over the entire length of your sequence => useful tool to discover domains.

Understanding your BLAST output: 2. Hit list

- **Sequence ac number and name**: Hyperlink to the database entry: useful annotations
- **Description**: better to check the full annotation
- **Bit score (normalized score)**: A measure of the similarity between the two sequences: the higher the better (matches below 50 bits are very unreliable)
- **E-value**: The lower the E-value, the better. Sequences identical to the query have an E-value of 0. Matches above 0.001 are often close to the twilight zone. As a rule-of-thumb an E-value above 10^{-4} (0.0001) is not necessarily interesting. If you want to be certain of the homology, your E-value must be lower than 10^{-4}
Understanding your BLAST output: 3. Alignment

A good alignment should not contain too many gaps and should have a few patches of high similarity, rather than isolated identical residues spread here and there.

Percent identity
25% is good news

Length of the alignment

Percent identity

XXX: low complexity regions masked

Mismatch

Identical aa

Similar aa

Understanding your BLAST output: 4. Parameters

Database: swarm_nr
Posted date: Jan 13, 2002 0:19 AM
Number of entries in database: 59,057,040
Number of sequences in database: 1,616,155

Database: swarm_vespillae_pir
Posted date: Jan 13, 2002 0:19 AM
Number of entries in database: 2,261,893
Number of sequences in database: 575,184

Length of query: 149
Length of database: 40,576,601
Effective length query: 119
Effective length of database: 38,054,160

Search details (at the bottom of the results)

- Size of the database searched
- Scoring system parameters
- Details about the number of hits found
A Blast for each query

Different programs are available according to the type of query

<table>
<thead>
<tr>
<th>Program</th>
<th>Query</th>
<th>Database</th>
</tr>
</thead>
<tbody>
<tr>
<td>blastp</td>
<td>protein</td>
<td>protein</td>
</tr>
<tr>
<td>blastn</td>
<td>nucleotide</td>
<td>nucleotide</td>
</tr>
<tr>
<td>blastx</td>
<td>nucleotide</td>
<td>protein</td>
</tr>
<tr>
<td>tblastn</td>
<td>nucleotide</td>
<td>protein</td>
</tr>
<tr>
<td>tblastx</td>
<td>nucleotide</td>
<td>nucleotide</td>
</tr>
</tbody>
</table>

BLASTing DNA sequences

- BLASTing DNA requires operations similar to BLASTing proteins BUT does not always work so well.

- It is faster and more accurate to BLAST proteins (blastp) rather than nucleotides. If you know the reading frame in your sequence, you’re better off translating the sequence and BLASTing with a protein sequence.

- Otherwise:

<table>
<thead>
<tr>
<th>Different BLAST Programs Available for DNA Sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Program</strong></td>
</tr>
<tr>
<td>-------------</td>
</tr>
<tr>
<td>blastn</td>
</tr>
<tr>
<td>tblastx</td>
</tr>
<tr>
<td>blastx</td>
</tr>
</tbody>
</table>

T= translated
BLASTing DNA sequences

**blastn** = Compares a DNA sequence with a DNA database;

Mapping oligonucleotides, cDNAs, and PCR products to a genome; annotating genomic DNA; screening repetitive elements; cross-species sequence exploration;

**tblastx** = Compares a DNA translated into protein with a DNA database translated into protein;

Cross-species gene prediction at the genome or transcript level (ESTs); searching for genes not yet in protein databases;

**blastx** = Compares a DNA translated into protein with a protein sequence database;

Finding protein-coding genes in genomic cDNA; determining if a cDNA corresponds to a known protein;

---

**BLASTing DNA sequences: choosing the right BLAST**

<table>
<thead>
<tr>
<th>Question</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Am I interested in non-coding DNA?</td>
<td>Yes: Use blastn. Never forget that blastn is only for closely related DNA sequences (more than 70% identical)</td>
</tr>
<tr>
<td>Do I want to discover new proteins?</td>
<td>Yes: Use tblastx.</td>
</tr>
<tr>
<td>Do I want to discover proteins encoded in my query DNA sequence?</td>
<td>Yes: Use blastx.</td>
</tr>
<tr>
<td>Am I unsure of the quality of my DNA?</td>
<td>Yes: Use blastx if you suspect your DNA sequence is the coding for a protein but it may contain sequencing errors.</td>
</tr>
</tbody>
</table>

- **Pick the right database:** choose the database that’s compatible with the BLAST program you want to use (in general)
- **Restrict your search:** Database searches on DNA are slower. When possible, restrict your search to the subset of the database that you’re interested in (e.g. only the Drosophila genome)
- **Shop around:** Find the BLAST server containing the database that you’re interested in
- **Use filtering:** Genomic sequences are full of repetitions: use some filtering
BLASTting DNA: BLASTN output

- DNA double-stranded molecule => genes may occur on either strand
- **plus** strand (the query sequence), **minus** strand (reverse complement)
- If the similarity between query and subject is on the same strand: **plus/plus**
- If the minus strand of the query sequence is similar to a database sequence: **plus/minus** with the subject sequence in reverse coordinates (flipped)

\[
\text{Score} = 87.7 \text{ bits (44)}, \text{ Expect} = 2e-15 \text{ Identities} = 57/60 (95\%), \text{ Gaps} = 1/60 (1%)
\]
\[
\text{Strand} = \text{Plus / Plus}
\]
\[
\text{Query:} \quad 1 \quad \text{ggtggtttagaacgatctggtcttaccctgctaccaactgttcatcggttattgttggag} \quad 60
\]
\[
\text{Sbjct:} \quad 96694 \quad \text{ggtgttttagaacgat-tggtcttaccctgctaccaactgttcatcggttattgttggag} \quad 96752
\]

\[
\text{Score} = 52.0 \text{ bits (26)}, \text{ Expect} = 1e-04 \text{ Identities} = 26/26 (100%)
\]
\[
\text{Strand} = \text{Plus / Minus}
\]
\[
\text{Query:} \quad 18 \quad \text{tggtcttaccctgctaccaactgttc} \quad 43
\]
\[
\text{Sbjct:} \quad 40758 \quad \text{tggtcttaccctgctaccaactgttc} \quad 40733
\]

BLASTting DNA: BLASTX output

- Query sequence: translated in the **3 reading frames**, on both **plus** and **minus** strand: +1,+2,+3 (plus strand) and -1, -2, -3 (minus strand)
- Matches on the plus strand: +1,+2,+3
- Matches on the minus strand: query coordinates are inverted

\[
\text{Score} = 790 \text{ bits (2040)}, \text{ Expect} = 0.0 \text{ Identities} = 520/1381 (37\%), \text{ Positives} = 745/1381 (53\%), \text{ Gaps} = 36/1381 (2\%)
\]
\[
\text{Frame} = +3
\]
\[
\text{Query:} \quad 156 \quad \text{SEMNVNMKYQLPNFTAETPIQNVVLHKHH--IYLGA} \quad 329
\]
\[
\text{S} \quad \text{+} \quad \text{N} \quad ++ \quad \text{Y} \quad +\text{P F A PIQN+V} \quad + \quad +\text{Y+} \quad + \quad \text{N I} \quad +\text{N + L+KV E +TGPV}
\]
\[
\text{Sbjct:} \quad 31 \quad \text{SPVNFSVYTMPFPFQAGGPIQNIVNNNSFYQEVYVASQNVIEAVN-QSLEKVWELRTGPV- 88}
\]

\[
\text{Score} = 64.5 \text{ bits (169)}, \text{ Expect} = 1.7e-258 \text{ Identities} = 30/34 (88\%), \text{ Positives} = 34/34 (100\%), \text{ Gaps} = 3/34 (2\%)
\]
\[
\text{Frame} = -1
\]
\[
\text{Query:} \quad 1071 \quad \text{SEMNVNMKYQLPNFTAETPIQNVVLHKHH--IYLGA} \quad 970
\]
\[
\text{S} \quad \text{+} \quad \text{N} \quad ++ \quad \text{Y} \quad +\text{P F A PIQN+V} \quad + \quad +\text{Y+} \quad + \quad \text{N I} \quad +\text{N + L+KV E +TGPV}
\]
\[
\text{Sbjct:} \quad 722 \quad \text{SPVNFSVYTMPFPFQAGGPIQNIVNNNSFYQEVYVASQNVIEAVN-QSLEKVWELRTGPV- 755}
\]
BLASTting DNA: TBLASTN output

- Alignments similar to BLASTX, except that the database and query are exchanged (e.g. on minus strand the database sequence has flipped coordinates)

Score = 47.8 bits (112), Expect = 5e-04
Identities = 20/21 (95%), Positives = 21/21 (99%)
Frame: +2
Query: 1 SQITRIPLNGLGCEHFQSCSQ 21
Sbjct: 108872 SQITKIPLNGLGCEHFQSCSQ 108934

Score = 45.8 bits (107), Expect = 0.002
Identities = 19/21 (90%), Positives = 20/21 (94%)
Frame: -2
Query: 1 SQITRIPLNGLGCEHFQSCSQ 21
Sbjct: 28239 SQITKIPLNGLGCEHFQSCSQ 28177

BLASTting DNA: TBLASTX output

- Both query and database have strand and frame
- Alignments may have any combination of frames

Score = 790 bits (2040), Expect = 0.0
Identities = 520/1381 (37%), Positives = 745/1381 (53%), Gaps = 36/1381 (2%)
Frame: +3/+3
Query: 156 SEMNVNMKYQLPNFTEPQIQQVVLHKKH--IYLGA VNYIYVLNDKLQVAEQYKTPVVL 329
S +N ++ Y +P F A PIQN+V + + +Y+ + N I +N +L+KV E +TGPV
Sbjct: 31 SVPNFSVVTMFPPQAGGPIQNIVNNSFYQEVYVASQNEAVN-QSLEKVWELERTGPV- 88

Score = 64.5 bits (169), Expect = 1.7e-258
Identities = 30/34 (88%), Positives = 34/34 (100%), Gaps = 3/34 (2%)
Frame: -1/+2
Query: 1071 SEMNVNMKYQLPNFTEPQIQQVVLHKKH--IYLGA VNYIYVLNDKLQVAEQYKTPVVL 970
S +N ++ Y +P F A PIQN+V + + +Y+ + N I +N +L+KV E +TGPV
Sbjct: 722 SVPNFSVVTMFPPQAGGPIQNIVNNSFYQEVYVASQNEAVN-QSLEKVWELERTGPV- 755
Choosing the right Parameters

- The **default** parameters that BLAST uses are quite optimal and well tested. However for the following reasons you might want to change them:

### Some Reasons to Change BLAST Default Parameters

<table>
<thead>
<tr>
<th>Reason</th>
</tr>
</thead>
<tbody>
<tr>
<td>The sequence you're interested in contains many identical residues; it has a biased composition.</td>
</tr>
<tr>
<td>BLAST doesn’t report any results</td>
</tr>
<tr>
<td>Your match has a borderline E-value</td>
</tr>
<tr>
<td>BLAST reports too many matches</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameters to Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequence filter (automatic masking)</td>
</tr>
<tr>
<td>Change the substitution matrix or the gap penalties.</td>
</tr>
<tr>
<td>Change the substitution matrix or the gap penalties to check the match robustness.</td>
</tr>
<tr>
<td>Change the database you’re searching OR filter the reported entries by keyword OR increase the number of reported matches OR increase Expect, the E-value threshold.</td>
</tr>
</tbody>
</table>

Choosing the right Parameters: sequence masking

- When BLAST searches databases, it makes the assumption that the average composition of any sequence is the same as the average composition of the whole database.

- However this assumption doesn’t hold all the time, some sequences have biased compositions, e.g. many proteins contain patches known as low-complexity regions: such as segments that contain many prolines or glutamic acid residues.

- If BLAST aligns two proline-rich domains, this alignment gets a very good E-value because of the high number of identical amino acids it contains. **BUT** there is a good chance that these two proline-rich domains are not related at all.

- In order to avoid this problem, sequence masking can be applied.
Choosing the right Parameters: DNA masking

- DNA sequences are full of sequences repeated many times: most of genomes contain many such repeats, especially the human genome (60% are repeats).

- If you want to avoid the interference of that many repeats, select the Human Repeats check box that appears in the blastn page of NCBI or the Xblast-repsim filter.

Controlling the BLAST output

- If your query belongs to a large protein family, the BLAST output may give you troubles because the databases contain too many sequences nearly identical to yours => preventing you from seeing a homologous sequence less closely related but associated with experimental information; so how to proceed?

1) Choosing the right database
   If BLAST reports too many hits, search for Swiss-Prot (100 times smaller) rather than NR; or search only one genome.

2) Limit by Entrez query (NCBI)
   For instance, if you want BLAST to report proteases only and to ignore proteases from the HIV virus, type “protease NOT hiv1[Organism].”

3) Expect
   Change the cutoff for reporting hits, to force BLAST to report only good hits with a low cutoff.
Changing the BLAST alignment parameters

• Among the parameters that you can change on the BLAST servers two important ones have to do with the way BLAST makes the alignments: the gap penalties (gap costs) and the substitution matrix (matrix) or match/mismatch parameters (DNA).

• Use a substitution matrix adapted to the expected divergence of the searched sequences (nevertheless most of the time BLOSUM62 works well):
  - BLOSUM 80: increase selectivity (exclude false positive, missing true positives) (closest to PAM120)
  - BLOSUM 45: increase sensitivity (more true matches, include false positives) (closest to PAM250)

<table>
<thead>
<tr>
<th>BLOSUM 80</th>
<th>BLOSUM 62</th>
<th>BLOSUM 45</th>
</tr>
</thead>
<tbody>
<tr>
<td>PAM 1</td>
<td>PAM 120</td>
<td>PAM 250</td>
</tr>
</tbody>
</table>

Less divergent → More divergent

Changing the BLAST alignment parameters

Most of the BLAST searches fall into one of two categories: 1. **mapping** and 2. **exploring**;

1. **Mapping**: finding the position of one sequence within another (e.g. finding a gene within a genome) ⇒ you can expect the alignments to be nearly identical, and the coordinates are generally the focus of the results;

2. **Exploring**: the goal is usually to find functionally related sequences ⇒ the alignment and alignment statistics (score, E-value, percent identity, ...) are often of greatest importance

-\( W \) = word size
-\( G \) = open gap penalty
-\( E \) = extension gap penalty
(\( X \) = default value)
Alignment parameters: BLASTN protocols

1. When sequences are expected to be nearly identical (mapping): +1/-3 match-mismatch parameters:
   - **Mapping oligos**: filtering (turned off): we want the entire oligo to match; -G 2(5) –E 1(2)
   - **Mapping nonspliced DNA to a genome**: mask repeats; increase the word size (faster but specific): -W 30 (11); -G 1(5) –E 3(2);
   - **Mapping cDNA/EST (determine exon-intron structure)**: mask repeats; reduce word size (-W 15) to see short exons; -G 1(5) –E 3(2); low E-value to cut down false positives (-e 1e-20); See also other programs, e.g. EST2GENOME, SIM4 and SPIDEY

2. Cross-species exploration (search for genes, regulatory elements, RNA genes): +1/-1 match-mismatch parameters (sequences expected to be similar but not identical), -W 9(11) to increase the sensitivity:
   - **Annotating Genomic DNA with ESTs** (similar transcripts for genes where no transcripts have been isolated yet): mask repeats; -G 1(5) –E 2; set low E-value to cut down false positives (-e 1e-20);

Alignment parameters: BLASTP protocols

Most BLASTP searches fall under the exploring category: try to learn about your query sequence by comparing it to other proteins:

- **Standard search (default parameters)**: balances speed and sensitivity; not ideal for very distant proteomes;

- **Fast insensitive search**: when performing multiple searches (but not for sequences that have less than 50 percent identity); sequences are expected to be very similar: BLOSUM80, set low E-value (-e 1e-5); -G 9(11) –E 2(1); -f 999 (11) identical word;

- **Slow, sensitive search**: looking for distant relatives; set E higher (-e 100); -f 9 (11) BLOSUM45; See also other program, e.g. HMMER, PSI-BLAST

- **f** = (T parameter of the Blast algorithm) word threshold score; only those words scoring equal or greater than the threshold will seed the alignment.
Changing the BLAST alignment parameters


**Step 3. Choose the appropriate search parameters or use default settings.**
Choosing Parameters for Protein-Based BLAST Searches.

<table>
<thead>
<tr>
<th></th>
<th>Default</th>
<th>Special Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Filter</td>
<td>on</td>
<td>off</td>
</tr>
<tr>
<td>Scoring Matrix</td>
<td>BLOSUM62</td>
<td>PAM30 for 35 and under</td>
</tr>
<tr>
<td>Word Size</td>
<td>3</td>
<td>3, or reduce to 2</td>
</tr>
<tr>
<td>E value</td>
<td>10</td>
<td>1000 or more</td>
</tr>
<tr>
<td>Gap costs</td>
<td>11,1</td>
<td>11,1</td>
</tr>
<tr>
<td>Alignments</td>
<td>50</td>
<td>50</td>
</tr>
</tbody>
</table>

**BLAST substitution matrix: short sequences**

* In particular, short query sequences can only produce short alignments, and therefore database searches with short queries should use an appropriately tailored matrix. The BLOSUM series does not include any matrices with relative entropies suitable for the shortest queries, so the older PAM matrices may be used instead.
* For proteins, a provisional table of recommended substitution matrices and gap costs for various query lengths is:

<table>
<thead>
<tr>
<th>Query Length</th>
<th>Substitution Matrix</th>
<th>Gap Costs</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;35</td>
<td>PAM-30</td>
<td>(9,1)</td>
</tr>
<tr>
<td>35-50</td>
<td>PAM-70</td>
<td>(10,1)</td>
</tr>
<tr>
<td>50-85</td>
<td>BLOSUM-80</td>
<td>(10,1)</td>
</tr>
<tr>
<td>85</td>
<td>BLOSUM-62</td>
<td>(10,1)</td>
</tr>
</tbody>
</table>
Alignment parameters: BLASTX protocols

BLASTX is generally used to find protein coding genes in genomic DNA or to identify proteins encoded by transcripts (exploring, but sometimes mapping):

- **Gene finding in genomic DNA**: mask repeats; BLOSUM62; higher E-value (-e 100) don’t want to miss low-scoring genes; -f 14(12), which increases the speed while being still quite sensitive;

- **Annotating ESTs (determine what protein they encode)**: slightly less sensitive parameters than the default ones, good compromise for speed and sensitivity: set low E-value (-e 1e-10) to prevent misclassification; -f 14(12)

Alignment parameters: TBLASTN protocols

Similar to BLASTX but with TBLASTN you map a protein to a genome or search EST databases for related protein not yet in the protein database:

- **Mapping a protein to a genome (to study related homologs or the genomic environment for regulatory elements)**: set a low E-value (-e 1e-5) to cut down the number of low scoring hits; -f 999 (13)

- **Annotating ESTs**: what protein do they encode?; -f 15 (13)
Alignment parameters: TBLASTX protocols

Coding sequences evolve slowly compared to the DNA: TBLASTX for gene-prediction for genomes that are appropriately diverged: not too much (human vs. E.coli) or not enough (human vs. chimpanzees)

- **Finding undocumented genes in genomic DNA**: mask repeats; -f 999(13)

- **Transcript of unknown function**: first BLASTX and then (if no results) TBLASTX with ESTs databases; -f 999(13)

Changing the BLAST alignment parameters

- Guidelines from BLAST tutorial at the swiss EMBnet server

<table>
<thead>
<tr>
<th>gap opening</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>Pam30</td>
<td>Pam30</td>
<td>Pam30</td>
</tr>
<tr>
<td>4</td>
<td>Pam30</td>
<td>Pam30</td>
<td>Pam30</td>
</tr>
<tr>
<td>5</td>
<td>Pam30</td>
<td>Pam30</td>
<td>Pam30</td>
</tr>
<tr>
<td>6</td>
<td>Pam30</td>
<td>Pam30</td>
<td>Pam30</td>
</tr>
<tr>
<td>7</td>
<td>Pam30</td>
<td>Pam30</td>
<td>Pam30</td>
</tr>
</tbody>
</table>
Conclusions

**Blast: the most used database search tool**
- Fast and very reliable even for a heuristic algorithm
- Does not necessarily find the best alignment, but most of the time it finds the best matching sequences in the database
- Easy to use with default parameters
- Solid statistical framework for the evaluation of scores

**but...**
- The biologist's expertise is still essential to the analysis of the results!

**Tips and tricks**
- For coding sequences always search at the protein level
- Mask low complexity regions
- Use a substitution matrix adapted to the expected divergence of the searched sequences (nevertheless most of the time BLOSUM62 works well)
- If there are only matches to a limited region of your query, cut out that region and rerun the search with the remaining part of your query

**BLAST Family**

- Faster algorithm for genomic search:
  - and SSAHA (Ensembl): http://www.ensembl.org/
  This program is optimized for aligning sequences that differ slightly as a result of sequencing or other similar "errors". (larger word size is used as default to speed up the search)

- PSI-BLAST and PHI-BLAST-> Wednesday
Acknowledgments & References

Volker Flegel, Frédérique Galisson

References

- Ian Korf, Mark Yandell & Joseph Bedell, BLAST, O’Reilly
- David W. Mount, Bioinformatics, Cold Spring Harbor Laboratory Press
- Jean-Michel Claverie & Cedric Notredame, Bioinformatics for Dummies, Wiley Publishing