Database Searching

Why search databases?

- To find out if a new DNA sequence already is deposited in the databanks.
- To find proteins homologous to a putative coding ORF.
- To find similar non-coding DNA stretches in the database,
  (for example: repeat elements, regulatory sequences).
- To locate false priming sites for a set of PCR oligonucleotides.

What databases are available?

- DNA (nucleotide sequences):
  The big databases: Genbank, Embl, DDBJ an their weekly updates. These databases exchange information routinely.
- Genomic databases like the: Human (GDB), Mouse (MGB), Yeast (SGB), etc…
- Special databases:
  ESTs (expressed sequence tags)
  STSs (sequence-tagged sites)
  EPD (eukaryotic promotor database
  REPBASE (repetitive sequence database)
  and many others.

What databases are available?

- Protein (amino acid sequences):
  The big databases are:
  Swiss-Prot (high level of annotation)
  PIR (protein identification resource)
- Translated databases like:
  SPTREMBL (translated EMBL)
  GenPept (translation of coding regions in GenBank)
- Special databases like:
  PDB(sequences derived from the 3D structure Brookhaven PDB)

What is a homologous sequence?

- A homologous sequence, in molecular biology, means that the sequence is similar to another sequence. The similarity is derived from common ancestry.
- Homologous proteins means that they are similar in their folding or their structure.

DNA vs. Protein searches

- DNA is composed of 4 characters: A,G,C,T It is anticipated that on the average, at least 25% of the residues of any 2 unrelated aligned sequences, would be identical.
- Protein sequence is composed of 20 characters (aa). The sensitivity of the comparison is improved. It is accepted that convergence of Proteins is rare, meaning that high similarity between 2 proteins always means homology.

DNA vs. Protein searches

- What should we use to search for similarity, the nucleotide or the protein sequences?
- If we have a nucleotide sequence, should we search the DNA databases only? Or should we translate it to protein and search protein databases?

  Note, that by translating into aa sequence, we'll presumably lose information, since the genetic code is degenerate, meaning that two or more codons can be translated to the same amino acid.

DNA vs. Protein searches

- What about very different DNA sequences that code for similar protein sequences? We certainly do not want to miss those.
- Conclusion: We should use proteins for database similarity searches when possible.

DNA vs. Protein searches

- The reasons for this conclusion are:
  - When comparing DNA sequences, we get significantly more random matches than we get with proteins.
  - The DNA databases are much larger, and grow faster than Protein databases. Bigger database means more random hits!
  - For DNA we usually use identity matrices, for protein more sensitive matrices like PAM and BLOSUM, which allow for better search results.
  - The conservation in evolution, protein are rarely mutated.

Main algorithms for database searching

- FastA
  - – Better for nucleotides than for proteins
- BLAST - Basic Local Alignment Search Tool
  - – Better for proteins than for nucleotides
- Smith-Waterman
  - – More sensitive than FastA or BLAST.

Specificity and sensitivity

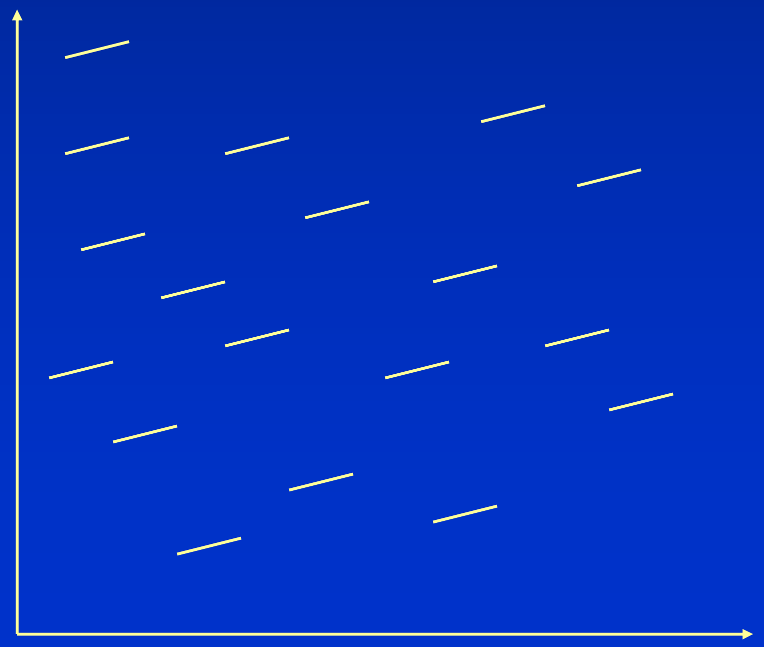
- Definitions:
- Sensitivity: the ability to detect "true positive" matches. The most sensitive search finds all true matches, but might have lots of false positives
- Specificity: the ability to reject "false positive" matches. The most specific search will return only true matches, but might have lots of false negatives
- view url

The FastA software package

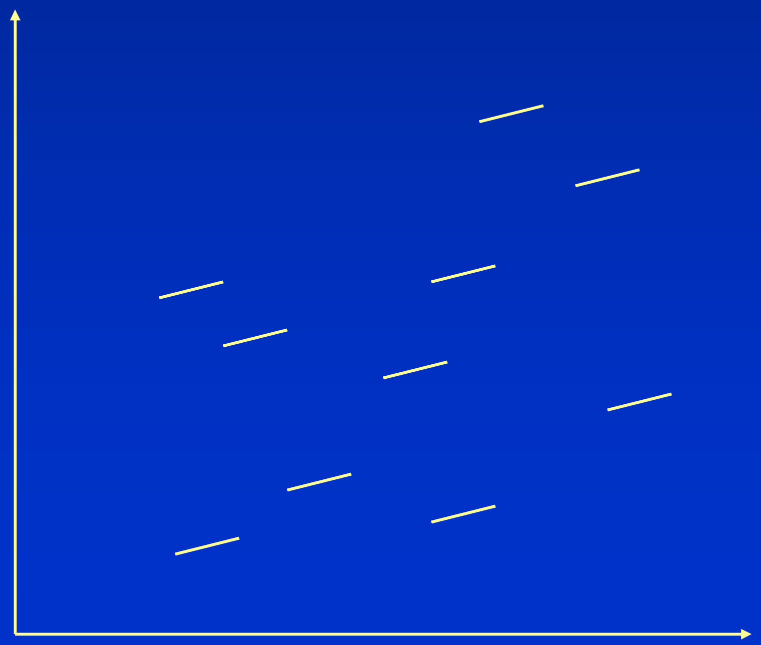
- FastA uses the method of Pearson and Lipman (PNAS 85: 2444-2448, 1988).
- FastA compares a DNA sequence to a DNA database or a protein sequence to a protein database.
- FastA is a family of programs, which include:
  - FastA, TFastA, Ssearch, etc...

General view of how the fasta program works:

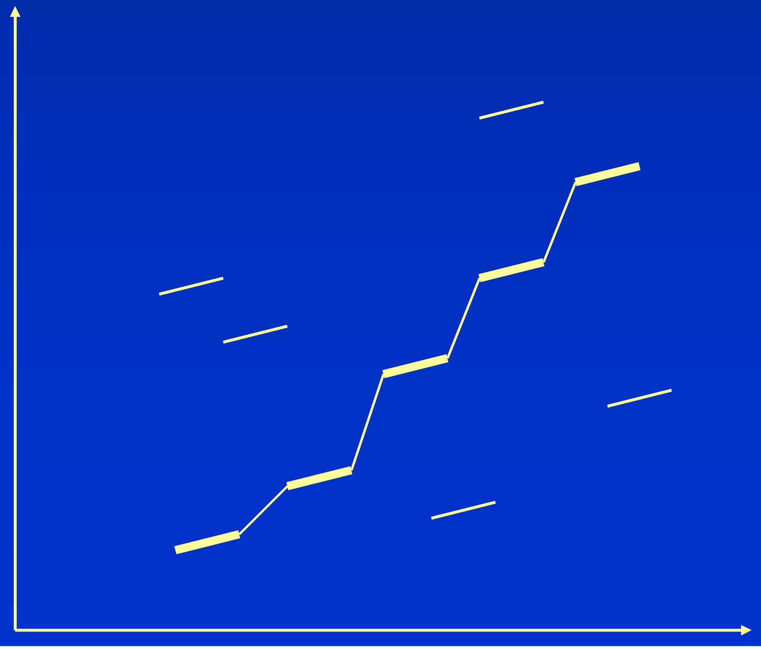
FastA locates regions of the query sequence and the search set sequence that have high densities of exact word matches.



The ten highest-scoring regions are rescored using a scoring matrix. The score of the highest scoring initial region is saved as the **init1 score**.



Next: FastA determines if any of the initial regions from different diagonals may be joined together to form an approximate alignment with gaps. Only non-overlapping regions may be joined. The score for the joined regions is the sum of the scores of the initial regions minus a joining penalty for each gap. The score of the highest scoring region, at the end of this step, is saved as the **initn score**.



After computing the initial scores, FastA determines the best segment of similarity between the query sequence and the search set sequence, using a variation of the Smith-Waterman algorithm. The score for this alignment is the **opt score**.

- Last: FastA uses a simple linear regression against the natural log of the search set sequence length to calculate a normalized **z-score** for the sequence pair.
- Using the distribution of the z-score, the program can estimate the number of sequences that would be expected to produce, purely by chance, a z-score greater than or equal to the z-score obtained in the search. This is reported as the **E() score**.

Where to find the FastA programs?

- FastA searches can be done on the WWW FastA server at EBI: http://www2.ebi.ac.uk/fasta3/
- On a stand alone computer such as dapsas1 at the Weizmann institute.
- From the GCG software package.

Comparison programs in the FastA3 package

- Fasta3 - Compare a protein sequence to a protein database, or a DNA sequence to a DNA database, using the fasta algorithm. Search speed and selectivity are controlled with the “ktup” (wordsize) parameter.
  - Tips for ktup: For proteins, the defualt, ktup=2, ktup=1 is more sensitive but slower.
  - For DNA, ktup=6, the defualt, ktup=3 or ktup=4 more sensitivity, ktup=1 for oligonucleotides (length <20).

Comparison programs in the FastA3 package

- Ssearch3 - Compare a protein sequence to a protein database, or a DNA database, using the Smith-Waterman algorithm. It is very slow but much more sensitive for full-length proteins comparison.
- Fastx3 - Compare a DNA sequence to a protein database, by comparing the translated DNA sequence in three frames and allowing gaps and frameshifts.

Which program When?

- Identify unknown protein -
  fasta3, ssearch3, tfastx3
- Identify structural DNA sequence - (repeated DNA, structural RNA)
  fasta3, (first with ktup=6 than ktup=3)
- Identify EST sequence -
  fastx3 (check first if the EST encodes a protein homologous to a known protein).

Running FastA

```
[~]% fasta -batch
```

FastA does a Pearson and Lipman search for similarity between a query sequence and a group of sequences of the same type (nucleic acid or protein). For nucleotide searches, FastA may be more sensitive than BLAST.

```
 FASTA with what query sequence ?  gb:y00762

                  Begin (* 1 *) ?
                End (*  1667 *) ?

 Search for query in what sequence(s) (* GenEMBL:* *) ?
embl:*
```

Running FastA

```
What word size (* 6 *) ?

 Don't show scores whose E() value exceeds: (* 2.0 *):

 What should I call the output file (* y00762.fasta *) ?

 ** fasta will run as a batch or at job.

 ** fasta was submitted using the command:
    "  atnow  "

job bfbecker.874320923.a at Mon Sep 15 13:55:23 1997

 [~]%
```

Output of FastA

```
!!SEQUENCE_LIST 1.0

(Nucleotide) FASTA of: y00762  from: 1 to: 1667  September
15, 1997 17:33

LOCUS       HSACHRA      1667 bp    RNA             PRI       23-
MAR-1995
DEFINITION  Human mRNA for muscle acetylcholine receptor
alpha-subunit.
ACCESSION   Y00762
NID         g28308
KEYWORDS    acetylcholine receptor alpha.
SOURCE      human. . . .
```

Fasta output

- The distribution of scores graph of frequency of observed scores
- expected curve (asterisks) according to the extreme value distribution
- the theoretic curve should be similar to the observed results
- deviations indicate that the fitting parameters are wrong
  - too weak gap penaltie
  - compositional biases

Fasta output

- The list of hits
- name, description, and length (between parentheses), general information about the hit.
- initn, init1 and opt scores. The scores calculated at the various stages of the comparison
- z-score, the score normalised by sequence length
- expectation value E(), how many hits we expect to find by chance with such a score, while comparing this query to this database. It is important to keep in mind that the E() value does not represent a measure of similarity between the two sequences!

Fasta output

- The information for each hit
- general information and statistics
- the Smith-Waterman score between the query and this hit
- the percent of identity and the length of overlap
- the alignment itself
- Statistics on the query, the database, and the search

Output of FastA

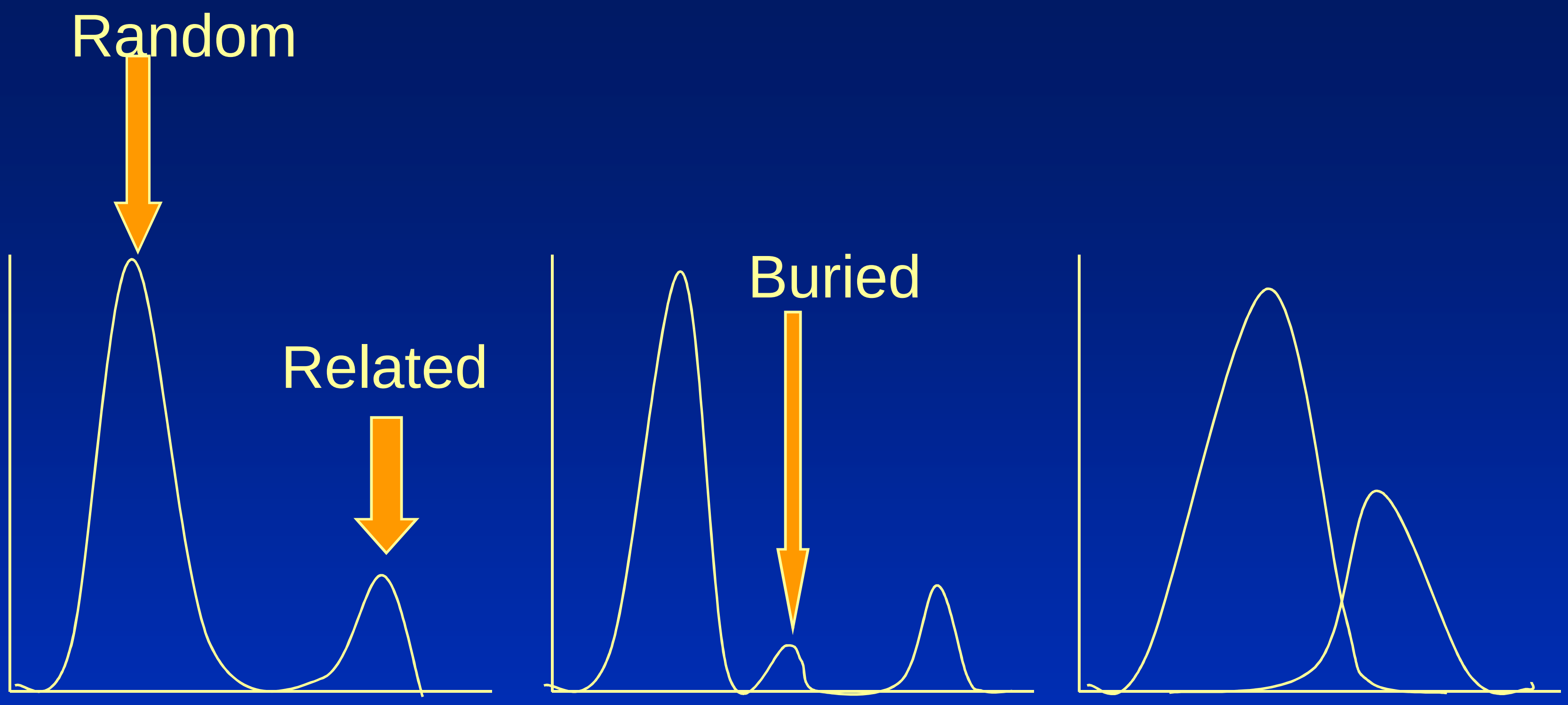
```
Sequences too short to analyze: 26 (110 symbols)
 Databases searched:
   EMBL, Release 51.0, Released on 25Jun1997, Formatted
on 13Jul1997

 Searching with both strands of the query.
 Scoring matrix: GenRunData:fastadna.cmp
 Constant pamfactor used
 Gap creation penalty: 16      Gap extension penalty: 4
```

Output of FastA

```
< 20   222      0 :*
  22    30      0 :*
  24    18      1 :*
  26    18     15 :*
  28    46    159 :*
  30   207    963 :*
  32  1016   3724 :=  *
  34  4596  10099 :====    *
  36  9835  20741 :=========        *
  38 23408  34278 :====================        *
  40 41534  47814 :===================================    *
  42 53471  58447 :============================================   *
  44 73080  64473 :====================================================*=======
  46 70283  65667 :=====================================================*====
  48 64918  62869 :===================================================*==
  50 65930  57368 :===============================================*=======
  52 47425  50436 :=======================================  *
  54 36788  43081 :===============================    *
  56 33156  35986 :============================ *
  58 26422  29544 :======================  *
  60 21578  23932 :================== *
  62 19321  19187 :===============*
  64 15988  15259 :============*=
  66 14293  12060 :=========*==
  68 11679   9486 :=======*==
  70 10135   7434 :======*==
```

Output of FastA

```
72   8957   5809 :====*===
  74   7728   4529 :===*===
  76   6176   3525 :==*===
  78   5363   2740 :==*==
  80   4434   2128 :=*==
  82   3823   1628 :=*==
  84   3231   1289 :=*=
  86   2474    998 :*==
  88   2197    772 :*=
  90   1716    597 :*=
  92   1430    462 :*=          :===============*========================
  94   1250    358 :*=          :============*===========================
  96    954    277 :*          :=========*=======================
  98    756    214 :*          :=======*===================
 100    678    166 :*           :=====*==================
 102    580    128 :*           :====*===============
 104    476     99 :*          :===*=============
 106    367     77 :*          :==*==========
 108    309     59 :*          :==*========
 110    287     46 :*          :=*========
 112    206     36 :*          :=*======
 114    161     28 :*          :*=====
 116    144     21 :*          :*====
 118    127     16 :*          :*====
```



Ideal

Borderline

No Good

Change Matrix

Output of FastA

```
The best scores are:                    init1 initn   opt    z-sc E(699079

EM_HUM1:HSACHRA    Begin: 1  End:   1667
! Y00762 Human mRNA for muscle acetyl... 8335  8335  8335  9159.3       0
EM_HUM2:S77094    Begin: 1  End:   1667
! S77094 nicotinic acetylcholine rece... 8299  8299  8299  9119.6       0
EM_OM:BTACHRA1    Begin: 10  End:   1422
! X02509 B.Taurus mRNA for acetylchol... 6018  6244  6048  6636.8       0
EM_RO:MMACHRAM    Begin: 4  End:   1634
! X03986 Mouse mRNA for muscle nicoti... 5570  5630  5881  6457.6       0
EM_RO:MMACHRAB    Begin: 59  End:   1731
! M17640 Mus musculus acetylcholine r... 5552  5607  5873  6448.4       0
EM_RO:RNACRA1    Begin: 27  End:   1678
! X74832 R.norvegicus mRNA for acetyl... 5550  5713  5807  6375.8       0
EM_OV:XLACHRA    Begin: 32  End:   1416
! X07067 Xenopus mRNA for muscle aety... 3309  3309  3558  3901.8       0
EM_OV:FSACHRA    Begin: 243  End:   1572
! J00963 Ray (T.californica) acetylch... 3345  3345  3527  3865.4       0
EM_OV:TMACHR    Begin: 120  End:   1449
! M25893 T.marmorata acetylcholine re... 3318  3318  3500  3836.4       0
EM_OV:DRU70438    Begin: 180  End:   1536
! U70438 Danio rerio muscle nicotinic... 3129  3129  3426  3753.7       0
EM_OV:XLACHRA1    Begin: 16  End:   1397
```

Output of FastA

```
y00762
EM_HUM1:HSACHRA

ID   HSACHRA    standard; RNA; HUM; 1667 BP.
AC   Y00762;
NI   g28308
DT   02-APR-1988 (Rel. 15, Created)
DT   23-MAR-1995 (Rel. 43, Last updated, Version 6)
DE   Human mRNA for muscle acetylcholine receptor alpha-subunit .


SCORES Init1:8335  Initn:8335  Opt: 8335 z-score: 9159.3 E(): 0
 100.0% identity in 1667 bp overlap

                     10        20        30        40        50
y00762       AAGCACAGGCCACCACTCTGCCCTGGTCCACACAAGCTCCGGTAGCCCATGGA
             |||||||||||||||||||||||||||||||||||||||||||||||||||||
HSACHRA      AAGCACAGGCCACCACTCTGCCCTGGTCCACACAAGCTCCGGTAGCCCATGGA
                     10        20        30        40        50
```

Output of FastA

```
y00762
EM_RO:MMACHRAB
ID   MMACHRAB   standard; RNA; ROD; 1860 BP.
AC   M17640;
NI   g2073542
DT   16-JUL-1988 (Rel. 16, Created)
DT   13-MAY-1997 (Rel. 51, Last updated, Version 3)
DE   Mus musculus acetylcholine receptor alpha-subunit mRNA, comple

SCORES  Init1: 5552  Initn: 5607  Opt: 5873 z-score: 6448.4 E(): 0
  84.1% identity in 1675 bp overlap
                                                  10         20
y00762                                    AAGCACAGGCCACCACTC-TGCCC
                                           || || || || ||| | || ||
MMACHRAB     CTTTACAGCTCACTTCCTTTCTCAGGCAGTAGGACCGG-CAGCACACGTGGCC
             30        40        50        60         70        80

                     40        50        60        70        80
y00762       CACAAGCTCCGGTAGCCCATGGAGCCCTGGCCTCTCCTCCTGCTCTTTAGCCT
             ||      ||   |||||||||||| || | || | ||||||||  |  ||||
MMACHRAB     CAGCGCGACCCACAGCCCATGGAGCTCTCGACTGTTCTCCTGCTGCTAGGCCT
```

Tips for FastA results

- When init1=init0=opt:
  100 % homology over the matched stretch.
- When initn > init1:
  more than 1 matching region in the database with poorly matching separating regions.
- When opt > initn:
  the matching regions are greatly improved by adding gaps in one or both of the sequences.

Statistical evaluation of results

- When the program finds a similarity between your query sequence and a database sequence it is not always clear how significant this similarity really is.
- To evaluate if this similarity is statistically significance, you can run any of these programs:
- (From GCG) gap -rand=100
- From the FastA package: prss or prdf

BLAST - Basic Local Alignment Search Tool

- Blast programs use a heuristic search algorithm. The programs use the statistical methods of Karlin and Altschul (1990,1993).
- Blast programs were designed for fast database searching, with minimal sacrifice of sensitivity to distant related sequences.

BLAST - Basic Local Alignment Search Tool

- BLAST programs search databases in a special compressed format.
  To use your own privat database with blast, you need to format it to the blast format.

BLAST Programs

- BLAST is actually a family of programs
  - BLASTN - Nucleotide query searching a nucleotide database.
  - BLASTP - Protein query searching a protein database.
  - BLASTX - Translated nucleotide query sequence (6 frames) searching a protein database.
  - TBLASTN - Protein query searching a translated nucleotide (6 frames) database.
  - TBLASTX - Translated nucleotide query (6 frames) searching a translated nucleotide (6 frames) database.

Where to find the BLAST programs?

- BLAST searches can be done on the WWW BLAST server at NIH: http://www.ncbi.nlm.nih.gov/BLAST/
- On a stand alone computer such as dapsas1 at the Weizmann institute.
- From the GCG software package.

Blast method

- Compare query to each sequence in database
- Use heuristic to speed pairwise comparison
- Create 'sequence abstraction' by listing exact and similar words
  - on the fly for the query
  - in advance for the database
- Find similar words between query and each database sequence
- Extend such words to obtain high-scoring sequence pairs (HSPs)
- Calculate statistics analytically

Gapped BLAST

- BLAST 2.0 is a new version with new capabilities such as Gapped-Blast and Psi-Blast.
- The Gapped Blast algorithm allows gaps to be introduces into the alignments. That means that similar regions are not broken into several segments (as in the older versions).
- This method reflects biological relationships much better.

PSI - BLAST

- PSI (Position Specific Iterated) Blast provides a new automatic “profile like” search.
- The program first performs a gapped blast search of the database. The information of the significant alignments are then used by the program to construct a “position specific” score matrix. This matrix replaces the query sequence in the next round of database searching.
- The program may be iterated until no new significant alignments are found.

Blast output

- The list of hits
- Database accession codes, name, description, general information about the hit
- Score in bits, the alignment score expressed in units of information. Usually 30 bits are required for significance
- Expectation value E(), how many hits we expect to find by chance with this score, when comparing this query to the database.

  It is important to keep in mind that the E() value does not represent a measure of similarity between the two sequences.

Blast output

- The information for each hit
- A header including hit name, description, length
- The same for all additional entries removed due to redundancy
- Composite expectation value
- Each hit may contain several HSPs
- score and expectation value
- – how many identical residues
- – how many residues contributing positively to the score
- The local alignment itself

The Smith-Waterman Tools

- Smith-Waterman searching method:
- Compare query to each sequence in database
- Do full Smith-Waterman pairwise comparisons
- Use search results to generate statistics

Where to find the SW programs?

- Since SW searching is exhaustive, it is the slowest method we use a special hardware + software (Bioccelerator) to run the programs.
- Bioccelerator is available here inTAU at the
- at the Weizmann Institute http://dapsas1.weizmann.ac.il/bcd/bcd_parent/ bcd_bioccel/bioccel.html
- The Bioccelerator from the command line on dapsas1 or life2.

Comparison of programs

- Concept:
- SW and BLAST: local alignments
- FASTA: global alignments

  BLAST can report more than one HSP per database entry, FASTA reports only one segment (match).
- Speed:
- BLAST > FASTA >> SW
- Sensitivity: SW > FASTA > BLAST (old version!)

Comparison of programs

- Sensitivity:
- FASTA is more sensitive, misses less homologues, (the opposite can also happen - if there are no identical residues conserved, but this is infrequent).
- FASTA gives a better separation between true homologues and random hits.
- Usually when FASTA gives an unexpected hit, it is an even farther homologue.

Comparison of programs

- Statistics:
- BLAST calculates probabilities
- sometimes fails entirely if some assumptions are invalid
- FASTA calculates significance 'on the fly' from the given dataset
- more relevant
- problematic if the dataset is small

Tips for DB searches

- Use latest database version
- Run Blast first, then depending on your results run a finer tool (fasta, ssearch, SW, blocks, etc..)
- Where possible use translated sequence.
- E() < 0.05 is statistically significant, usually biologically interesting. Check also 0.05 < E() <10 because you might find interesting hits.
- Pay attention to abnormal composition of the query sequence, it usually causes biased scoring.

Tips for DB searches

- Split large query sequence (if >1000 for DNA, >200 for protein).
- If the query has repeated segments, remove them and repeat the search.