

# SEQUENCE MATCHING, SIMPLE SEARCHING

---

PGA Course in Bioinformatics  
Tools for Comparative Analysis  
June 11, 2001

## Outline

- Sequence alignment algorithms
  - Rigorous Optimality: Needleman-Wunsch and Smith-Waterman
  - Rapid, heuristic algorithms
    - BLAST
    - FASTA
    - and their relatives
- Databases and Search Tools

## MAJOR SITES WE WILL USE

☞ <http://www.ncbi.nlm.nih.gov/>

☞ <http://workbench.sdsc.edu>

## MEDICAL SUBJECT HEADINGS

- ☞ CONTROLLED Vocabulary
- ☞ Indexing of articles, books, etc.
- ☞ Current version has over 300,000 terms
- ☞ Can download list and make your own assortment

## Needleman Wunsch Algorithm

- ☛ Global alignment
- ☛ Guaranteed to calculate an Optimal similarity score
- ☛ Begin at the beginning of each sequence and go to the end.
- ☛ Cannot detect domains

## Smith-Waterman Algorithm

- ☛ Optimal Local Alignment
- ☛ Guaranteed to find all significant matches to a given query
- ☛ Takes the query sequence versus every sequence in the database
- ☛ Can be used with arbitrary scoring systems
- ☛ **COMPUTATIONALLY EXPENSIVE!!!**

## Scoring Matrices

- Relatively simple for DNA-gap penalties or mismatches-can be made to look at Pu/Py
- Protein matches look also at similarity (leu/ileu)

## Protein Scoring Matrices

- Chemical similarity: 210 pairs of aa
- Nearness in Genetic Code
- Chemical similarity, e.g., hydrophobicity
- Observed Substitution Schemes

## Observed AA Substitution Matrices

- ☛ PAM

- ☛ BLOSUM

### PAM: Point Accepted Mutation

- ☛ DAYHOFF et al.
- ☛ Residue replacement in related proteins
- ☛ A model of molecular evolution
  
- ☛ 1 PAM = average change in 1% of all amino acid possibilities
  
- ☛ 100 PAMs does not mean every residue is changed.

## PAM continued

☛ TIME is NOT correlated with PAM

Means different families of proteins  
evolve at different rates

## BLOSUM

☛ Block Substitution Matrix

☛ Henikoff and Henikoff, PNAS, 1992

☛ Number following indicates per cent  
identity within set, BLOSUM62=62% id

☛ Finds short, highly similar sequences

## BLAST - Basic Local Alignment Sequence Tool

- Objective: find all local regions of similarity distinguishable from random
- Only local alignments permitted,
- Gaps permitted in version 2
- Statistically sound (Karlin and Altschul), but no guarantee of optimality

## BLAST: Three Step Algorithm

- Compile a list of high scoring words of length  $w$  ( $w=4$  for proteins, 12 for nucleic acids)
- Scan for word hits of score greater than threshold,  $T$
- Extend word hit in both directions to find High Scoring Pairs with scores greater than  $S$

## Other BLAST Programs

- BLASTN: nucleic acid query to NA database
- BLSATP: Protein query to Protein database
- BLASTX: Translated nucleic acid query to Protein database
- TBLASTN: Protein query against (translated) nucleic acid database
- TBLASTX: Translated nucleic acid

## OTHER BLAST VARIATIONS

- Gapped BLAST (BLAST 2.0) -extend words from no-gap to gap, generate gapped alignments
- PSI-BLAST- Position Specific Iterated BLAST-use gapped BLAST, generate a Profile from multiple iterations used instead of the input and Distance Matrix

## Limitations to BLAST

- ☞ Needs islands of strong homology
- ☞ Limits on the combination of scoring and penalty values
- ☞ The variants (blastx, tblastn, tblastx) use 6-frame translation-miss sequences with frameshifts)
- ☞ Finds and reports ONLY local alignments

## A WALK THROUGH BLAST

---

## BLAST RULES OF THUMB

- ☛ For short amino acid sequences (20-40), 50% identity happens by chance
- ☛ If A and B are homologous, and B and C are homologous, then A and C are, even if you can't see it.
- ☛ You can get similarity in the absence of homology for low complexity, transmembrane and coiled-coil regions. These have to be eliminated by you.

## BLAST Significance

- ☛ If you change scoring systems, you can still compare search results if you normalize the score.

$S' = (\lambda S - \ln K) / \ln 2$ .  $\lambda$  and  $K$  are associated with the scoring system.

$S'$ , with a given  $E$ , is significant if it is greater than  $N/E$ ,  $N$  the size of the search space.

## FASTA: FAST Alignment

- <http://alpha10.bioch.virginia.edu/fasta/>
- <http://www2.ebi.ac.uk/fasta3>
- <http://workbench.sdsc.edu>
- Rapid Global alignment
- Not a strong mathematical basis

## FASTA: WHY USE IT?

- Allow alignments to shift frames

## LALIGN

- Essentially a FASTA derivative for local alignments
- Compares two proteins to identify regions of similarity
- Will report several sequence alignments within a given sequence
- Works for internal repeats that are missed by FASTA because of gaps.

## SITEs for LALIGN

- <http://fasta.bioch.virginia.edu/fasta/lalign.htm>
- <http://xylian.igh.cnrs.fr/bin/lalign-guess.cgi>
- <http://biowb.sdsc.edu> (registration necessary but painless)
- PALIGN  
<http://fasta.bioch.virginia.edu/fasta/palign.htm>  
(plots a graph of the areas of alignment)

## ENTREZ: Linked Databases

<http://www.ncbi.nlm.nih.gov/Entrez/>

- ☛ Concept of Neighbor-usually BLAST relationship
- ☛ Precomputed=Fast
- ☛ Related sequence, structure neighbors, related articles

## EST DATABASES:Quality issues

- ☛ SEQUENCE QUALITY
  - calculated error less than 1% (Phred-20) is the rule
  - frameshifts and stops common
  - Rules are usually observed by exception
  - There are lots of exceptions in the public data
  - Many 3' UTRs

## EST Databases: Quality #2

### ☞ CLONE QUALITY

- Over-representation
- Tissue specificity
- Developmental stage specificity
- Unprocessed mRNA clones
- Chimeras
- Contamination

## EST Cluster Databases

- ☞ STACK-at SANBI <http://sanbi.ac.za>
- ☞ TIGR-animals, plants, other  
<http://www.tigr.org/tdb/tgi.shtml>
- ☞ Unigene-NCBI
  - Human, mouse, rat, cow, zebrafish
  - mRNAs
  - predicted mRNAs

## UNIGENE

### ☞ A LIST OF LISTS

- The cluster and known EST, mRNA pieces
- Additional annotation-gene name, etc.
- Distributed as a subset of dbest

NOT included in the BLAST searchable DB at NCBI

## Caveats on Clusters

- ☞ Not stable
- ☞ Can go to complete cDNAs as available

## LOCUSLINK

(<http://www.ncbi.nlm.nih.gov/LocusLink>)

- A useful, searchable compendium of loci across human, mouse, rat, Drosophila and zebrafish
- Linked for PubMed, OMIM, RefSeq, Homologene data, Unigene, and Variation Data

## Resources for Genomic Comparison

- GLASS-<http://plover.lcs.mit.edu>
- PipMaker: <http://bio.cse.psu.edu>
- Rosetta: [http:// plover.lcs.mit.edu/genes](http://plover.lcs.mit.edu/genes))
- SGP: <http://soft.ice.mpg.de/sgp-1>
- VISTA: <http://www-gsd.lbl.gov/VISTA>
- WABA:  
<http://www.cse.ucsc.edu/~kent/xenoAli/index.html>

## EFFICIENT SEARCHING

☞ Use Wild Cards: #,\$,?,\*

☞ Use Boolean Operators

- Not
- And
- Or
- Nor

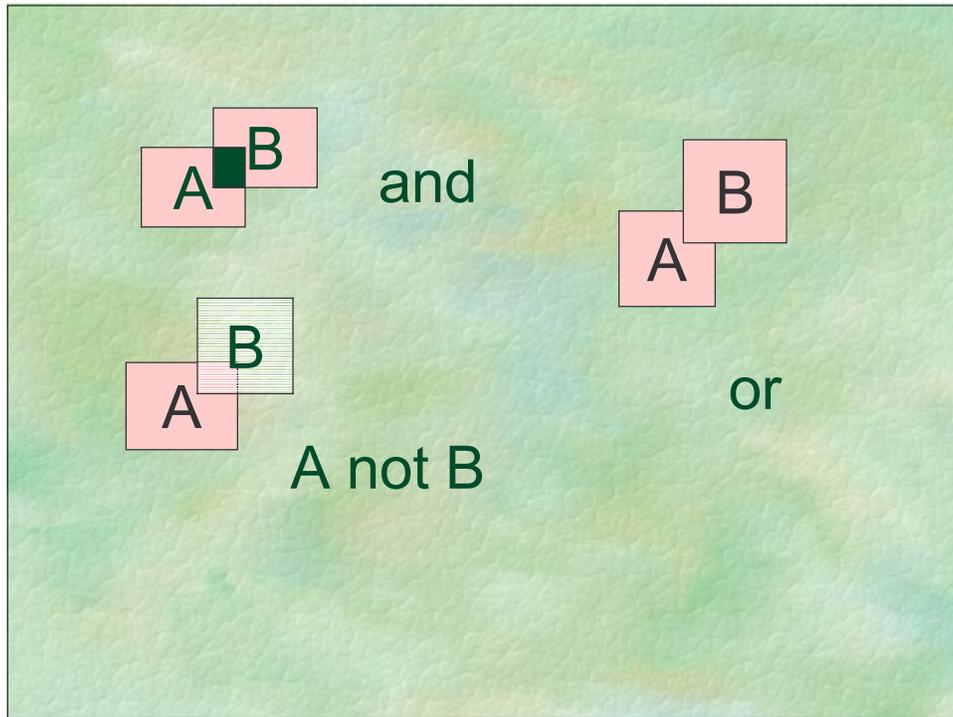
## Boolean Operators

☞ **AND** A and B BOTH

☞ **OR** A or B EITHER

☞ **NOT** B not A Have B, do not have A

☞ **NOR** A nor B A but not B OR B but  
not A



## WILD CARDS

- ☞ Match one character-NCBI uses #
- ☞ Match zero or one character NCBI uses \$, others ?
- ☞ Match zero or more characters-usually \*

## RULES OF THUMB

- ☛ Use an up-to-date database; repeat often
- ☛ Choose a fast algorithm
- ☛ Use the most recent version
- ☛ Work at the protein level--for a small amount of evolutionary change, DNA sequence contains less information about homology
- ☛ Respect your own *intuition*

## Other Resources

- ☛ NCBI Education Page  
<http://www.ncbi.nlm.nih.gov/Education/index.html>
- ☛ BCM Gene Finder  
[http://searchlauncher.bcm.tmc.edu/docs/sl\\_links.html](http://searchlauncher.bcm.tmc.edu/docs/sl_links.html)
- ☛ EBI-SwissProt, TrEMBL, PIR, SRS, Tools <http://www.ebi.ac.uk>
- ☛ ExPASy-SwissProt, TrEMBL  
<http://www.expasy.ch/>
- ☛ DISC-DNA Information and Stock Center <http://www.dna.affrc.go.jp>