Some Specific “Informatics” tools of Bioinformatics

• Databases
  ◦ NCBI GenBank - Protein and DNA sequence
  ◦ NCBI Human Map - Human Genome Viewer
  ◦ NCBI Ensembl - Genome browsers for human, mouse, zebra fish, mosquito
  ◦ TIGR - The Institute for Genome Research
  ◦ SwissProt - Protein Sequence and Function
  ◦ ProDom - Protein Domains
  ◦ Pfam - Protein domain families
  ◦ ProSite - Protein Sequence Motifs
  ◦ Protein Data Base (PDB) - Coordinates for Protein 3D structures
  ◦ SCOP Database - Domain structures organized into evolutionary families
  ◦ HSSP - Domain database using Dali
  ◦ FlyBase
  ◦ WormBase
  ◦ PubMed / MedLine

• Sequence Alignment Tools
  ◦ Clustal
  ◦ FASTA
  ◦ Simple Blast
  ◦ Gapped Blast
  ◦ PSI-Blast
  ◦ Hidden Markov Models

• 3D Structure Alignments / Classifications
  ◦ Dali
  ◦ VAST
  ◦ PRISM
  ◦ CATH
  ◦ SCOP
Multiple Sequence Alignments

- One of the most essential tools in molecular biology
- It is widely used in:
  - Phylogenetic analysis
  - Prediction of protein secondary/tertiary structure
  - Finding diagnostic patterns to characterize protein families
  - Detecting new homologies between new genes and established sequence families
- Practically useful methods only since 1987
- Before 1987 they were constructed by hand
- The basic problem: no dynamic programming approach can be used
- First useful approach by D. Sankoff (1987) based on phylogenetics

---

![Sequence Alignment Example](LEFT, adapted from Sonhammer et al. (1997). “Pfam,” Proteins 28:405-20. ABOVE, G Barton AMAS web page)
Progressive Multiple Alignments

(adapted from Sonhammer et al. (1997). "Pfam," Proteins 28:405-20)
Popular Multiple Alignment Programs
Clustering approaches for multiple sequence alignment: All vs All Transitive Relationships
Clustal Alignment
Community Assembly Through Adaptive Radiation in Hawaiian Spiders.

Phylogeny of spiny leg spider clade based on combined mitochondrial cytochrome oxidase I, 12S ribosomal DNA, and 16S ribosomal DNA sequences
Sequence-based Dendrograms
Fuse multiple alignment into:

- **Motif**: a short signature pattern identified in the conserved region of the multiple alignment
- **Profile**: frequency of each amino acid at each position is estimated
- **HMM**: Hidden Markov Model, a generalized profile in rigorous mathematical terms

Can get more sensitive searches with these multiple alignment representations (Run the profile against the DB.)
- several proteins are grouped together by similarity searches
- they share a conserved motif
- motif is stringent enough to retrieve the family members from the complete protein database
- PROSITE: a collection of motifs (1135 different motifs)
Motifs

- Each element in a pattern is separated from its neighbor by a “-”.
- The symbol “x” is used for a position where any amino acid is accepted.
- Ambiguities are indicated by listing the acceptable amino acids for a given position, between brackets “[]”.
- Ambiguities are also indicated by listing between a pair of braces “{}” the amino acids that are not accepted at a given position.
- Repetition of an element of the pattern is indicated by with a numerical value or a numerical range between parentheses following that element.

<table>
<thead>
<tr>
<th>Motif Name</th>
<th>Description</th>
<th>Pattern</th>
<th>Domain or Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>PKC_PHOSPHO_SITE</td>
<td>Protein kinase C phosphorylation site</td>
<td>[ST]-x-[RK]</td>
<td>Post-translational modifications</td>
</tr>
<tr>
<td>RGD</td>
<td>Cell attachment sequence</td>
<td>R-G-D</td>
<td>Domains</td>
</tr>
<tr>
<td>SOD_CU_ZN_1</td>
<td>Copper/Zinc superoxide dismutase</td>
<td>[GA]-[IMFAT]-H-[LIVF]-H-x(2)-[GP]-[SDG]-x-[STAGDE]</td>
<td>Enzymes_Oxidoreductases</td>
</tr>
<tr>
<td>THIOL_PROTEASE_ASN</td>
<td>Eukaryotic thiol (cysteine) proteases active site</td>
<td>[FYCH]-[WI]-[LIVT]-x-[KRQAG]-N-[ST]-W-x(3)-[FYW]-G-x(2)-G-[LFYW]-[LIVMFYG]-x-[LIVMF]</td>
<td>Enzymes_Hydrolases</td>
</tr>
<tr>
<td>TNFR_NGFR_1</td>
<td>TNFR/CD27/30/40/95 cysteine-rich region</td>
<td>C-x(4,6)-[FYH]-x(5,10)-C-x(0,2)-C-x(2,3)-C-x(7,11)-C-x(4,6)-[DNEQSKP]-x(2)-C</td>
<td>Receptors</td>
</tr>
</tbody>
</table>
Prosite Pattern -- EGF like pattern

A sequence of about thirty to forty amino-acid residues long found in the sequence of epidermal growth factor (EGF) has been shown [1 to 6] to be present, in a more or less conserved form, in a large number of other, mostly animal proteins. The proteins currently known to contain one or more copies of an EGF-like pattern are listed below.

- Bone morphogenic protein 1 (BMP-1), a protein which induces cartilage and bone formation.
- Caenorhabditis elegans developmental proteins lin-12 (13 copies) and glp-1 (10 copies).
- Calcium-dependent serine proteinase (CASP) which degrades the extracellular matrix proteins type ...
- Cell surface antigen 114/A10 (3 copies).
- Cell surface glycoprotein complex transmembrane subunit .
- Coagulation associated proteins C, Z (2 copies) and S (4 copies).
- Coagulation factors VII, IX, X and XII (2 copies).
- Complement C1r/C1s components (1 copy).
- Complement-activating component of Ra-reactive factor (RARF) (1 copy).
- Complement components C6, C7, C8 alpha and beta chains, and C9 (1 copy).
- Epidermal growth factor precursor (7-9 copies).

```plaintext
+-------------------+ +-------------------------+
| | | | | | | | |
|\ C\ -\ x\ (0,48) -\ C\ -\ x\ (3,12) -\ C\ -\ x\ (1,70) -\ C\ -\ x\ (1,6) -\ C\ -\ x\ (2) -\ G\ -\ a\ -\ x\ (0,21) -\ G\ -\ x\ (2) -\ C\ -\ x
| | | | | | | | |
+-------------------+

'C': conserved cysteine involved in a disulfide bond.
'G': often conserved glycine
'a': often conserved aromatic amino acid
'*': position of both patterns.
'x': any residue
-Consensus pattern: C-x-C-x(5)-G-x(2)-C

[The 3 C's are involved in disulfide bonds]

FASTA

• Hash table of short words (3 - 7 residues) in the query sequence

• Go through DB and look for matches in the query hash; hits with Eval < Threshold

• Speed is linear in size of DB

VLICT = ___

VLICT AVLM VLICT AAAA VLICT MSDFFD
Join together query lookups into diagonals and then a full alignment.

$S_{1,3} = S_1 + S_3 \cdot JP$

$S_{1,3,5} = S_1 + S_3 + S_3 \cdot 2 \times JP$

JP = Joining penalty

(Adapted from D Brutlag)
Basic Blast


- BLAST employs substitution matrix which specifies a score \( s(i, j) \) for aligning each pair of amino acids.

- Starts with all overlapping words (3 residues) from query

- Scans DB for “hits” that score at least \( T \) when aligned with some word in the query sequence

- Each such hit is then extended to test if it is contained in a high scoring alignment; Extends High Scoring Pairs (HSPs) left and right to maximal length, until the running alignment score drops to \( S_{\text{max}} - X \)

- Basic Blast does not permit gaps in alignments

- Extension time accounts for > 90% of execution time; desirable to reduce the number of extensions performed
Blast: Extension of Hash Hits

- Identify High Scoring Segment Pairs (HSPs) with match values $S > T$.

- Extend hits, computing score $S$ for each extended sequence match.

- Stop extension when total score drops below $S_{\text{max}} - X$

- Parameters $T$ and $X$ define coverage and specificity.

- Extension is $O(N)$. This takes most of the time in Blast.
Blasting against a Sequence DB

- In searching database
  - BLAST Algorithm finds HSP (High Scoring Pairs)
  - Then extends this using BLAST Extension Rules
  - Best scoring segment in each DB sequence is then reported
  - Extensions do not allow for Gaps
Gapped BLAST and PSI-BLAST: a new generation of protein database search programs

Stephen F. Altschul*, Thomas L. Madden, Alejandro A. Schäffer¹, Jinghui Zhang, Zheng Zhang², Webb Miller² and David J. Lipman

ABSTRACT
The BLAST searching program is described. A new search algorithm, Gapped BLAST, is introduced that uses extended matches or gaps to increase search sensitivity and speed. PSI-BLAST, a related program, uses expectation values from the PSI-BLAST search algorithm as search targets in successive rounds of database searches.

Figure 3. A gapped extension generated by BLAST for the comparison of broad bean leghemoglobin I (87) and horse β-globin (88). (a) The region of the path graph explored when seeded by the alignment of alanine residues at respective positions 60 and 62. This seed derives from the HSP generated by the leftward of the two ungapped extensions illustrated in Figure 2. The Y dropoff parameter is the nominal score, 40, used in conjunction with BLOSUM-62 substitution scores and a cost of 10 + 4 for gaps of length 4. (b) The path corresponding to the optimal local alignment generated, superimposed on the bits described in Figure 2. The original BLAST program, using the one-hit heuristic with T = 11, is able to locate three of the five HSPs included in this alignment, but only the first and last achieve a score sufficient to be reported. (c) The optimal local alignment, with nominal score 75 and normalized score 32.4 bits. In the context of a search of SWISS-PROT (26), release 34 (21 219 459 residues), using the leghemoglobin sequence (143 residues) as query, the E-value is 0.54 if no edge-effect correction (22) is invoked. The original BLAST program locates the first and last ungapped segments of this alignment. Using sum-statistics with no edge-effect correction, this combined result has an E-value of 3.1 (21,22). On the central lines of the alignment, identities are echoed and substitutions to which the BLOSUM-62 matrix (18) gives a positive score are indicated by a "*" symbol.
Gapped Blast

• An HSP of interest is much longer than a single word pair, and may therefore entail multiple hits on the same diagonal and within a relatively short distance of one another. “Islands of Certainty”

• Choose window of length A, and invoke extension only when two non-overlapping hits are found within a distance A of one another on the same diagonal.

• This “two hit method” requires the existence of two non-overlapping word pairs on the same diagonal, within a distance A of one another, before extension is invoked.

• If the HSP generated has a normalized score of at least $S_g$, then a gapped extension (using dynamic programming algorithm) is triggered

• Runs ~ 3X faster than Original Blast, with higher sensitivity and coverage
Blast2: Gapped Blast

- Gapped Extension on Diagonals with two Hash Hits
PSI Blast: Iterated Application of BLAST to Position-Specific Matrices
[Position Specific Iterated (PSI) BLAST]

- Database searches using position-specific score matrices (profiles or motifs) are often much better able to detect weak relationships that are database searches that use a simple sequence as a query

- Compile a set of N sequences hit by Gapped-BLAST with E-val < Threshold (default 0.01)

- The Query is used as a Master to Construct a Multiple Sequence Alignment (MSA).

- An N x 21 (21st “residue” is a gap) Position-Specific Score Matrix is then computed from the MSA

- Use the new Position-Specific Score Matrix to rerun Gapped-BLAST, generate MSA, compile new Position Specific Score Matrix

- Iterate
Ψ-Blast

Parameters: overall threshold (T), inclusion threshold (X), iterations

• Automatically builds profile and then searches with this

Gapped BLAST and PSI-BLAST: a new generation of protein database search programs

Stephen F. Altschul*, Thomas L. Madden, Alejandro A. Schäffer¹, Jinghui Zhang, Zheng Zhang², Webb Miller² and David J. Lipman

National Center for Biotechnology Information, National Library of Medicine, Bethesda, MD 20894, USA, ¹Laboratory of Computational Biology, National Institutes of Health, Bethesda, Maryland, and ²Department of Computer Science, Johns Hopkins University, Baltimore, Maryland

Received June 20, 1997; Revised and Accepted July 1, 1997

ABSTRACT

The BLAST programs are widely used for searching protein and DNA databases for similarities. For protein comparisons, the algorithms are based on an empirical method for defining similarities, which is described in this paper. The search algorithm uses a dynamic programming method that is highly scalable and can be used to search large databases. The algorithm is implemented in the software package BLAST, which is available for download from the Nucleic Acids Research website.
PSI-Blast

Iteration Scheme

Blast
FASTA
Smith-Waterman
Profiles
HMMs

Sensitivity
Speed