APPLICATIONS OF MULTIPLE ALIGNMENT

PATTERNS, MOTIFS, BLOCKS AND PSI-BLAST

Shifra Ben-Dor    Irit Orr
PAIRWISE ALIGNMENT

DATABASE SEARCHING

MULTIPLE ALIGNMENT
MULTIPLE ALIGNMENT

Homology Modeling

Phylogenetic Analysis

Advanced Database Searches Patterns, Motifs, Promoter Elements
Why run similarity searches?

Similarity searches of databases are used in order to:

- Gain knowledge and understanding of a gene or protein, in terms of evolution, structure or function.
- Try to find homologous sequences, where homologous sequences mean that those sequences are derived from common ancestry.
The tools used for similarity searches (e.g. blast, fasta) are known to miss 10% - 20% of “true hits”. This “area” of similarity is known as the “twilight zone”

- The proportion of missed similarities are even greater when searching modular proteins (that are composed of several, small domains.)

- So, other tools are needed for these specific searches.

Database searching doesn’t always find what we want…..
Database searching doesn’t always find what we want…..

Even using advanced methods such as Smith-Waterman and Framesearch, more distantly related, though biologically relevant sequences are often missed, due to the requirement for high sequence similarity.
Biologically relevant sequences that are hard to find

- Proteins with several similar, short regions of similarity
  
  aaa..................bbb.....................ccc
  aaa.........bbb...........ccc

- Proteins with extended motifs
  
  GV (X20) C (X30) C

- Proteins with ‘inexact’ motifs (structural, electrostatic, hydrophobic/philic motifs)
How do we usually find them?

Historically, these protein families were found by looking for functionally related sequences, either in the same species or in others.

The similarities can also be seen by performing multiple alignments on the more distantly related sequences that we can find.
The bottom line ......

What we would like to do is harness the power of multiple alignments to help us in our database searches.
New tools are needed for these similarity specific searches, based on the knowledge gained from multiple alignments, (eg protein families).

Tools like motifs, patterns, blocks or profiles searches can help. These tools use family information to improve the sensitivity to distant family members (homologs).
The Old Method: Scoring Matrices

- Most database search methods, pairwise, and multiple alignments use previously derived matrices (such as PAM or BLOSUM) for scoring the change of an amino acid or nucleotide.

- These matrices are based on known protein families and the probabilities drawn from them are generalized for all sequences.
Scoring Matrices

- The PAM family (Dayhoff) is based on evolutionary distance. The matrices were derived from closely related sequences and the mutations seen in them.

- The Blosum family (Henikoff and Henikoff) were derived from more distantly related sequences. The number of the matrix is percent identity threshold used for clustering.
New Method:

• What we’d like to do is derive a scoring matrix from our specific family of sequences

• This takes into account which positions are absolutely unchangeable, which are more flexible, and is not a generalized score based on all proteins available, but just those that are relevant to a specific family of proteins
Terminology

- Motif
- Pattern
- Profile
Terminology: Motif

- Motif - small conserved region within a large sequence.
- Also called domains

- Two types:
  - functional, no relation to context (SH2, glycosylation)
  - Indications of family relationship (cytokine receptor superfamily)
Terminology: Pattern

- Pattern (1)- small motifs

- Pattern (2)- a region containing several motifs and can also contain gaps.
Terminology: Profile

- Profile - position specific matrix built from multiple alignment of group of sequences.
- Different tools are used for each of the above.
We can search databases made of motifs and profiles, or use motifs and profiles to search sequence databases, and in some cases use profiles and motifs to search profile and motif databases.
What do we search with?

- Fixed expressions
- Patterns
- Profiles
Search with: fixed expressions

For example: SHIFRA or IRIT

Advantages: simple, fast searching, Can reduce noise of non-conserved residues

Problems:
1) Demands exact match, no provision for similarity (conservative change)
2) Only some of the information contained in a given protein or domain is used
3) An exact match is not necessarily a true hit, there is no context.
Search with: Patterns

For example: $C-X^{\{1,13\}}-C-[IVML]$ $[ST]-H-[IVML]-[FYW]-[RK]-A$

Advantages:
- More information, more likely to find distant matches

Disadvantages:
- More “noise”, may add irrelevant sequences
- Some context, but still demands exact matches
Search with: Profiles

Example: Position Specific Scoring Matrix

Advantages:
1) Profile searches include maximal information.
2) Use most rigorous algorithms

Problems:
1) Slow searches due to rigor. Demands powerful computer, lots of computer time
2) If a mistake enters the profile, may end up with irrelevant data.
Position Specific Scoring Matrix (PSSM)

- Specific for each family of sequences
- A matrix of vectors of the size 20 x the sequence length

- Many methods exist for deriving them
Advantages of PSSM

- Weights sequence according to observed diversity specific to the family of interest
- Minimal assumptions
- Easy to compute
- Can be used in comprehensive evaluations

**Position Specific Scoring Matrix**

- **PSSM** can be used to search against sequence or a group of sequences (db) for the location(s) of motif(s) represented by the PSSM.

- It is important that the **PSSM** represent the **expected motifs** (sites) as best as possible.

- When producing a **PSSM**, the larger the number of the sequences in the alignment, the greater guarantee that the PSSM will have the best representation of the motif.
Position Specific Scoring Matrix

• If the dataset used in building the PSSM is small, then unless the motif has almost identical AA in each column, the column frequencies in the motif may not be highly representative of all other occurrences of characters in the motif.

• This means we may miss true hits
What we expect from **Motif or Pattern Analysis Tools**

- Identification of very distant homologs.
- May point to important functional units in a sequence.
- Can be used to “anchor” or break-up a multiple alignment.
- Can generate a database of **motifs**
Steps of search

Initial similarity search of a query Against sequence database

Multiple alignment of the hits

Derivation of a motif/pattern/profile

Pattern/profile database search

To be considered

Borders of the query segment.
Scoring matrices.
Gap penalties.
Filters.
Choice of the databases.

Weighting sequences
Weighting positions
Types of search

- Search with a sequence against motif database
- Search with a pattern against a sequence database
- Build your own PSSM
- Search against a database of PSSMs
- Search with a PSSM against a database
Search with a sequence to find motifs

- The simplest search is to use a single sequence, and search against a database of motifs.
- This kind of search is very fast, but does not provide any significance estimations.

- An example of Motifs:
  ATP-binding: [AG]XXXXGK[ST]
  Phosphorylation site: [ST]X[RK]
Search with a Consensus Pattern

- A consensus pattern - a string of characters, where characters at certain positions are “conserved” and are separated by “unimportant” positions.

- This type of searching, where important positions are filtered, is successful in finding distantly related sequences.
Size is important...

<table>
<thead>
<tr>
<th>Name</th>
<th>Database</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>IRIT</td>
<td>swissprot</td>
<td>1,509</td>
</tr>
<tr>
<td>SARA</td>
<td>swissprot</td>
<td>3,661</td>
</tr>
<tr>
<td>AVIAD</td>
<td>swissprot</td>
<td>348</td>
</tr>
<tr>
<td>RACHEL</td>
<td>swissprot</td>
<td>0</td>
</tr>
<tr>
<td>SHIFRA</td>
<td>swissprot</td>
<td>3</td>
</tr>
<tr>
<td>IRIT</td>
<td>uniprot</td>
<td>30,300</td>
</tr>
<tr>
<td>SARA</td>
<td>uniprot</td>
<td>100,483</td>
</tr>
<tr>
<td>AVIAD</td>
<td>uniprot</td>
<td>7,608</td>
</tr>
<tr>
<td>RACHEL</td>
<td>uniprot</td>
<td>84</td>
</tr>
<tr>
<td>SHIFRA</td>
<td>uniprot</td>
<td>45</td>
</tr>
</tbody>
</table>

[ST]-H-[IVML]-[FYW]-[RK]-A in swissprot 38
[ST]-H-[IVML]-[FYW]-[RK]-A in uniprot 1,370
Pattern length

- The choice of pattern length is very important for database searches, and it should be chosen carefully to enable the program used to give the best results.

- The use of logical operators are also important for pattern searches, because they can change the results.

For example: enabling the use of mismatches in the pattern searched.
FindPatterns in GCG

- The program used for pattern searching in GCG is **findpatterns**.
- The program can be used with a single sequence or group of sequences (e.g. a database).
- By default, the program will look for a perfect match but the user can also use mismatches.
FindPatterns in GCG

- Findpatterns can read patterns from the keyboard, or from a file.

- You can search for several patterns simultaneously.

- You can use consensus sequence. For example: \( \text{FCT}(V,I) \times \{2,10\}CA \)
Motif Databases

- There are motifs databases such as Prosite, Prints, Pfam, Smart etc.

- Functional sites of Protein families are stored in these databases. Usually the database provides an excellent description for the motifs.
DNA Motifs

- There are several databases with transcription factor binding sites

- The two biggest commercial databases are Transfac and MatBase (Genomatix)

- The largest free collection is JASPAR
DNA Motifs

- TransFac Database - is a database of eukaryotic cis-acting regulatory DNA elements and trans-acting factors. It covers the whole range from yeast to human.

- TransFac contains the following data types: Sites, Consensus patterns, and Matrices

http://www.gene-regulation.com
Protein Motif Databases

- There are many, some of the most commonly used are:
  - Prosite
  - Pfam
  - Prints
  - Smart
  - Interpro (a metaserver)
ProSite Database of Proteins families & domains

PROSITE is a method of determining what is the function of uncharacterized proteins translated from genomic or cDNA sequences.

PROSITE database consists of biologically significant sites and patterns formulated in such a way that with appropriate computational tools it can rapidly and reliably identify to which known family of proteins, (if any), the new sequence belongs.

http://www.expasy.org/prosite
Prosite Database

- Prosite has patterns, matrices (profiles), and can group related patterns or profiles into one family.
- You can input either a sequence to search against the whole database of patterns and profiles, or choose one of the patterns or profiles to search against a database.
- You can filter the hits to exclude the most common hits (usually a result of random chance).

Release 20.72, of 04-May-2011 (1612 documentation entries, 1308 patterns, 928 profiles and 920 ProRule)
PFAM database

- Pfam is a collection of multiple proteins alignments and profiles built using HMMs (Hidden Markov Models).
- Pfam is divided into 2 sections:
  - PfamA – set of manually curated and annotated models (seeds)
  - PfamB – fully automated models create from alignments generated by ADDA - automatic protein clustering of UniProt.
PFAM database

- The database is organized into families (clans) and subfamilies
- Information on domain organization and taxonomic distribution is available for each family

- Pfam version 25 from March 2011, has 12273 families.

http://www.pfam.org
Prints Database

PRINTS is a compendium of protein fingerprints. A fingerprint is a group of conserved motifs used to characterize a protein family.

Usually the motifs do not overlap, but are separated along a sequence, though they may be contiguous in 3D-space.

Fingerprints are observed in sequence alignments; taken together, the motifs characterize the aligned family and hence provide a specific diagnostic signature.
Prints Database

**Fingerprints** thus derive much of their potency from the biological context afforded by matching multiple motifs; this makes them at once more flexible and more powerful than single-motif approaches.

The technique further departs from other pattern-matching methods by readily allowing the creation of discriminators at super-family, family and sub-family-specific levels.
Prints Database

Prints, (version 41 from 8/10) includes 2050 fingerprints, encoding 12,121 motifs, covering a range of globular and membrane proteins, modular polypeptides and so on.

The PRINTS-S database models relationships between families, including those beyond the reach of conventional sequence analysis approaches.

The database is accessible for BLAST, fingerprint and text searches at: http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/
SMART database
(Simple Modular Architecture Research Tool)

- SMART is based on curated HMMs models of multiple proteins alignments of representative members of protein families found with PSI-Blast.

- Once a model is created it is being used to search the databases for additional family members. When found, these additions are entered to the multiple alignment and a new HMM is built.
SMART database

- **Genomic SMART** contains the proteomes of completely sequenced genomes.

- SMART is accessible from:
  
  http://smart.embl-heidelberg.de/
InterPro Database

**InterPro** is a meta-database of protein domains and functional sites, that combines the search strategies of several signature-recognition methods for best results.

These various methods address different sequence analysis problems, resulting in rather different and, for the most part, independent databases.

Diagnostically, each method has different areas of optimum application owing to the different strengths and weaknesses of their underlying analysis methods.
InterPro Database

InterPro (The InterPro Consortium 2001) is a collaborative project aimed at providing an integrated layer on top of the most commonly used signature databases by creating a unique, non-redundant characterisation of a given protein family, domain or functional site.
InterPro Database

InterPro data is distributed in XML format and it is freely available under the InterPro Consortium copyright.

The InterPro project home page is available at http://www.ebi.ac.uk/interpro

The current version (32) of InterPro (April 2011) contains 21516 entries.
InterPro Data Sources
Sequence Logos

- Sometimes it helps to have a visual representation of a multiple alignment to help identify critical residues
- Sequence Logos are one format that is easy to build, and easy to understand

http://weblogo.berkeley.edu
http://weblogo.threeplusone.com
De Novo Motif Definition

- Block Maker
- MEME
- HMMER
- and many more...
PSI-BLAST

- POSITION-SPECIFIC ITERATED
- Runs one round of gapped-Blast, and then builds a PSSM
- The PSSM is used as the input for the following rounds of Blast
Producing the PSSM

- PSSM equals the length of the query sequence
- All database segments with an E score of less than 0.005 are taken for the multiple alignment
- The query sequence is the template for the alignment
- Identical sequences are discarded
Producing the PSSM

- One copy of sequences with more than 98% identity to each other is used.
- Gaps are ignored in the alignment, and treated as an independent character in the alignment weighting (no additional penalty).
- Reduce the size of the matrix per base to only those columns that are contained in all rows.
Producing the PSSM
Producing the PSSM

- Can have different numbers of sequences in each row

- Weights are calculated over the whole alignment, gaps are counted as an independent character, Columns with identical bases are ignored in the weight calculation
Iteration

- PSI-Blast continues until no new proteins with E-value of less than 0.005 are found.
- Adds the new sequences in each round to the PSSM.
- User has the choice to manually edit (force sequences in or out) the input to the alignment.
PHI-Blast

- Pattern Hit Initiated
- Uses a pattern as an input sequence
- Output can be used as an input for PSI-Blast
“PHI-BLAST helps answer the question:

What other protein sequences both contain an occurrence of P and are homologous to S in the vicinity of the pattern occurrences?

PHI-BLAST may be preferable to just searching for pattern occurrences because it filters out those cases where the pattern occurrence is probably random and not indicative of homology.”
Blocks database and tools

- Blocks are multiply aligned **ungapped** segments corresponding to the most highly conserved regions of proteins.
- The Blocks web server tools are: Block Searcher, Get Blocks and Block Maker. These are aids to detection and verification of protein sequence homology.
- They compare a protein or DNA sequence to a database of protein blocks, retrieve blocks, and create new blocks, respectively.
The BLOCKS web server

http://blocks.fhcrc.org/

http://bioinfo.weizmann.ac.il/blocks

Refs: Henikoff S and Henikoff JG
The Block Maker Tool

Block Maker finds conserved blocks in a group of two or more unaligned protein sequences, which are assumed to be related.

The input file must contain at least 2 sequences.

Input sequences must be in FastA format.
The Block Maker tool

This program uses different algorithms to create regions of local alignment. There are two steps to the program:

The first step finds candidate alignments. This is done using two different algorithms.

The best alignments from both methods are passed on to the second step, an algorithm called MOTOMAT (Henikoff 1991)
The Block Maker tool

This algorithm extends the alignments, scores them, and then sorts them in such a way that a best set ("best path") is chosen.

Motomat will not attempt to realign sequences that don’t fit - it discards them.

At the end it produces two sets of blocks, one for each of the original alignment methods.
**BLOCKS from MOTIF**

>dynamin MX3_RAT P18590 rattus norvegicus (rat). interferon... family
7 sequences are included in 4 blocks

dynaminA, width = 30

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Position</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>DYN1_HUMAN</td>
<td>24</td>
<td>GQNADLDLPQIAVGGQSGKSSVLENFVG</td>
</tr>
<tr>
<td>MGM1_YEAST</td>
<td>168</td>
<td>SSSAHLTLPSSIVGQSSGKSSVLESIVG</td>
</tr>
<tr>
<td>MX1_ANAPL</td>
<td>137</td>
<td>GIEKDLSDLPAIAVIGDQSSGKSSILEALSG</td>
</tr>
<tr>
<td>MX2_HUMAN</td>
<td>140</td>
<td>GVEQDLALPAIAVIGDQSSGKSSVLEALSG</td>
</tr>
<tr>
<td>MX3_RAT</td>
<td>61</td>
<td>GVEQDLALPAIAVIGDQSSGKSSVLEALSG</td>
</tr>
<tr>
<td>MX_SHEEP</td>
<td>58</td>
<td>GVEQDLALPAIAVIGDQSSGKSSVLEALSG</td>
</tr>
<tr>
<td>VPS1_YEAST</td>
<td>27</td>
<td>GSQSPIDLIQITVGVQSAGKSSVLENIVG</td>
</tr>
</tbody>
</table>

**BLOCKS from GIBBS**

>dynamin MX3_RAT P18590 rattus norvegicus (rat). interferon... family
6 sequences are included in 6 blocks

dynaminA, width = 30

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Position</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>DYN1_HUMAN</td>
<td>24</td>
<td>GQNADLDLPQIAVGGQSGKSSVLENFVG</td>
</tr>
<tr>
<td>MX1_ANAPL</td>
<td>137</td>
<td>GIEKDLSDLPAIAVIGDQSSGKSSILEALSG</td>
</tr>
<tr>
<td>MX2_HUMAN</td>
<td>140</td>
<td>GVEQDLALPAIAVIGDQSSGKSSVLEALSG</td>
</tr>
<tr>
<td>MX3_RAT</td>
<td>61</td>
<td>GVEQDLALPAIAVIGDQSSGKSSVLEALSG</td>
</tr>
<tr>
<td>MX_SHEEP</td>
<td>58</td>
<td>GVEQDLALPAIAVIGDQSSGKSSVLEALSG</td>
</tr>
<tr>
<td>VPS1_YEAST</td>
<td>27</td>
<td>GSQSPIDLIQITVGVQSAGKSSVLENIVG</td>
</tr>
</tbody>
</table>
ID  ADH_IRON_1; BLOCK
AC  BL00913C; distance from previous block=(56,76)
DE  Iron-containing alcohol dehydrogenases proteins.
BL  HHG motif; width=22; seqs=11; 99.5%=492 strength =1428

---

ADHE_CLOAB (720) CHSMAIKLSSEHNIPSGIANAL 66
FUCO_ECOLI (262) VHGMAHPLGAFYNTPHGVANAI 44
GLDA_BACST (259) HNGFTALEEGIHLHTHGEKVAF 100
GLDA_ECOLI (269) VHNGLTAIPDAHHYYHGEKVAF 100
MEDH_BACMT (259) VHSISHQVGVVYKLOHQGCNSV 78
ADH1_CLOAB (258) CHSMAHTGAVFHIPHGCANAI 47
ADHE_ECOLI (721) CHSMAHKLGSQFHIPHGANAL 47
ADH2_ZYMMO (261) VHAMEHQLGYYNLPHGVCNAV 36
ADH4_YEAST (263) VHALAHQLGGFYHLPHGVCNAV 41
ADHA_CLOAB (266) CHPMEHELCSAYDITHGVGLAI 50
ADHB_CLOAB (266) VHLMEHELCSAYDITHGVGLAI 49