

APPLICATIONS OF MULTIPLE ALIGNMENT

MOTIFS, PATTERNS and DOMAINS

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.



Database searching doesn't always find what we want.....

- The tools used for similarity searches (e.g BLAST, FASTA) are known to miss 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.....

Even using advanced methods such as a full Smith-Waterman, more distantly related, though biologically relevant sequences are often missed, due to the requirement for overall 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?

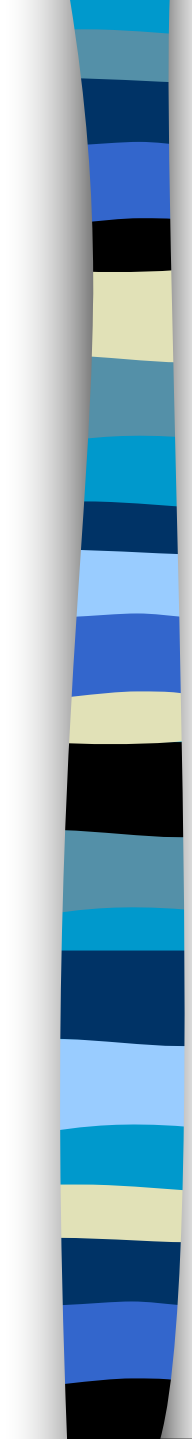
Historically, these protein families were found by looking for functionally related sequences, either in the same species or in others.

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 (for example, protein families).
 - Tools like motif, pattern or profile 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 Blosom 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)



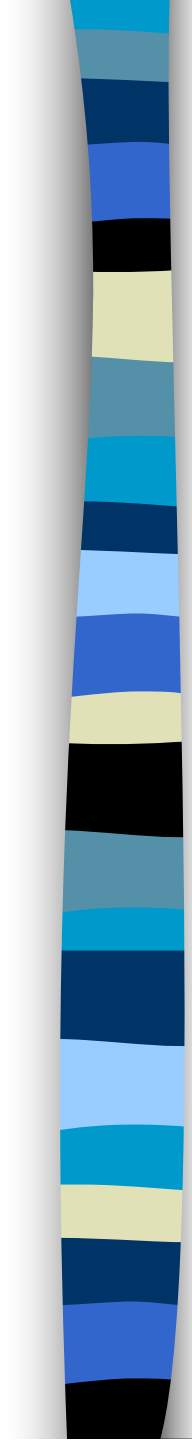
Terminology: Pattern

- 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

Henikoff and Henikoff (1994) J.Mol Biol. 243:574-578



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

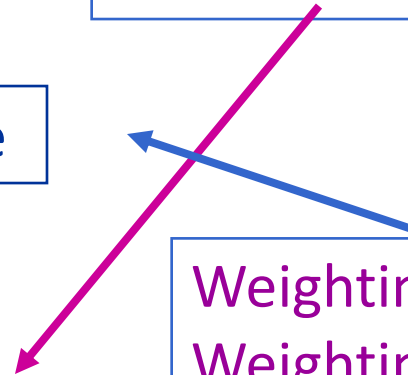


Pattern/profile database search

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

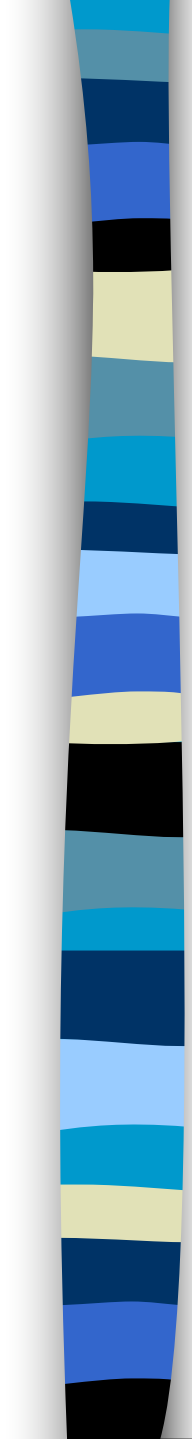
[ST] X [RK]



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...



IRIT	swissprot	1,615
SARA	swissprot	3,972
AVIAD	swissprot	377
RACHEL	swissprot	0
SHIFRA	swissprot	3

IRIT	uniprot	538,375
SARA	uniprot	2,140,405
AVIAD	uniprot	160,539
RACHEL	uniprot	1509
SHIFRA	uniprot	1041

[S]-[T]-[H]-[I]-[V]-[M]-[L]-[F]-[Y]-[W]-[R]-[K]-[A] in swissprot 46
[S]-[T]-[H]-[I]-[V]-[M]-[L]-[F]-[Y]-[W]-[R]-[K]-[A] in uniprot 27,499



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.



FuzzPro/FuzzNuc in EMBOSS

- The EMBOSS package has tools for protein (FuzzPro) and nucleotide (FuzzNuc)
- 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 ambiguity codons, or specify multiple amino acids per position.



Motif Databases

- There are motif databases such as Prosite, Pfam, Smart, Prints 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



RNA motifs

Rfam is the largest collection

(Release 14.8, May 2022, 4094 families)

Latest updates include:

- microRNA families
- Viral RNA families

<http://rfam.xfam.org>



Other Nucleotide Motifs

- UTRdb: sites in UTRs with functional significance
- Splice signal databases
- Others.....



Protein Motif Databases

- There are many, some of the most commonly used are:
 - Prosite
 - Pfam
 - Prints
 - Smart
 - Interpro (a metasever)
 - CDD (database + some others)



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 2022_02, of 25-May-2022 (1906 documentation entries, 1311 patterns, 1341 profiles and 1356 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 33.1 from May 2020, has 18259 families. (now on Pfam 35)

<http://pfam.xfam.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 42 from 2/12) includes 2156 fingerprints, encoding 12,444 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/>

CDD

The Conserved Domain Database is a resource for the annotation of functional units in proteins. Its collection of domain models includes a set curated by NCBI, which utilizes 3D structure to provide insights into sequence/structure/function relationships.



CDD uses 3D structural information to build better definitions of the motifs, with gaps allowed in between structural elements.

<https://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml>



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) 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 (89) of **InterPro** (May 2022) contains 40,124 **entries**.

InterPro Data Sources



CDD (NCBI)



HAMAP



ProSite



PIRSF



Pfam



Gene3D



PRINTS



SFLD



PANTHER



TIGRFAM



SMART



Superfamily



CATH-Gen3D



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>





PSI-BLAST

- **P**osition-**S**pecific **I**terated
- Runs one round of gapped-Blast, and then builds a PSSM
- The PSSM is used as the input for the following rounds of Blast
- Ref: Altschul et al (1997) Nucleic Acids Research 25 (17) 3389-3402



Producing the PSSM

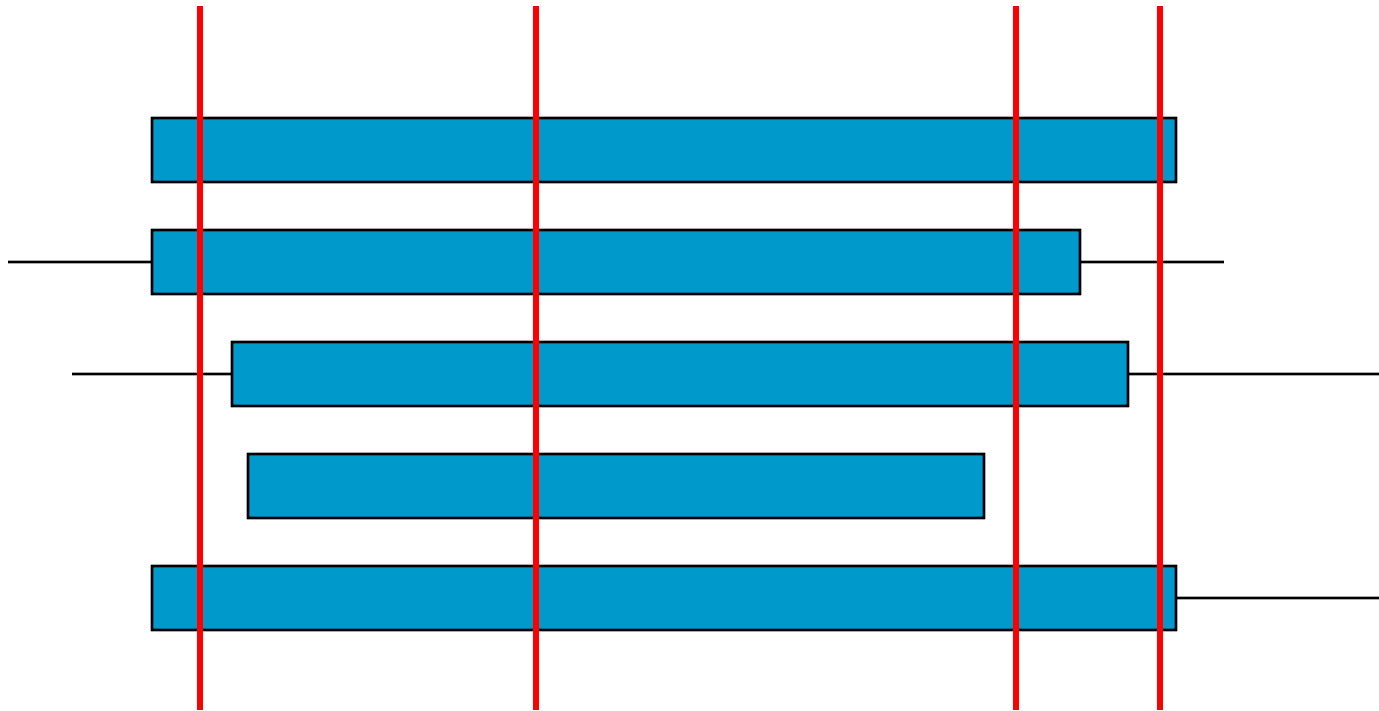
- PSSM equals the length of the query sequence
- All database segments with an E score of less than 0.005 are taken for the multiple alignment
- The query sequence is the template for the alignment
- Identical sequences are discarded



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.”



De Novo Motif Definition

- MEME
- HMMER
- Dilimot
- and many more...