In the Name of God, the Merciful, the Compassionate

Introduction to Bioinformatics
08 - Protein Motif and Domain Prediction

Instructor: Hossein Zeinali

Amirkabir University of Technology

Motifs and Domains

- Often achieving significant similarity of proteins with database sequences of known functions over their entire length is difficult.
  - Solution: identification of short consensus sequences related to known functions.
  - These consensus sequence patterns are termed *motifs* and *domains*.
- A motif is a short conserved sequence pattern associated with distinct functions of a protein or DNA.
  - It is often associated with a distinct structural site performing a particular function. Example: *Zn-finger* motif with ten to twenty amino acids.

Motifs and Domains (Cont.)

- A *domain* is also a conserved sequence pattern, defined as an independent functional and structural unit.
  - Are normally longer than motifs with an average length of 100 residues.
  - A domain may or may not include motifs within its boundaries.
- Motifs and domains are evolutionarily more conserved than other regions of a protein.
- The identification of motifs and domains in proteins is an important aspect of the classification of protein sequences and functional annotation.

Motifs and Domains (Cont.)

- Because of evolutionary divergence, functional relationships between proteins often cannot be distinguished through simple BLAST or FASTA database searches.
- Identification of motifs and domains heavily relies on multiple sequence alignment as well as profile and hidden Markov model (HMM) construction.
- Motifs and domains serve as diagnostic features for a protein family.
  - The consensus sequence information of motifs and domains are stored in a database for later searches.

Approaches to Representing Motifs and Domains

- Reduce the MSA from which motifs or domains are derived to a consensus sequence pattern, known as a *regular expression*.
  - Example: protein phosphorylation motif can be expressed as [ST]-X-[RK].
- Use a statistical model such as a *profile or HMM* to include probability information.

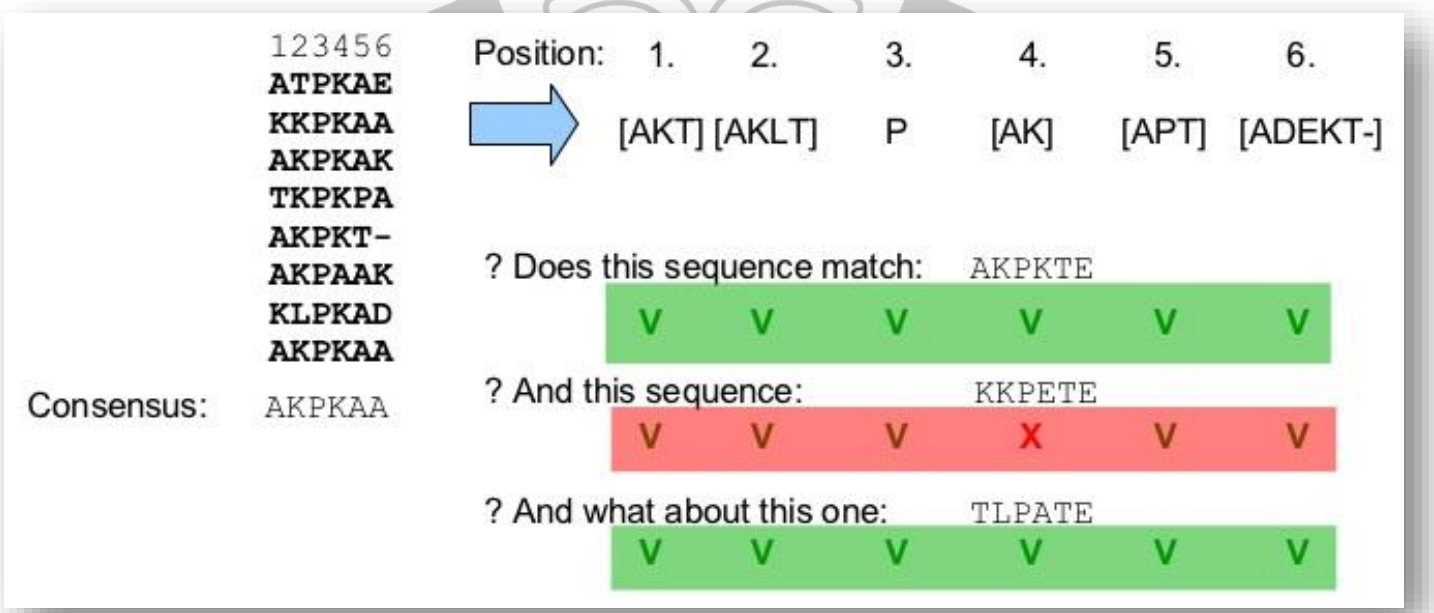
Regular Expressions

- A regular expression is a concise way of representing a sequence family by a string of characters. The used rules:
  - Single conserved amino acid residue in a position => one letter code.
  - Multiple alternative conserved residues => included residues are placed within brackets [].
  - If the position excludes certain residues => excluded residues are placed in curly braces {}.
  - *X* is used to indicate all possible residues in a given position.
  - If a sequence element within the pattern is repetitive => the number of pattern repetitions is indicated within parentheses: (n) or (n, m)
  - Each position is linked by a hyphen.
  - Example:

E-X(2)-[FHM]-X(4)-{P}-L

Matching Regular Expressions with a Query

- Exact matching:
  - There must be a strict match of sequence patterns
- Fuzzy matching also called *approximate matches*:
  - Allows more flexible matching of residues of similar biochemical properties.



Example

ADLGAVFALCDRYFQ

SDVGPRSCFCERFYQ

ADLGRTQNRCDRYYQ

ADIGQPHSLCERYFQ

- Regular Expression:

[AS]-D-[IVL]-G-X(4)-{PG}-C-[DE]-R-[FY](2)-Q

Create RE Database

Sequence alignment

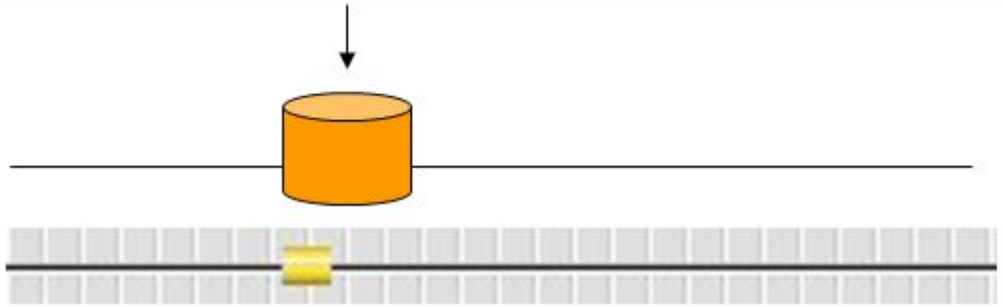
Define pattern

Extract pattern sequences

Build regular expression

C-C-{P}-x(2)-C-[STDNEKPI]-x(3)-[LIVMFS]-x(3)-C

Pattern signature



Motif & Domain Databases

- Based on regular expressions:
  - Prosite (Interpro includes Prosite, PRINTS, etc)
  - Emofit

  Limitation: these don't take probability info into account

- Based on statistical models:
  - PRINTS
  - BLOCKS
  - ProDom
  - Pfam
  - SMART
  - CDART
  - Reverse PsiBLAST

PROSITE

- The first established sequence pattern database
  - www.expasy.ch/prosite/
- It primarily uses a single consensus pattern or “sequence signature” to characterize a protein function and a sequence family.
- Patterns are derived from conserved regions of protein
  - Represented with regular expressions
- For searching, it uses exact matches to the sequence patterns.
- The database also constructs profiles to complement some of the sequence patterns.
- The major pitfall: some of the sequence patterns are too short to be specific.
  - The resulting match is very likely to be a result of random events.
- The database is relatively small and has a greater than 20% error rate.

Emotif

- Emotif is a motif database
  - Uses multiple sequence alignments from both the BLOCKS and PRINTS databases.
  - Has a alignment collection much larger than PROSITE.
  - http://motif.stanford.edu/emotif/emotif-search.html
- It identifies patterns by allowing fuzzy matching.
  - Produces fewer false negatives than PROSITE.

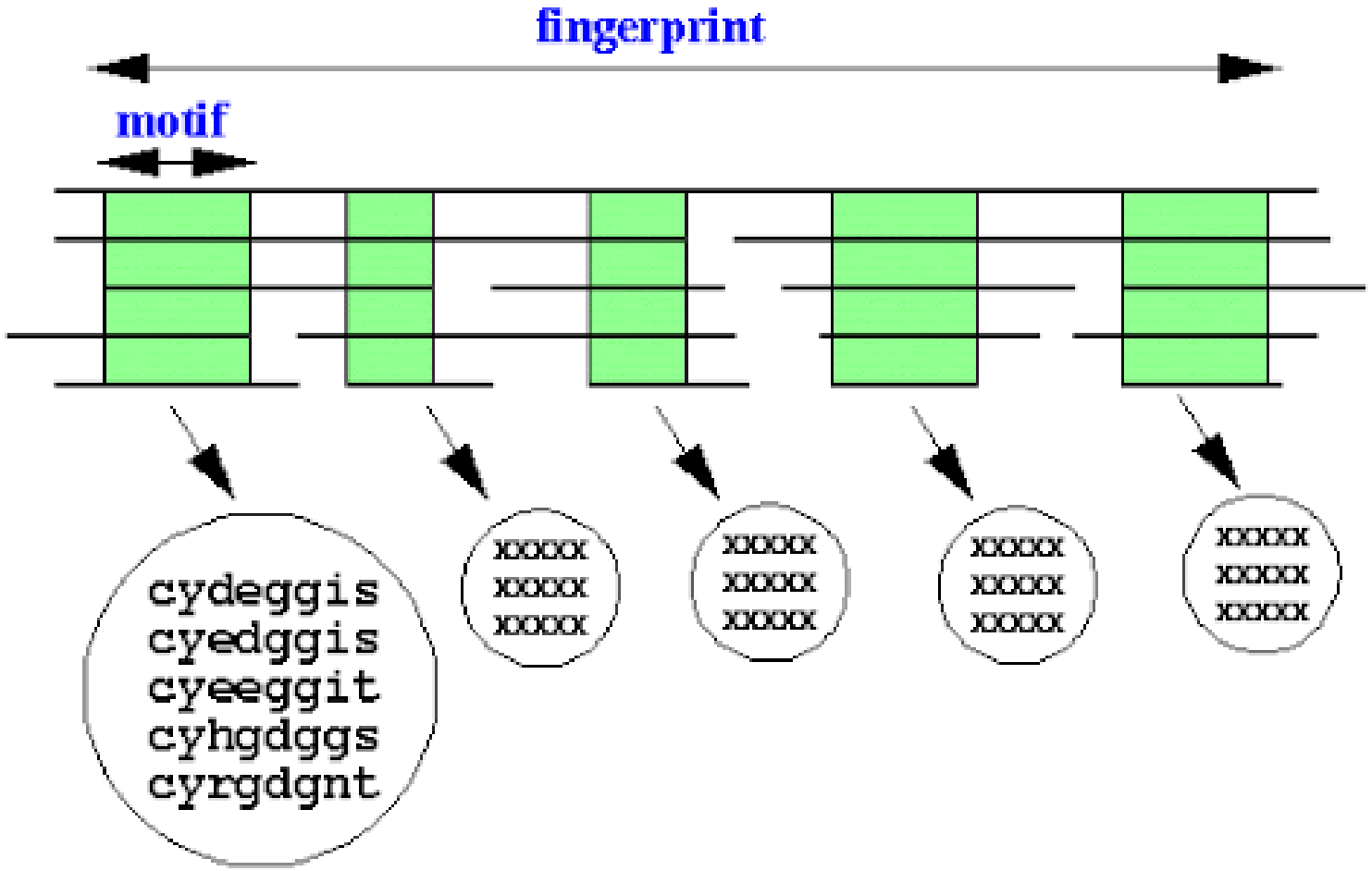
Using Statistical Models

- The major limitation of regular expressions is that this method does not take into account sequence probability information.
  - If a regular expression is derived from an incomplete sequence set, it has less predictive power.
- Unlike regular expressions, PSSMs, profiles, and HMMs preserve the sequence information from a MSA and express it with probabilistic models.
  - They allow partial matches and compensate for unobserved sequence patterns using pseudocounts.
  - Have stronger predictive power than the regular expression.

PRINTS

- PRINTS is a collection of so-called **fingerprints**:
  - Contains ungapped, manually curated alignments corresponding to the most conserved regions among related sequences.
  - A *fingerprint* is a group of *conserved motifs* taken from a multiple sequence alignment.
  - http://bioinf.man.ac.uk/dbbrowser/PRINTS/
- Drawbacks:
  - The difficulty to recognize short motifs when they reach the size of single fingerprints.
  - Relatively small database, which restricts detection of many motifs.

PRINTS (Cont.)


fingerprint
motif
cydeggis
cyedggis
cyeeggit
cyhgdggs
cyrgdgnt
xxxxx
xxxxx
xxxxx

BLOCKS

- BLOCKS is a database that uses MSA derived from the most conserved, ungapped regions of homologous proteins.
  - Automatically generated using the same data sets used for deriving the BLOSUM matrices.
  - The derived ungapped alignments are called *blocks*.
    - The blocks, which are usually longer than motifs, are converted to PSSMs.
    - A weighting scheme and pseudocounts are applied to the PSSMs.
  - http://blocks.fhcrc.org/blocks
- Blocks often encompass motifs
  - the functional annotation of blocks is consistent with that for the motifs.

Pfam

- Pfam is a comprehensive database with protein domain alignments and families derived from sequences in SWISSPROT and TrEMBL.
  - Each motif or domain is represented by an HMM.
  - http://pfam.wustl.edu/hmmsearch.shtml
- The Pfam database is composed of two parts:
  - *Pfam-A*: involves manual alignments
  - *Pfam-B*: involves automatic alignment in a way similar to ProDom.
    - Only sequence families not covered in Pfam-A
- Each family is represented by 2 MSAs and 2 profile-HMMs.
- Pfam 32.0 was released in September 2018 and contains 17,929 families.

Other Databases

- ProDom:
  - Is a domain database generated from sequences in the SWISSPROT and TrEMBL databases.
  - The domains are built using recursive iterations of PSI-BLAST.
- SMART:
  - Contains HMM profiles constructed from manually refined protein domain alignments.
  - Alignments are further checked and refined by human annotators before HMM profile construction.
  - Protein functions are also manually curated.
- InterPro:
  - Integrates information from PROSITE, Pfam, PRINTS, ProDom, and SMART databases.
  - Included only overlapping motifs and domains in all five databases.

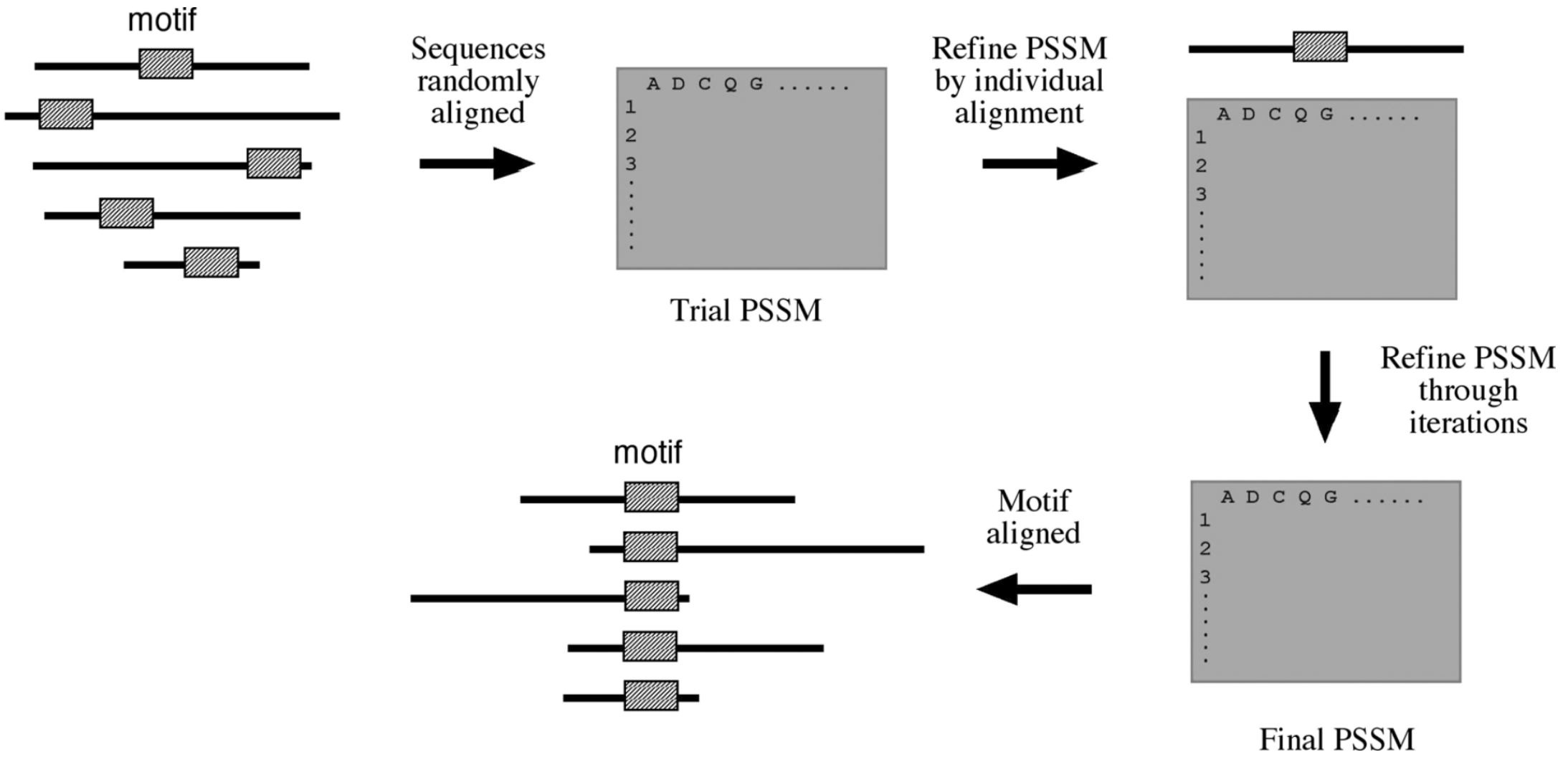
Protein Family Databases

- In addition to databases of "related" protein sequences, based on shared motifs or domains (Pfam, BLOCKS, CDART), some databases "cluster" sequences into families based on near full-length sequence comparisons
- *COGs* - Clusters of Orthologous Groups (at NCBI)
  - Mostly Prokaryotic sequences
  - KOG = newer Eukaryotic version
  - COGnitor - software to search database
- ProtoNet - also clusters of homologous protein sequences
  - Advantages: tree-like hierarchical structure
    - Provide GO (gene ontology) annotations
    - Provides InterPro keywords

Motif Discovery in Unaligned Sequences

- For a set of closely related sequences, commonly shared motifs can be discovered by using the MSA–based methods.
  - Distantly related sequences that share common motifs cannot be readily aligned.
- **Expectation Maximization** – generate "random" alignment of all sequences, derive a trial PSSM, iteratively match individual sequences to trial PSSM to edit & improve it
  - Problems? Can hit a local optimum (premature convergence)
    Sensitive to initial alignment
  - MEME - Multiple EM for Motif Elicitation - modified EM, avoids local optimum issues; two step procedure.
    - http://meme.sdsc.edu/meme/website/meme-intro.html

Schematic diagram of the EM algorithm



Gibbs Motif Sampling

- EM can get trapped in local minima
  - One approach to alleviate this limitation: try different (perhaps random) initial parameters
- Gibbs sampling exploits randomized search to a much greater degree
- In theory, Gibbs sampling is less susceptible to local minima than EM.
- The Gibbs sampling algorithm makes an initial guessed alignment of all but one sequence. Motif length should be defined as a parameter *w*.
- A trial PSSM is built to represent the alignment.
- The matrix is then aligned to the left-out sequence.
- The matrix scores are subsequently adjusted to achieve the best alignment with the left-out sequence.
- Gibbs sampler: http://bayesweb.wadsworth.org/gibbs/gibbs.html

References

- Mostly used:
  - Essential bioinformatics, Chapter 7 (Protein Motif and Domain Prediction)

Thanks for your attention