Chapter 7

Protein Motif and Domain Prediction

Overview

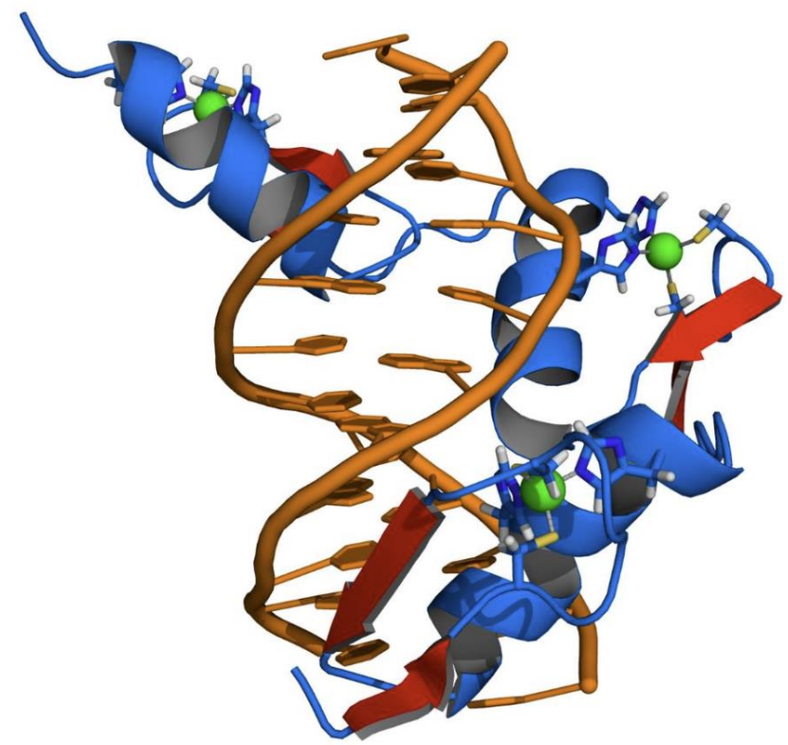
1. Introduction
2. Introduction to Biological Databases
3. Pairwise Sequence Alignment
4. Database Similarity Searching
5. Multiple Sequence Alignment
6. Profiles and Hidden Markov Models
7. Protein Motifs and Domain Prediction
8. Gene Prediction
9. Promoter and Regulatory Element Prediction
10. Phylogenetics Basics
11. Phylogenetic Tree Construction Methods and Programs
12. Protein Structure Basics
13. Protein Structure Visualization, Comparison and Classification
14. Protein Secondary Structure Prediction
15. Protein Tertiary Structure Prediction
16. RNA Structure Prediction
17. Genome Mapping, Assembly and Comparison
18. Functional Genomics
19. Proteomics

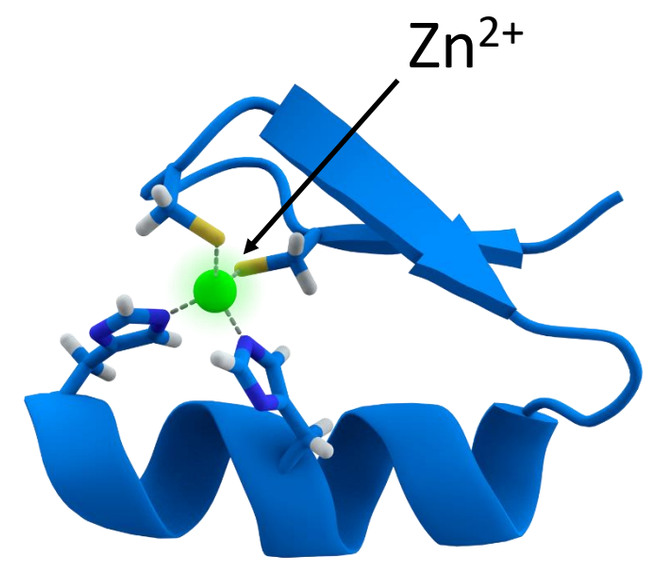
Motifs

Motif ≡ <u>short conserved sequence pattern</u> associated with distinct functions of a protein or DNA

- Often associated with a distinct structural site performing a particular function, *e.g.* a metal-binding motif or a catalytically active site
- Typically ten to twenty amino acids long
- *E.g.* Zn-binding *Zn-finger* motif:

  X_2-Cys-$X_{2,4}$-Cys-X_{12}-His-$X_{3,4,5}$-His



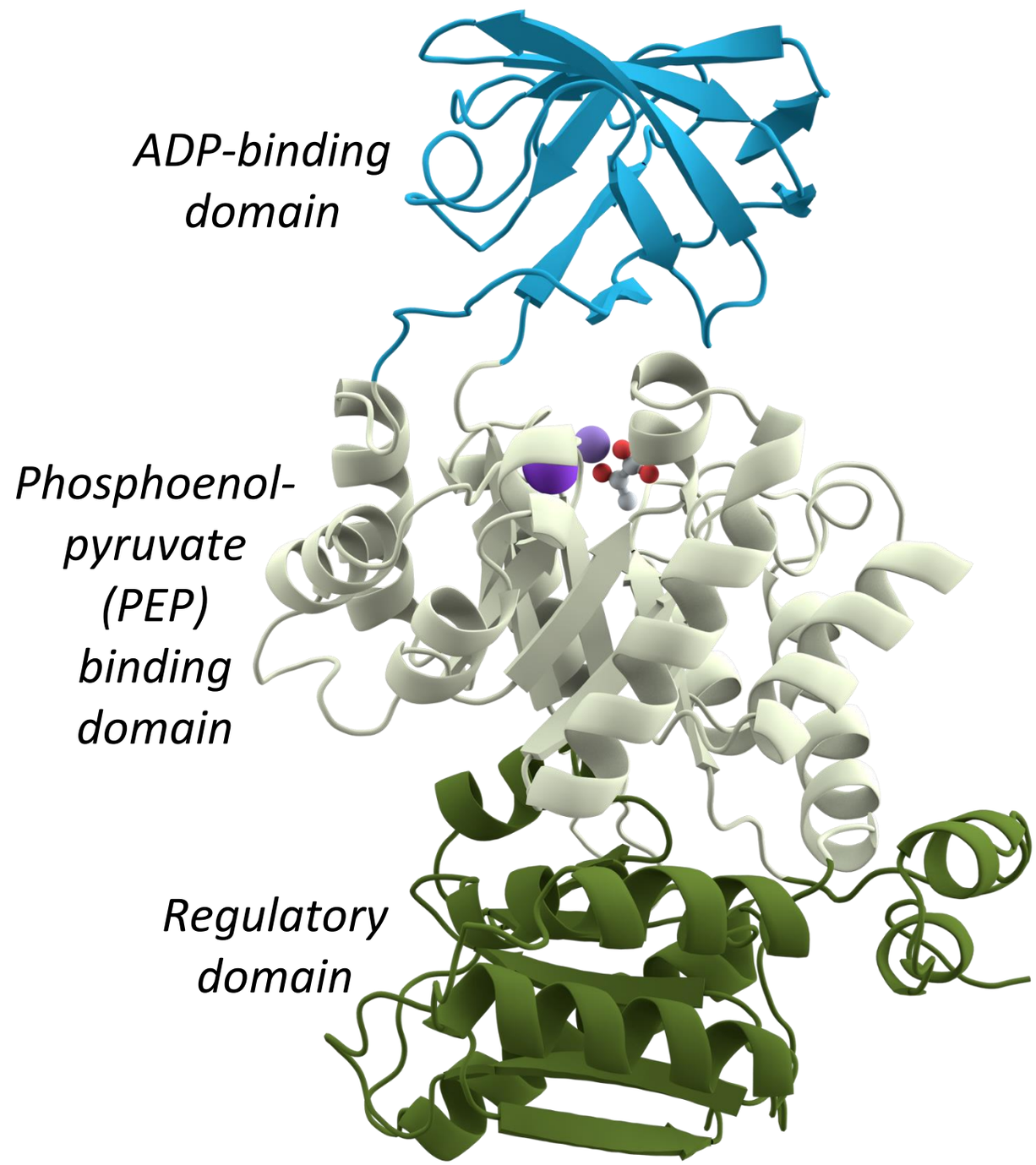


Domains

Domain ≡ conserved sequence pattern that forms an <u>independent functional and structural unit</u>

- Larger than motifs: typically 40 to 700 residues, average length 100 residues
- May include one or several motifs
- *E.g.* transmembrane domains, catalytic domains, ligand-binding domains

Example: pyruvate kinase



Identifying/predicting motifs and domains

- Why? Identification is important for classification of protein sequences and functional annotation
- However: motifs and domains usually cannot be distinguished through simple BLAST or FASTA database searches and pairwise alignment
- Identification requires more information, *i.e.* multiple sequence alignment, and ideally profiles or hidden Markov models (HMMs)

Domains are the objects of nature's tinkering

"Evolution is a tinkerer, not an engineer"

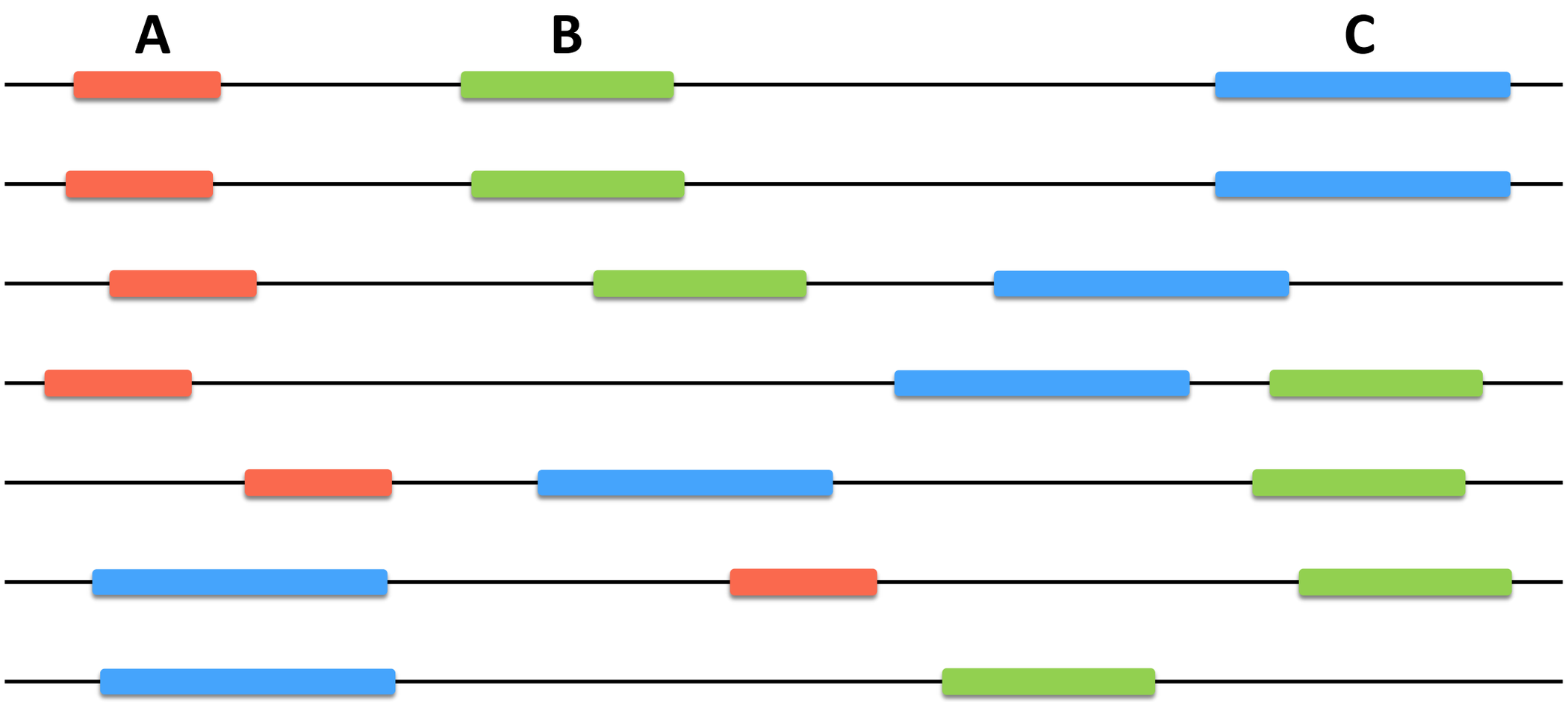
François Jacob,

Evolution and tinkering (1977)

Science **196** 1161-1166

Domain shuffling

- Domains are highly conserved objects in evolution that usually have a well-defined biochemical function (*e.g.* kinase, membrane-binding, ATPase, dimerisation, …)
- Domains therefore <u>tend to evolve as units</u>, which are gained, lost, or shuffled as one module



Identifying motifs and domains

- Commonly conserved regions can be identified by <u>multiple sequence alignment</u>
- Regions considered motifs and domains then serve as "diagnostic features" for a protein family
- Consensus sequence information that defines motifs and domains can be stored in database
- By looking for the presence of sequence patterns, associated functions can be rapidly attributed to a query sequence

Describing or “defining” motifs and domains

Two ways of representing the consensus information:

- Regular expressions: reduction to a <u>consensus sequence pattern</u>
- Statistical models, *i.e.* <u>profiles or HMMs</u>

Regular expressions

- A string of characters represents the sequence family
- A single conserved residue is indicated using the standard one-letter code for amino acids
- Multiple alternative conserved residues are placed within brackets **[ ]**
- Excluded residues at a position are indicate in curly braces **{ }**
- Non-specific residues in a given position are indicated by an **X**
- Repeats are indicated as a number within parentheses **( )**
- Each consecutive position is linked by a hyphen

Example of a regular expression

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

which means:

- E followed by two unspecific residues
- followed by F or H or M
- followed by four unspecific residues
- followed by a non-P residue
- followed by L

Exact matching of regular expressions

- No variations from the predefined patterns allowed
- The query sequence is either a match or a non-match
- High chance of false-negative results
- Has to be updated whenever new sequences of a motif are accumulated

Fuzzy matching of regular expressions

Fuzzy matching = <u>approximate</u> matching

- Also allows residues with similar biochemical properties as the ones specified
- More false positives, especially for short motifs
- Implemented in the *Emotif* database (which no longer exists; no other databases seem to use this approach)

PROSITE

- https://prosite.expasy.org
- The first sequence pattern database established
- Uses regular expressions and exact matching
- Functional information added, based on published literature

Problems:

- Some sequence patterns in PROSITE are too short to be specific (random matches are highly likely)
- Exact matching results in false negatives
- Error rate greater than 20%

Motif/domain databases using statistical models

- Statistical models contain <u>more information</u> and have <u>stronger predictive power</u> than approaches based on regular expressions
- Position-specific scoring matrices (PSSMs) / profiles and HMMs <u>preserve frequency information</u> from a multiple sequence alignment and express it with probabilistic models
- Result: an <u>enhancement of sensitivity</u> of motif discovery and detection of more divergent but truly related sequences

Pfam

- http://pfam.xfam.org
- Protein domain alignments derived from sequences in SWISSPROT and TrEMBL
- Each motif or domain is represented by an HMM
- Higher sensitivity than with approaches based on regular expressions
- Like PROSITE, also contains functional annotations, as well as links to other databases

SMART

- http://smart.embl-heidelberg.de/
- HMM profiles from manually refined protein domain alignments based on tertiary structures whenever available
- Manually curated protein function annotations
- Emphasis on specific (signaling-related, extracellular and chromatin-associated) motifs and domains
- Output contains information with respect to cellular localisation and tertiary structure

Unifying databases: InterPro

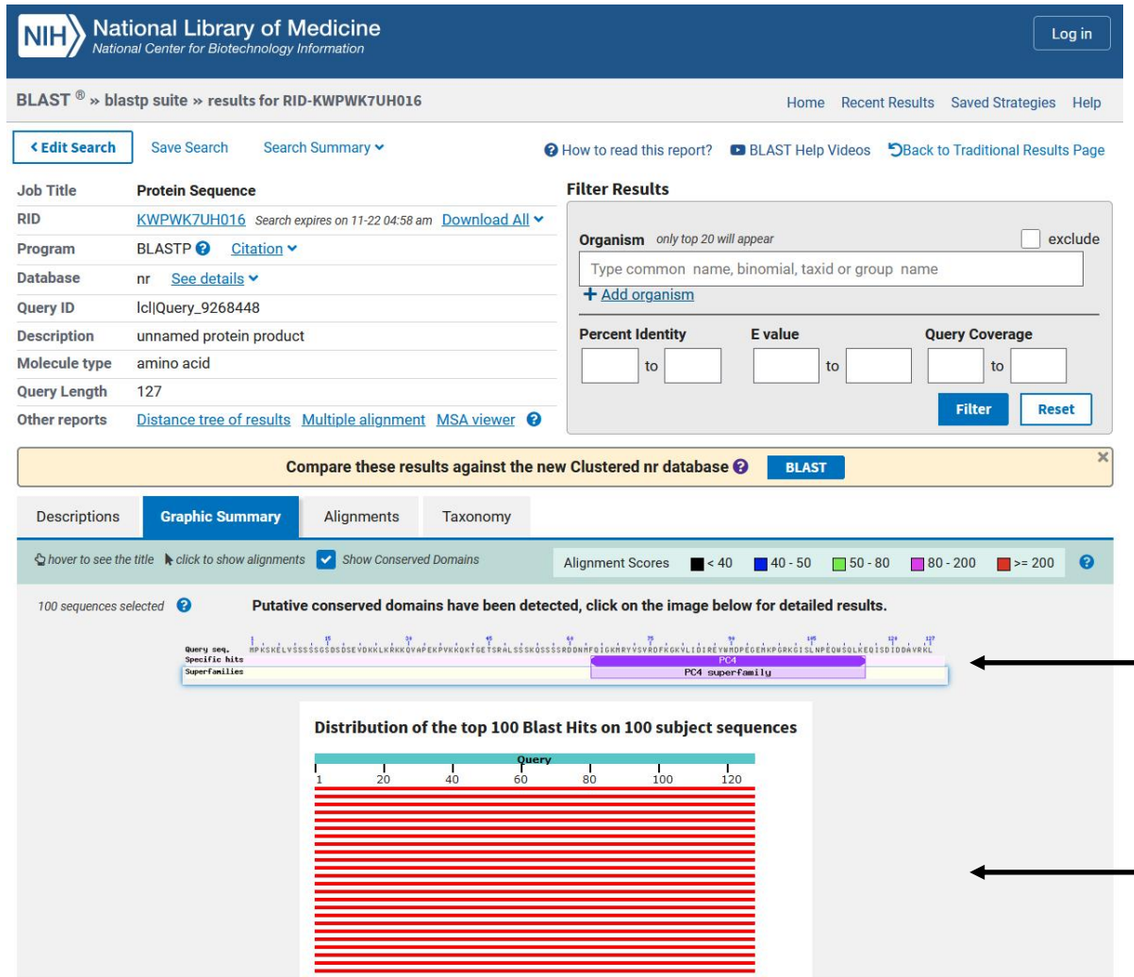
- http://www.ebi.ac.uk/interpro
- Designed to unify multiple databases for protein domains and conserved functional sites
- Integrates information from several other databases (*e.g.* PROSITE, Pfam, SMART)
- Only "overlapping" motifs and domains from all of these databases are included
- Uses a combination of regular expressions, profiles and HMMs in pattern matching

Reverse PSI-BLAST (RPS-BLAST)

- http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi
- Searches a query sequence against a pre-computed profile database using the PSI-BLAST method
- PSI-BLAST searches a profile against a sequence database, hence "reverse"

Domain searching at Entrez (NCBI)

- https://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml
- Uses RPS-BLAST
- Now an integral part of the regular BLAST search function:



Results of automatic domain search

BLAST hits

Domain searching at Entrez (NCBI)

NCBI

Conserved Domains

HOME | SEARCH | GUIDE | NewSearch | Structure Home | 3D Macromolecular Structures | Conserved Domains | Pubchem | BioSystems

Conserved domains on [lcl|Query_3089482]

View Standard Results

Local query sequence

Protein Classification

transcriptional coactivator p15/PC4 family protein(domain architecture ID 10491384)
transcriptional coactivator p15/PC4 family protein is a general coactivator that functions cooperatively with TAFs and mediates functional interactions between upstream activators and the general transcriptional machinery

CATH: 2.30.31.10 | **Gene Ontology:** GO:0003713|GO:0005515|GO:0003697|GO:0003713|GO:0140297|GO:0060261 | **PubMed:** 7628453|15692559|8062392 | **SCOP:** 4001029

Graphical summary

Zoom to residue level | show extra options »

Query seq. MPKSKELVSSSSSGSDSDSEVDKKLKRKKQVAPEKPVKKQKTGETSRALSSSKQSSSSRDDNMFQIGKMRYVSVRDFKGKVLIDIREYWMDPEGEMKPGRKGISLNPEQWSQLKEQISDIDDAVRKL

Specific hits: PC4

Superfamilies: PC4 superfamily

Search for similar domain architectures | Refine search

List of domain hits

Name	Accession	Description	Interval	E-value
[+] PC4	pfam02229	Transcriptional Coactivator p15 (PC4); p15 has a bipartite structure composed of an ...	64-115	1.32e-21

Blast search parameters

Data Source: Live blast search **RID =** KWPWKJ2X013

User Options: **Database:** CDSEARCH/cdd **Low complexity filter:** no **Composition Based Adjustment:** yes **E-value threshold:** 0.01 **Maximum number of hits:** 500

References:

- Wang J et al. (2023), *"The conserved domain database in 2023"*, **Nucleic Acids Res.51**(D)384-8.
- Lu S et al. (2020), *"The conserved domain database in 2020"*, **Nucleic Acids Res.48**(D)265-8.
- Marchler-Bauer A et al. (2017), *"CDD/SPARCLE: functional classification of proteins via subfamily domain architectures."*, **Nucleic Acids Res.45**(D)200-3.

Protein family databases

- Classification of full-length protein sequences into families
- Whole-genome comparisons and phylogenetic classification to identify true *orthologs* in fully sequenced genomes (orthologs: homologs with the same function, *i.e.* the opposite of *paralogs*)
- This approach does not depend on the presence of particular sequence signatures or conserved domains
- Example: COG (Cluster of Orthologous Groups), https://www.ncbi.nlm.nih.gov/research/cog

COG (Cluster of Orthologous Groups)

- Constructed by comparing protein sequences encoded in 66 completely sequenced (mainly) prokaryotic genomes
- Orthologous proteins shared by three or more lineages are identified and clustered together as orthologous groups
- If the function of one of the members is known, functionality of other members can be assigned
- Currently there are 4872 clusters in the COG database derived from unicellular organisms
- KOG: more recent eukaryotic version of COG

Expectation maximization and Gibbs motif sampling

Methods starting from random alignments and PSSMs, gradually improved by iterative optimisation:

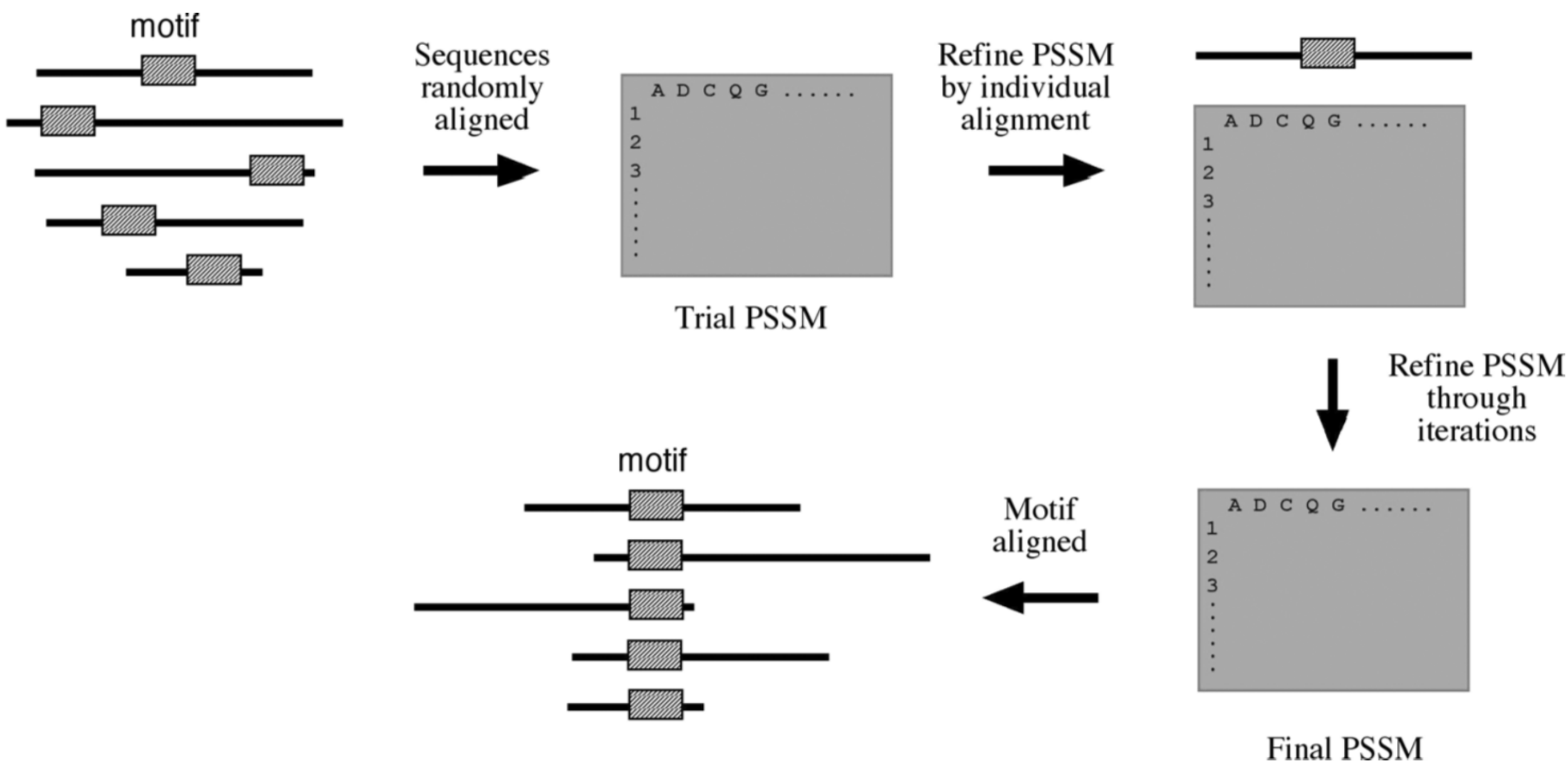


Figure 7.1: Schematic diagram of the EM algorithm.

Graphic representation: sequence logos

- Graphic representation of a multiple sequence alignment of a domain or motif, often encountered in literature
- Each position consists of stacked letters representing residues appearing in a particular column of a multiple alignment
- Over-all height of a logo position reflects how conserved the position is
- Height of each letter in a position reflects the relative frequency of the residue in the alignment
- Conserved positions have fewer residues and bigger symbols

Example of a sequence logo

```
GDLGAGKTT
GDLGAGKTT
GPLGAGKTS
GDLGAGKTS
GDLGAGKTT
GDLGAGKTT
GEVGSGKTT
GELGAGKTT
GDLGAGKTT
GNLGAGKTT
GELGAGKTT
GTLGAGKTT
GDLGAGKTT
GDLGAGKTT
GDLGAGKTT
GDLGAGKTT
GDLGAGKTT
```

Figure 7.2: Example of multiple alignment representation using a logo (produced using the WebLogo program).

http://weblogo.berkeley.edu

The (likely) future of domain identification

- The number of available experimental protein structures (in the PDB) continues to increase in near-exponential fashion (currently > 200 000)
- Structure prediction methods (*e.g.* AlphaFold) have matured to the point where there is nowadays a useful model for most protein sequences, even if there is no experimental structure
- Domain identification will <u>rely increasingly on 3D structural models</u>