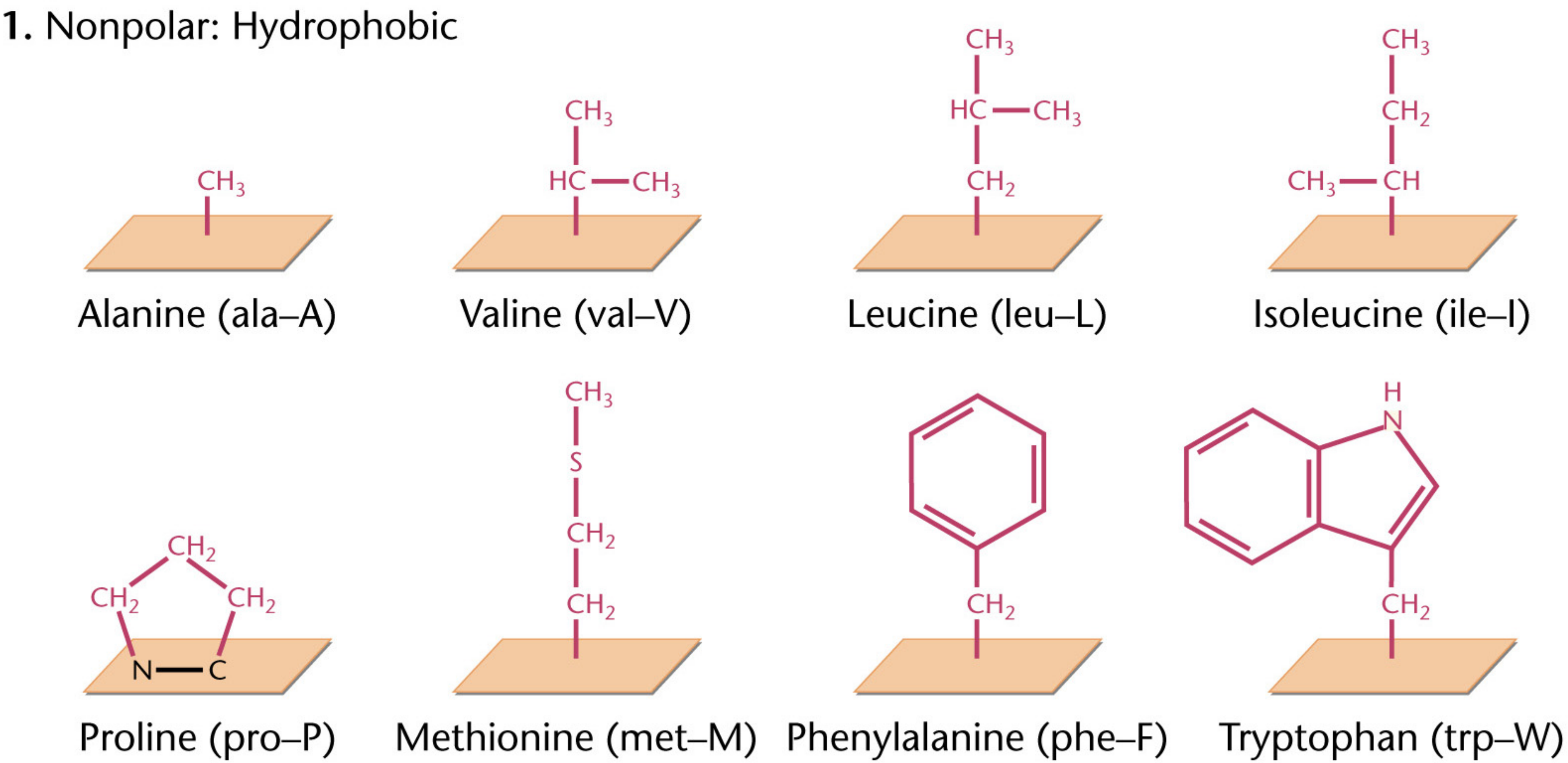
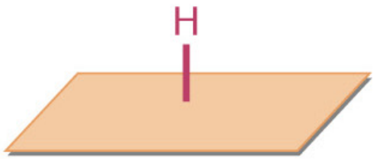
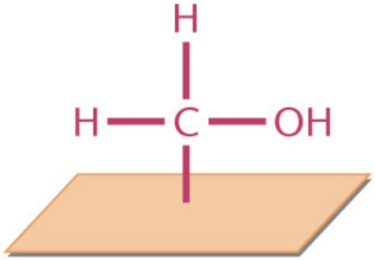


Amino Acid Structures from Klug & Cummings

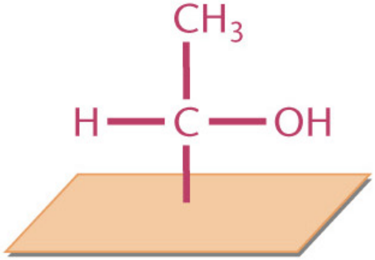
2. Polar: Hydrophilic



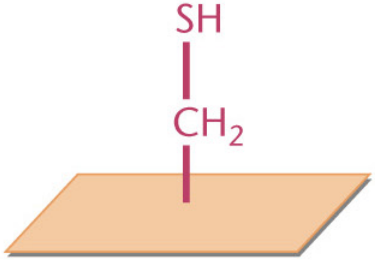
Glycine (gly–G)



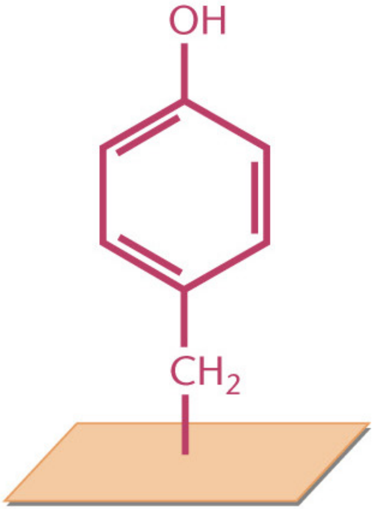
Serine (ser–S)



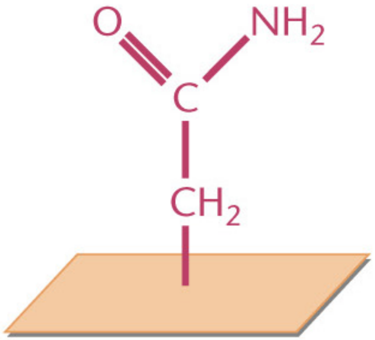
Threonine (thr–T)



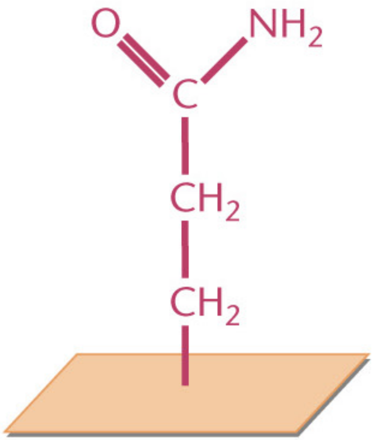
Cysteine (cys–C)



Tyrosine (tyr–Y)



Asparagine (asn–N)

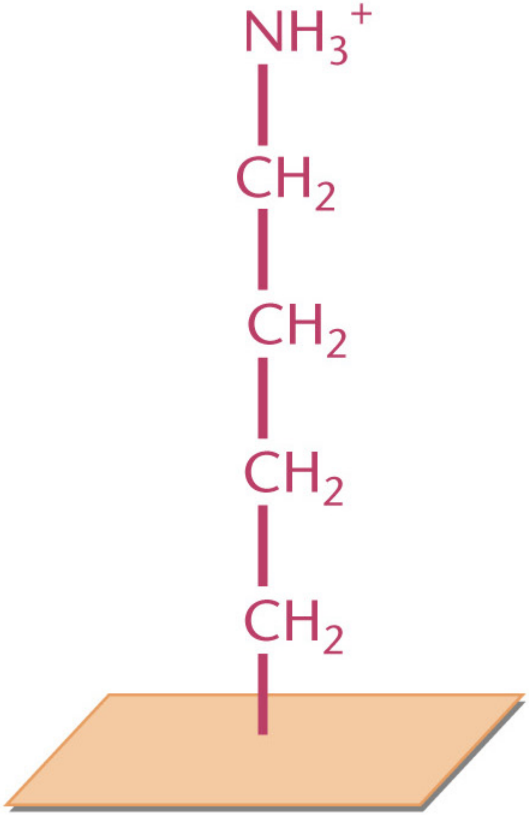


Glutamine (gln–Q)

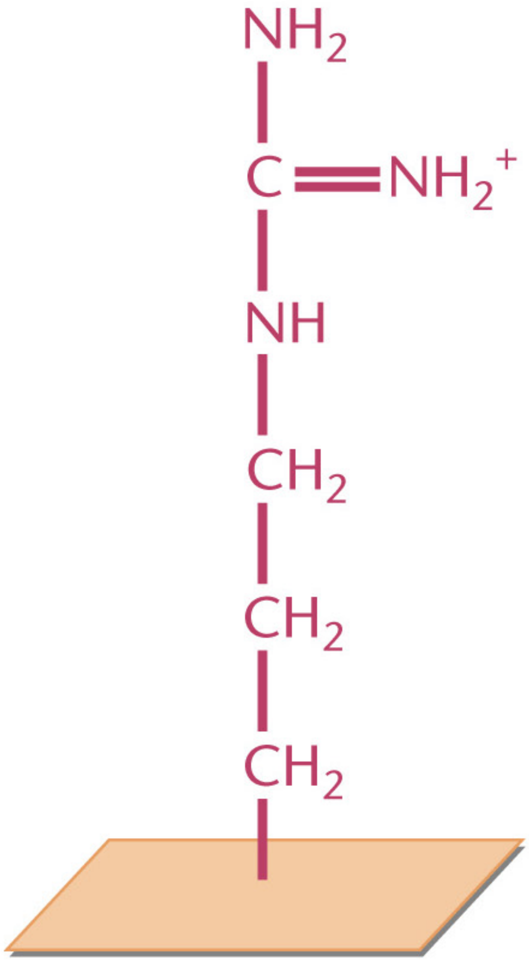
Amino Acid Structures from Klug & Cummings

3. Polar: positively charged (basic)

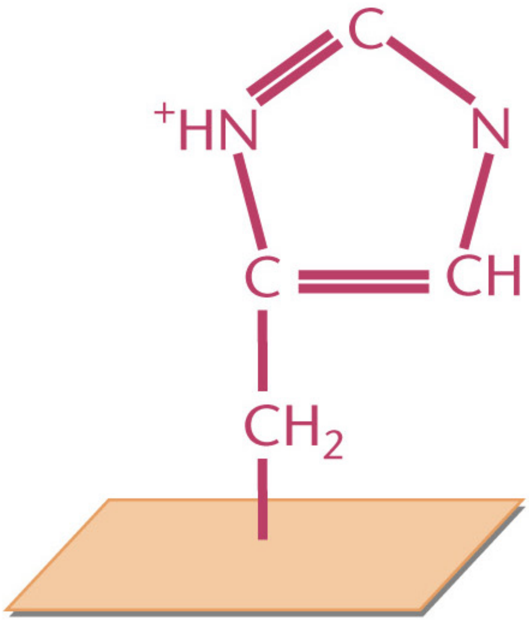
Amino Acid Structures from Klug & Cummings



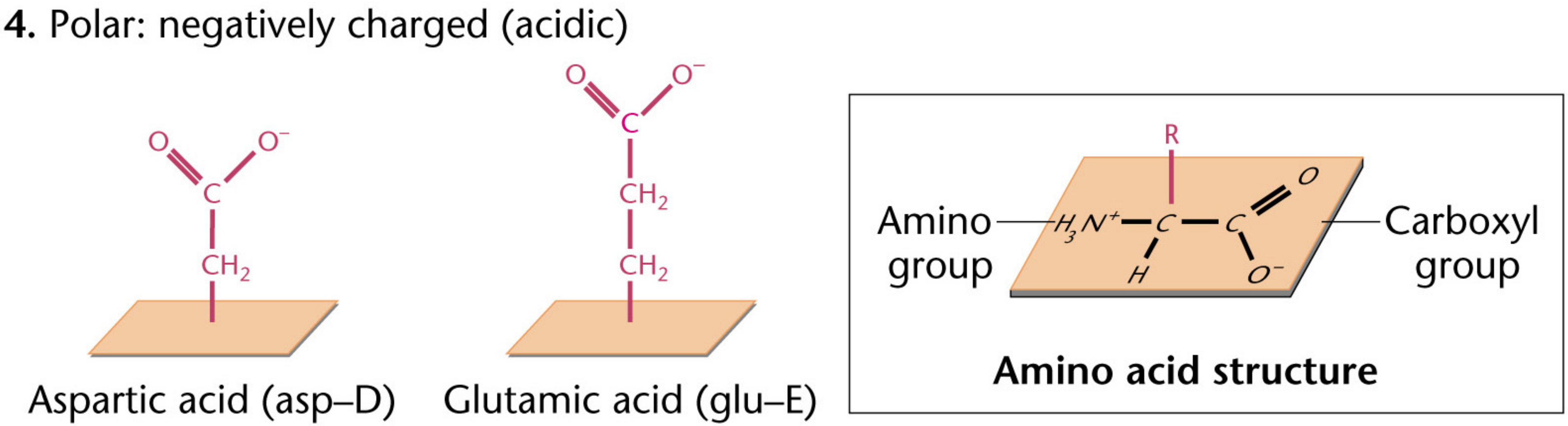
Lysine (lys–K)



Arginine (arg–R)



Histidine (his–H)



Amino Acid Structures from Klug & Cummings

Secondary Structure Prediction Software

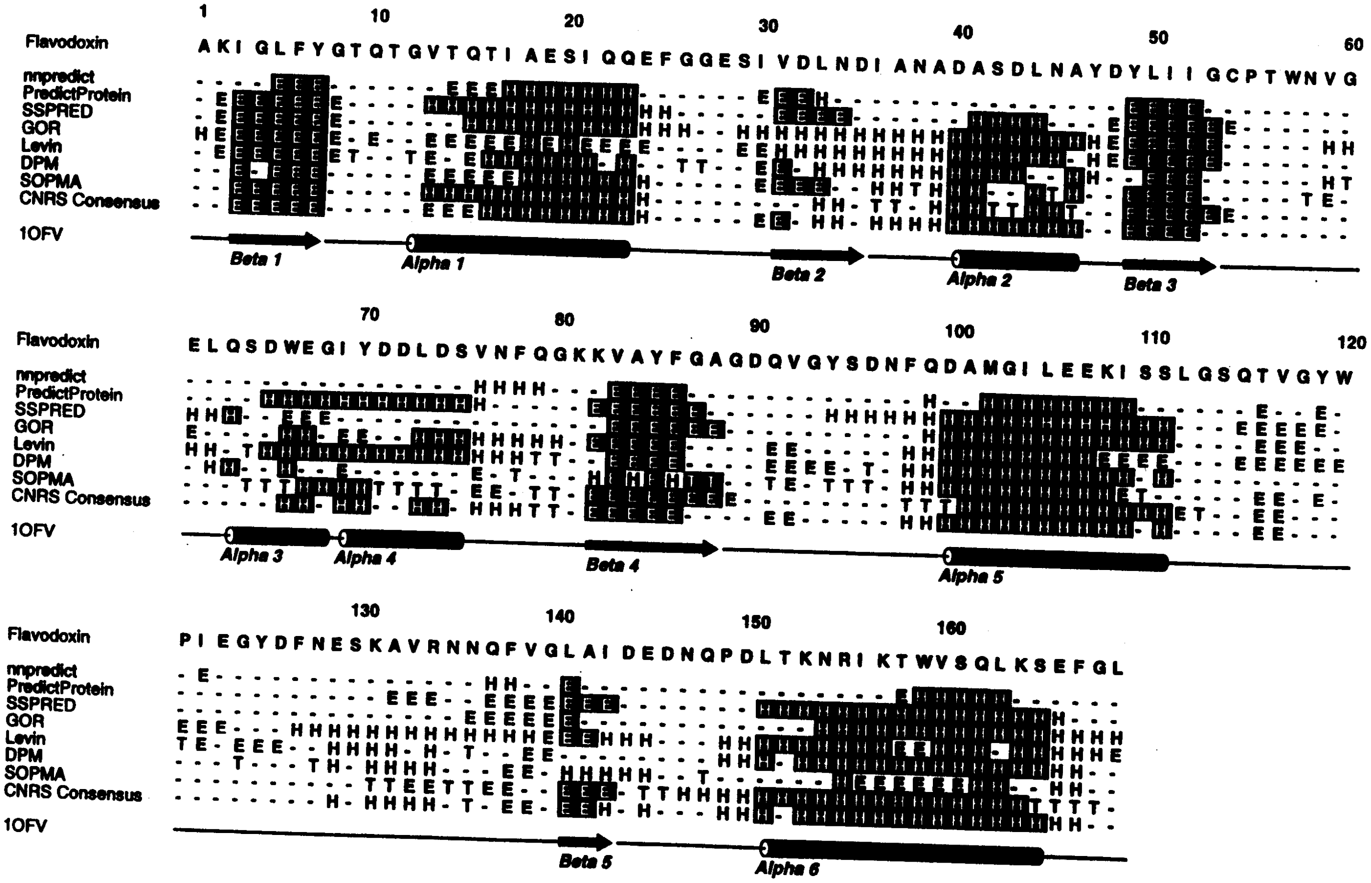

1
10
20
30
40
50
60
Flavodoxin
A K I G L F Y G T Q T G V T Q T I A E S I Q Q E F G G E S I V D L N D I A N A D A S D L N A Y D Y L I I G C P T W N V G
nnpredict
PredictProtein
SSPRED
GOR
Levin
DPM
SOPMA
CNRS Consensus
1OFV
Beta 1
Alpha 1
Beta 2
Alpha 2
Beta 3
70
80
90
100
110
120
Flavodoxin
E L Q S D W E G I Y D D L D S V N F Q G K K V A Y F G A G D Q V G Y S D N F Q D A M G I L E E K I S S L G S Q T V G Y W
nnpredict
PredictProtein
SSPRED
GOR
Levin
DPM
SOPMA
CNRS Consensus
1OFV
Alpha 3
Alpha 4
Beta 4
Alpha 5
130
140
150
160
Flavodoxin
P I E G Y D F N E S K A V R N N Q F V G L A I D E D N Q P D L T K N R I K T W V S Q L K S E F G L
nnpredict
PredictProtein
SSPRED
GOR
Levin
DPM
SOPMA
CNRS Consensus
1OFV
Beta 5
Alpha 6

Figure 11.3 Comparison of secondary structure predictions by various methods. The sequence of flavodoxin, an α/β protein, was used as the query and is shown on the first line of the alignment. For each prediction, H denotes an α helix, E a β strand, T a β turn; all other positions are assumed to be random coil. Correctly assigned residues are shown in inverse type. The methods used are listed along the left side of the alignment and are described in the text. At the bottom of the figure is the secondary structure assignment given in the PDB file for flavodoxin (1OFV, Smith et al., 1983).

Secondary Structure Prediction

- [NN based] PSI-pred, nnPredict (2-layer, feed-forward NN), Pred2ary
- [Consensus Approach] JPRED, SOPMA
- [K-nearest neighbor] NNSSP, PREDATOR
- [HMM] PSA
- ZPRED
- SSP
- PHD (See Sample)

Motif Detection Tools

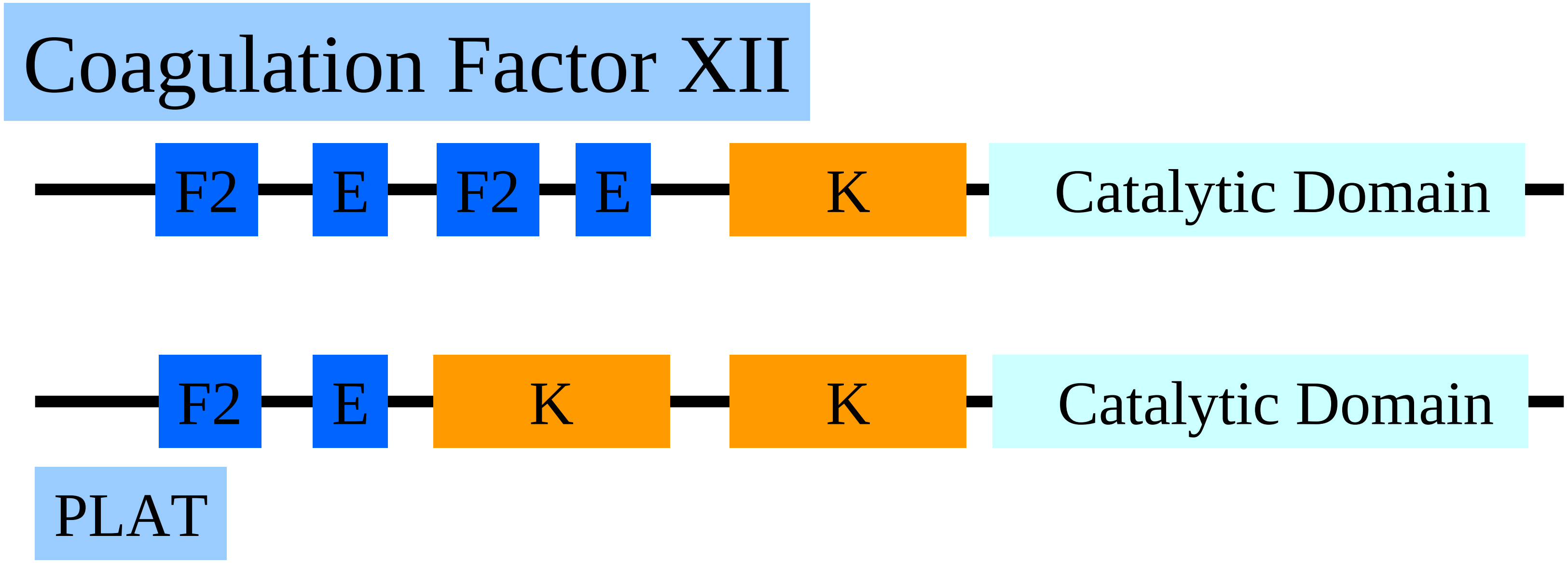
- PROSITE (Database of protein families & domains)
  - Try PDOC00040. Also Try PS00041
- PRINTS Sample Output
- BLOCKS (multiply aligned ungapped segments for highly conserved regions of proteins; automatically created) Sample Output
- Pfam (Protein families database of alignments & HMMs)
  - Multiple Alignment, domain architectures, species distribution, links: Try
- MoST
- PROBE
- ProDom
- DIP

Protein Information Sites

- SwissPROT & GenBank
- InterPRO is a database of protein families, domains and functional sites in which identifiable features found in known proteins can be applied to unknown protein sequences. See sample.
- PIR Sample Protein page

Modular Nature of Proteins

- Proteins are collections of “modular” domains. For example,



Domain Architecture Tools

- CDART
  - – Protein **AAH24495**; Domain Architecture;
  - – It's domain relatives;
  - – Multiple alignment for 2nd domain
- SMART

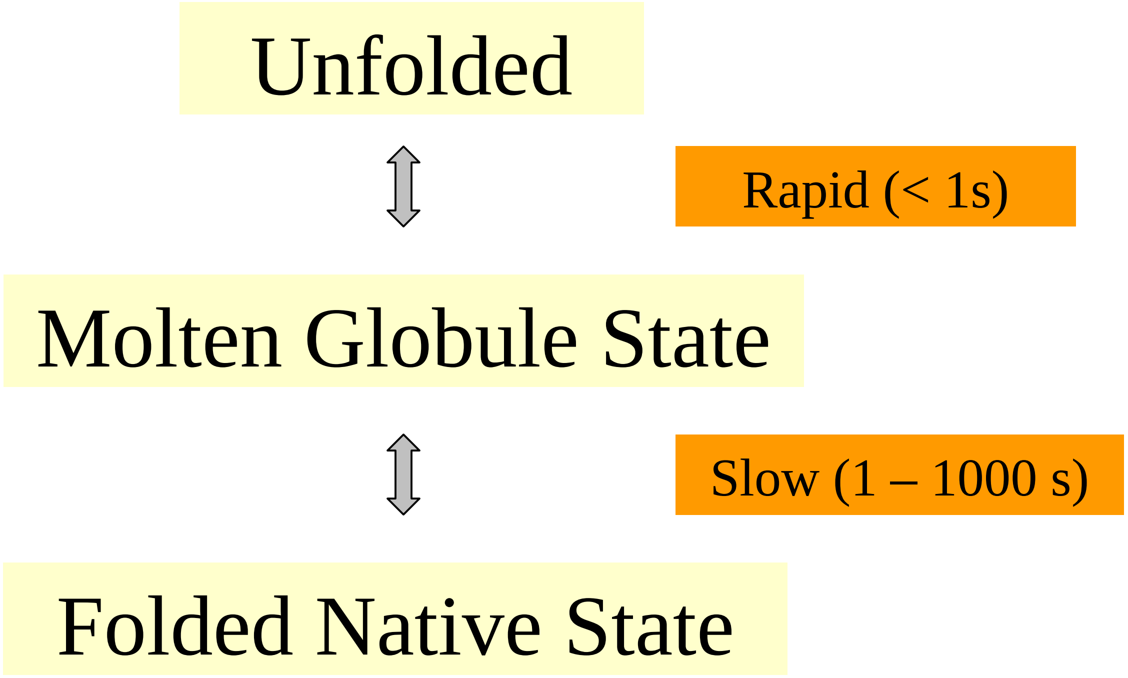
Predicting Specialized Structures

- COILS – Predicts coiled coil motifs
- TMPred – predicts transmembrane regions
- SignalP – predicts signal peptides
- SEG – predicts nonglobular regions

Tertiary & Quaternary Protein Structures

- Experimental methods
  - X-ray crystallography [More accurate!]
  - Nuclear Magnetic Resonance Spectroscopy (NMR)
- If protein “unfolded” (denatured) and “released”, then it goes back to its native 3-d structure.
- The tertiary structure is a structure of minimum energy.
- Angles ϕ and ψ are constrained.
- Proteins structures often have hydrophobic core.

Protein Folding



Unfolded

Rapid (< 1s)

Molten Globule State

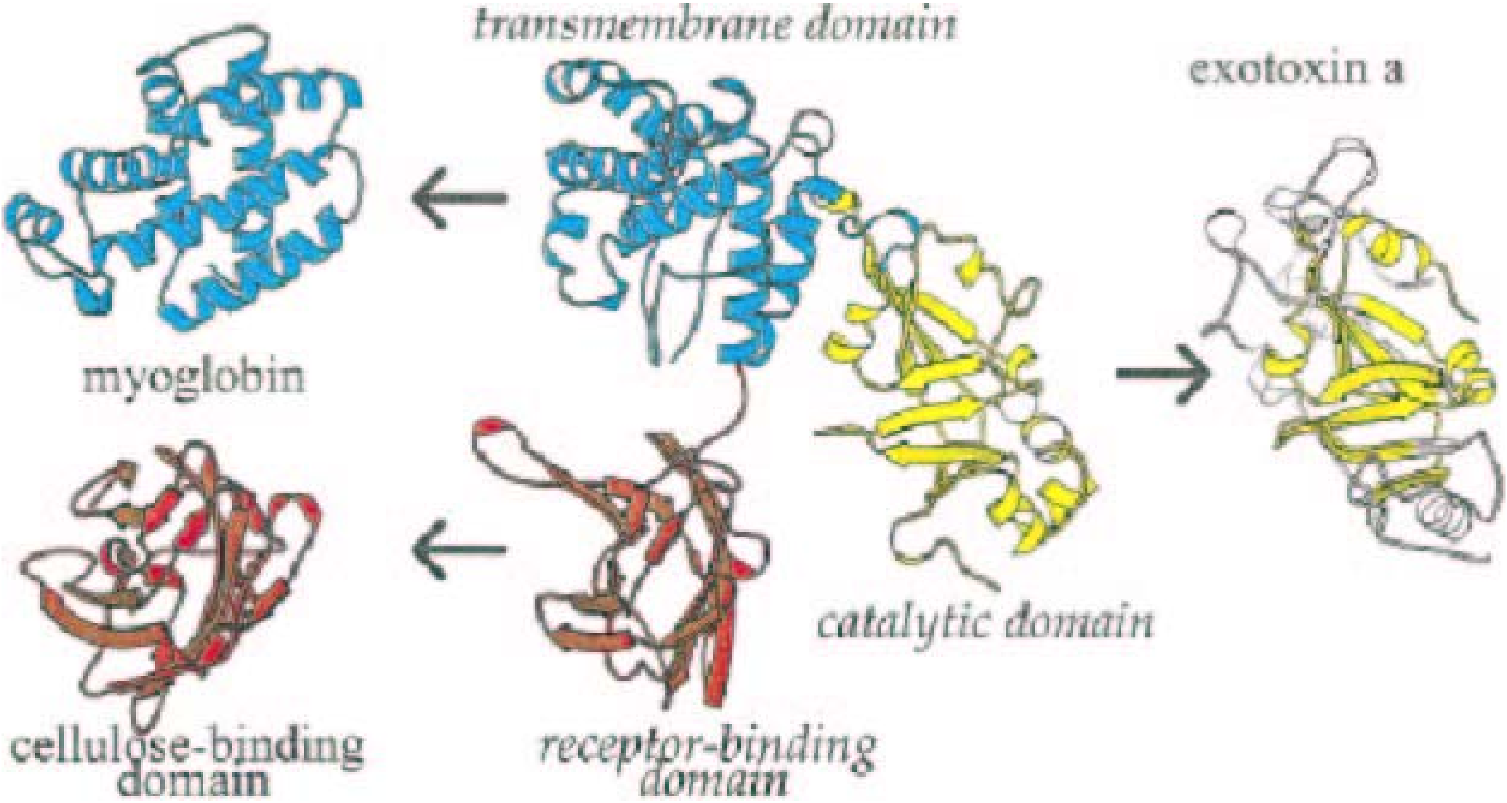
Slow (1 – 1000 s)

Folded Native State

- How to find minimum energy configuration?

Modular Nature of Protein Structures

Example: Diphtheria Toxin



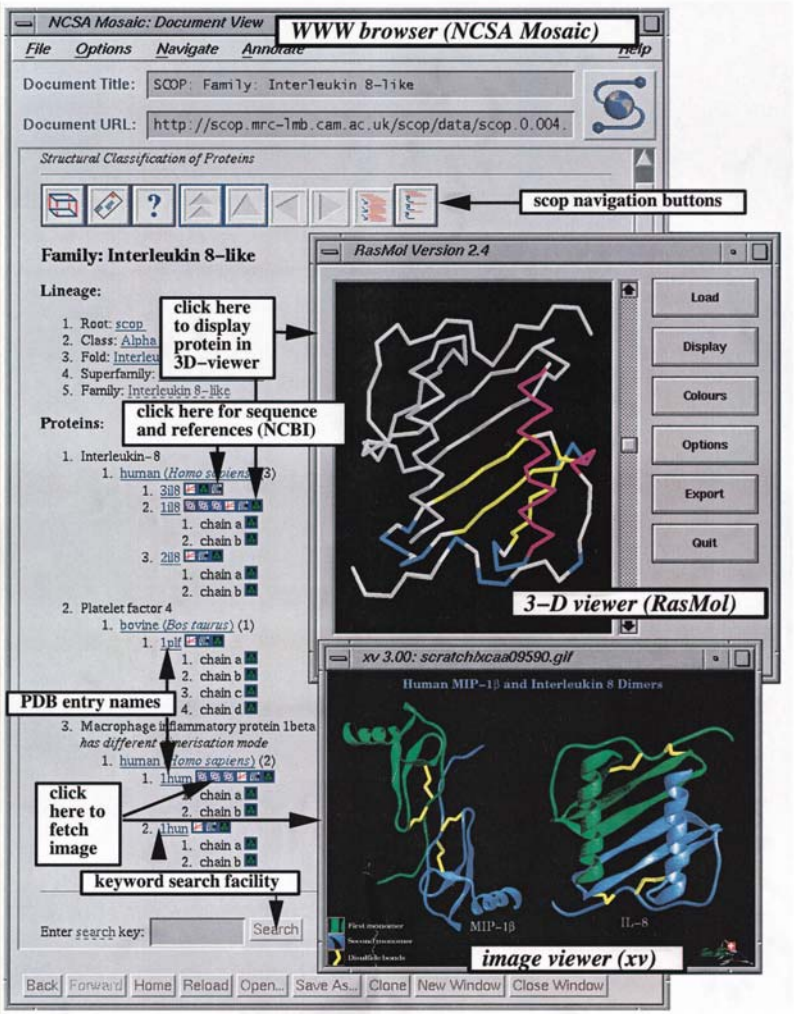
Structural Classification of Proteins

- SCOP (Structural Classification of Proteins)
  - Based on structurla & evolutionary relationships.
  - Contains ~ 40,000 domains
  - Classes (groups of folds), Folds (proteins sharing folds), Families (proteins related by function/evolution), Superfamilies (distantly related proteins)

SCOP Family View

JMB—MS 422

538 Communication


NCSA Mosaic: Document View
WWW browser (NCSA Mosaic)
File Options Navigate Annotate Help
scop navigation buttons
click here to display protein in 3D–viewer
click here for sequence and references (NCBI)
PDB entry names
click here to fetch image
keyword search facility
RasMol Version 2.4
Load
Display
Colours
Options
Export
Quit
3–D viewer (RasMol)
xv 3.00: scratch/xcaa09590.gif
Human MIP-1β and Interleukin 8 Dimers
First monomer
Second monomer
Disulfide bonds
MIP-1β
IL-8
image viewer (xv)
Back Forward Home Reload Open... Save As... Clone New Window Close Window

Document Title: SCOP: Family: Interleukin 8-like

Document URL: http://scop.mrc-lmb.cam.ac.uk/scop/data/scop.0.004.

Structural Classification of Proteins

Family: Interleukin 8–like

Lineage:

1. Root: scop
2. Class: Alpha
3. Fold: Interleu
4. Superfamily:
5. Family: Interleukin 8–like

Proteins:

1. Interleukin–8
   1. human (*Homo sapiens*) (3)
      1. 3il8
      2. 1il8
         1. chain a
         2. chain b
      3. 2il8
         1. chain a
         2. chain b
2. Platelet factor 4
   1. bovine (*Bos taurus*) (1)
      1. 1plf
         1. chain a
         2. chain b
         3. chain c
         4. chain d
3. Macrophage inflammatory protein 1beta
   has different dimerisation mode
   1. human (*Homo sapiens*) (2)
      1. 1hum
         1. chain a
         2. chain b
      2. 1hun
         1. chain a
         2. chain b

Enter search key: Search

Figure 2. A typical scop session is shown on a unix workstation. A scop page, of the Interleukin 8-like family, is displayed by the *WWW browser program (NCSA Mosaic)* (Schatz & Hardin, 1994). Navigating through the tree structure is accomplished by selecting any underlined entry, by clicking on buttons (at the top of each page) and by keyword searching (at the bottom of each page). The static image comparing two proteins in this family was downloaded by clicking on the icon indicated and is displayed by image-viewer program *xv*. By clicking on one of the green icons, commands were sent to a molecular viewer program (*RasMol*) written by Roger Sayle (Sayle, 1994), instructing it to automatically display the relevant PDB file and colour the domain in question by secondary structure. Since sending large PDB files over the network can be slow, this feature of scop can be configured to use local copies of PDB files if they are available. Equivalent WWW browsers, image-display programs and molecular viewers are also available free for Windows-PC and Macintosh platforms.

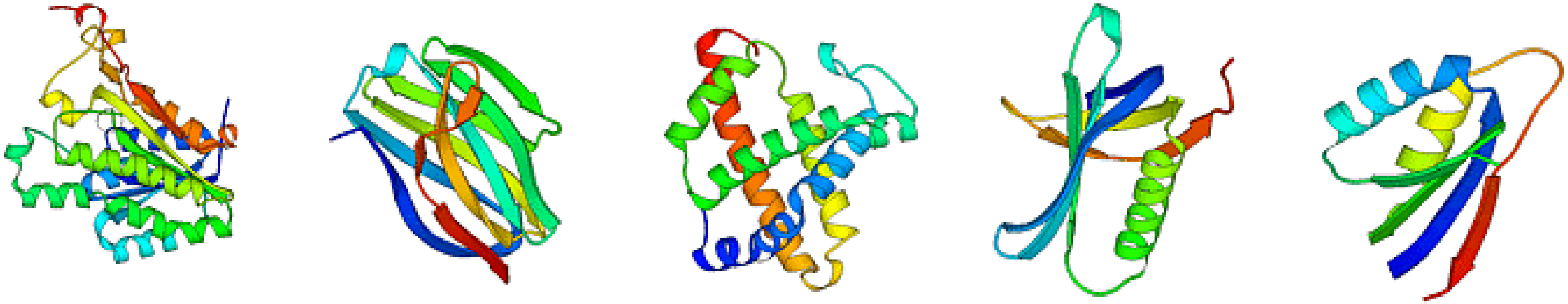
CATH: Protein Structure Classification

- Semi-automatic classification; ~36K domains
- 4 levels of classification:
  - Class (C), depends on sec. Str. Content
  - Architecture (A), orientation of sec. Str.
  - Topolgy (T), topological connections &
  - Homologous Superfamily (H), similar str and functions.

DALI Domain Dictionary

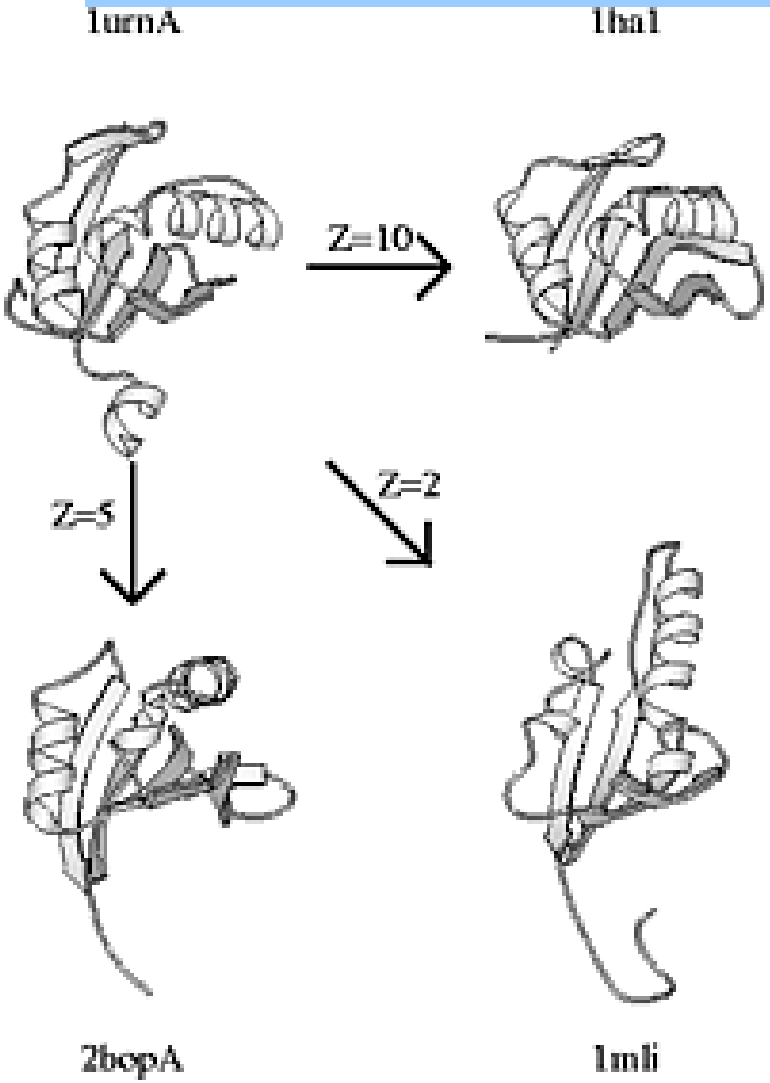
- Completely automated; 3724 domains
- Criteria of compactness & recurrence
- Each domain is assigned a Domain Classification number DC_l_m_n_p representing fold space attractor region (l), globular folding topology (m), functional family (n) and sequence family (p).

5 Fold Space classes



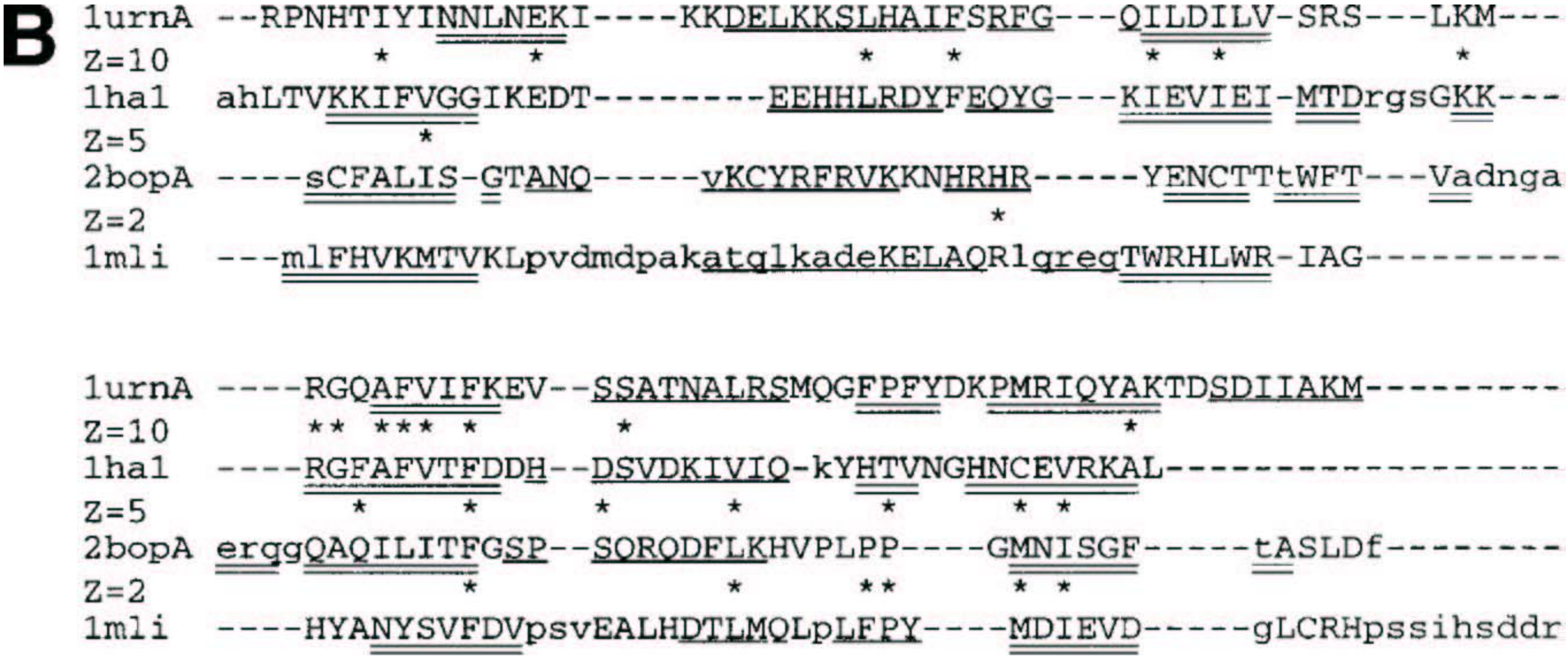
Attractor 1 can be characterized as alpha/beta,
attractor 2 as all-beta, attractor 3 as all-alpha,
attractor 5 as alpha-beta meander (1mli), and
attractor 4 contains antiparallel beta-barrels e.g. OB-fold (1prtF).

Fold Types & Neighbors



Structural neighbours of 1urnA (top left). 1mli (bottom right) has the same topology even though there are shifts in the relative orientation of secondary structure elements.

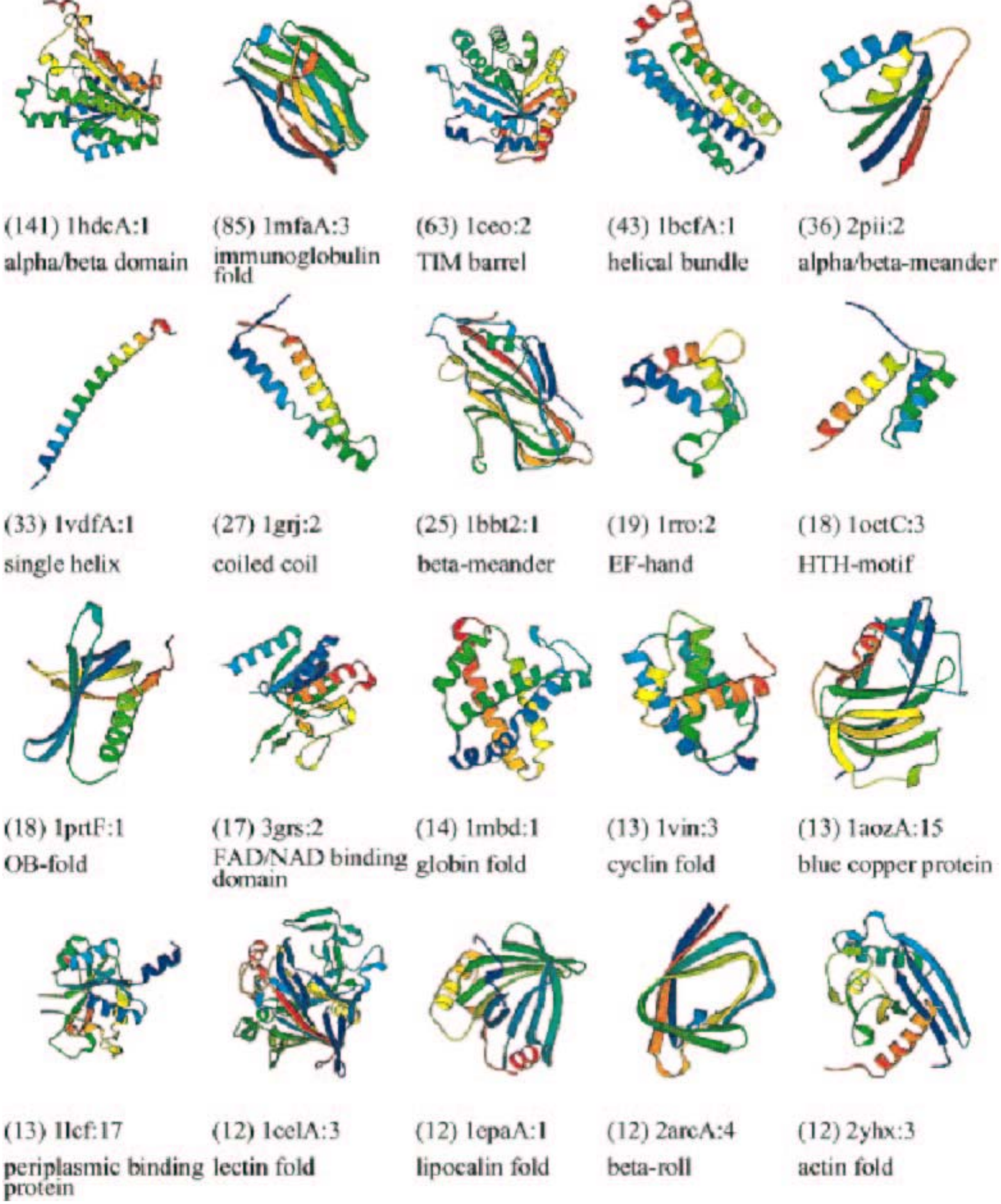
Sequence Alignment of Fold Neighbors



```
1urnA --RPNHTIYINNLNEKI----KKDELKKSLHAIFSRFG---QILDILV-SRS---LKM---
Z=10         *      *              *   *        *  *          *
1hal  ahLTVKKIFVGGIKEDT--------EEHHLRDYFEQYG---KIEVIEI-MTDrgsGKK---
Z=5            *
2bopA ----sCFALIS-GTANQ-----vKCYRFRVKKNHRHR-----YENCTTtWFT---Vadnga
Z=2                                      *
1mli  ---mlFHVKMTVKLpvdmdpakatqlkadeKELAQRlqreqTWRHLWR-IAG---------

1urnA ----RGQAFVIFKEV--SSATNALRSMQGFPFYDKPMRIQYAKTDSDIIAKM---------
Z=10      ** *** *      *                      *
1hal  ----RGFAFVTFDDH--DSVDKIVIQ-kYHTVNGHNCEVRKAL------------------
Z=5         *    *     *     *      *     * *
2bopA erggQAQILITFGSP--SQRQDFLKHVPLPP----GMNISGF-----tASLDf--------
Z=2              *           *     **      * *
1mli  ----HYANYSVFDVpsvEALHDTLMQLpLFPY----MDIEVD-----gLCRHpssihsddr
```

Frequent Fold Types



Motifs in Protein Sequences

Motifs are combinations of secondary structures in proteins with a specific **structure** and a specific **function**. They are also called **super-secondary structures**.

Examples: Helix-Turn-Helix, Zinc-finger,
Homeobox domain, Hairpin-beta motif,
Calcium-binding motif, Beta-alpha-beta motif,
Coiled-coil motifs.

Several motifs may combine to form **domains**.

- Serine proteinase domain, Kringle domain, calcium-binding domain, homeobox domain.

Motif Detection Problem

Input: Set, S, of known (aligned) examples of a motif M,
A new protein sequence, P.

Output: Does P have a copy of the motif M?

Example: Zinc Finger Motif

…**Y**K**C**GL**C**ERS**F**VEKSA**L**SR**H**0RV**H**KN…

3 6 19 23

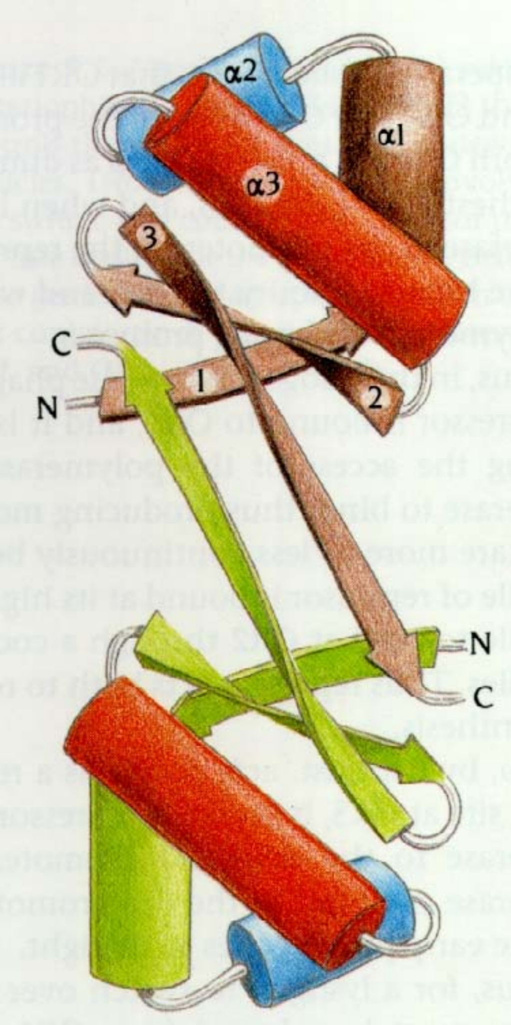
Input: Database, D, of known protein sequences,
A new protein sequence, P.

Output: What interesting patterns from D are present in P?

Helix-Turn-Helix Motifs

- Structure
  - 3-helix complex
  - Length: 22 amino acids
  - Turn angle

- Function
  - Gene regulation by binding to DNA



DNA Binding at HTH Motif

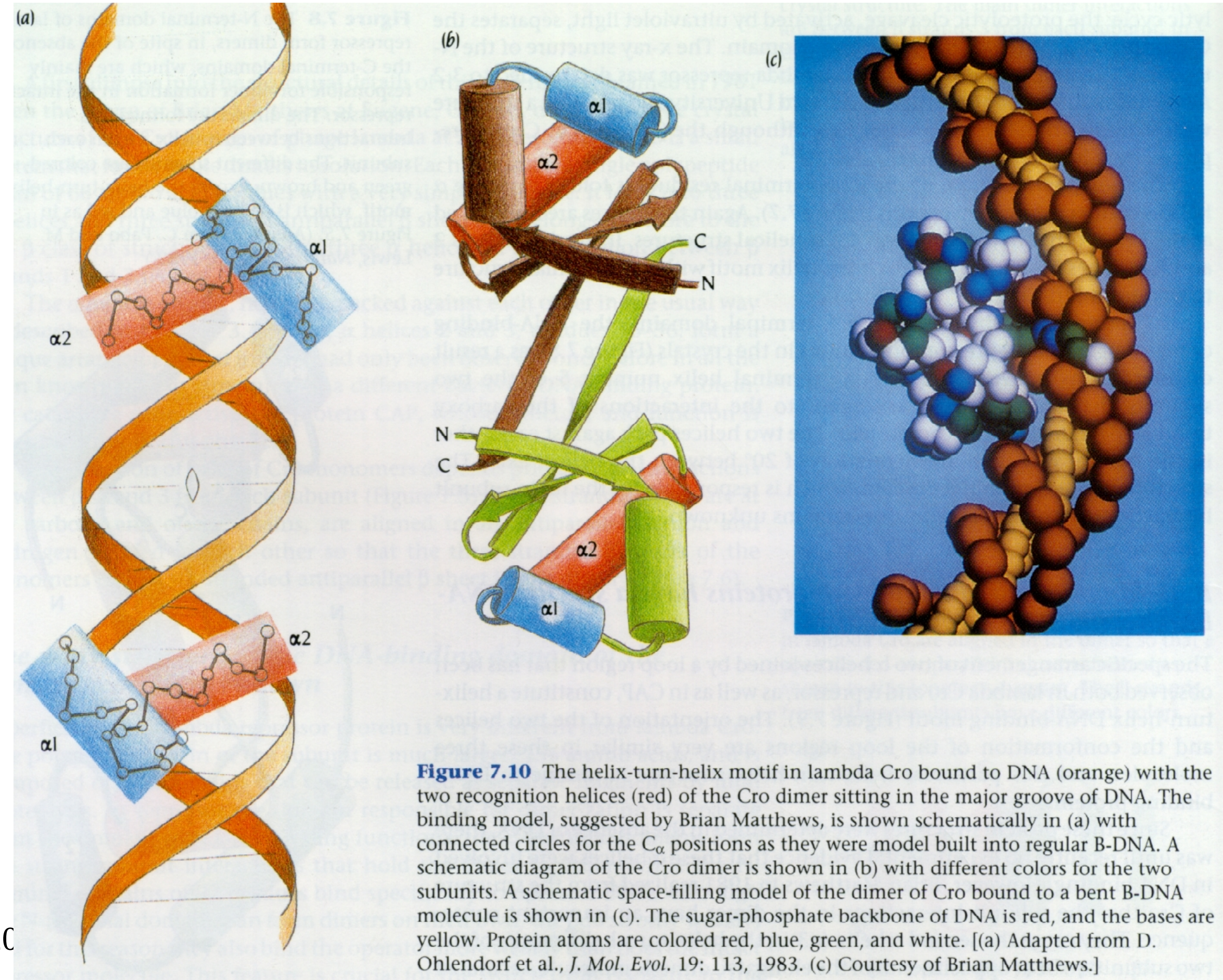


Figure 7.10 The helix-turn-helix motif in lambda Cro bound to DNA (orange) with the two recognition helices (red) of the Cro dimer sitting in the major groove of DNA. The binding model, suggested by Brian Matthews, is shown schematically in (a) with connected circles for the C_α positions as they were model built into regular B-DNA. A schematic diagram of the Cro dimer is shown in (b) with different colors for the two subunits. A schematic space-filling model of the dimer of Cro bound to a bent B-DNA molecule is shown in (c). The sugar-phosphate backbone of DNA is red, and the bases are yellow. Protein atoms are colored red, blue, green, and white. [(a) Adapted from D. Ohlendorf et al., *J. Mol. Evol.* 19: 113, 1983. (c) Courtesy of Brian Matthews.]

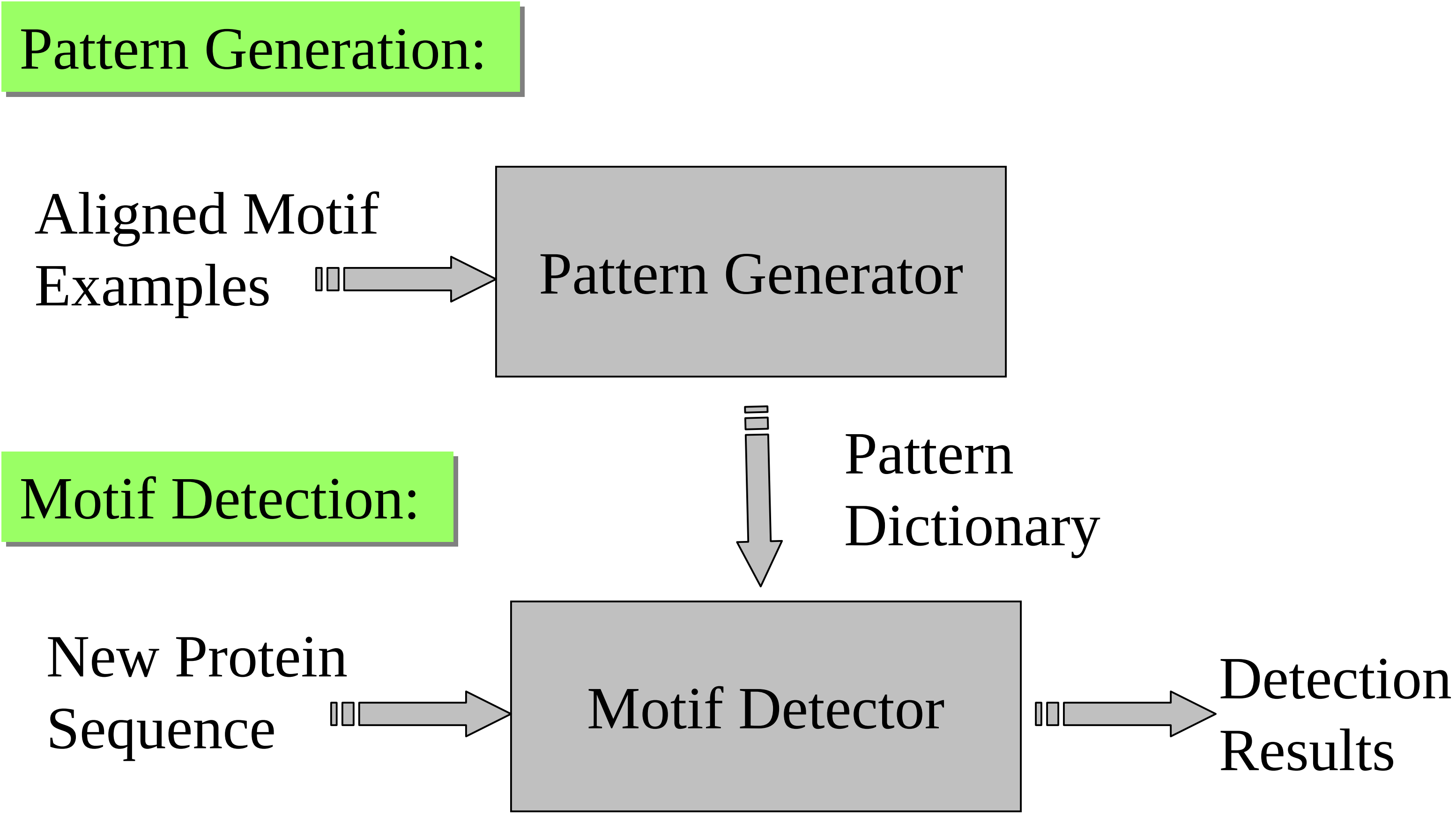
HTH Motifs: Examples

Loc		Helix 2	Turn

Basis for New Algorithm

- Combinations of residues in specific locations (may not be contiguous) contribute towards stabilizing a structure.
- Some reinforcing combinations are relatively rare.

New Motif Detection Algorithm



Pattern Generation:

Aligned Motif Examples

Pattern Generator

Pattern Dictionary

Motif Detection:

New Protein Sequence

Motif Detector

Detection Results

Patterns

Loc	Protein Name	Helix 2			Turn	Helix 3
		-1	0	1		

- Q1 G9 N20
- A5 G9 V10 I15

Pattern Mining Algorithm

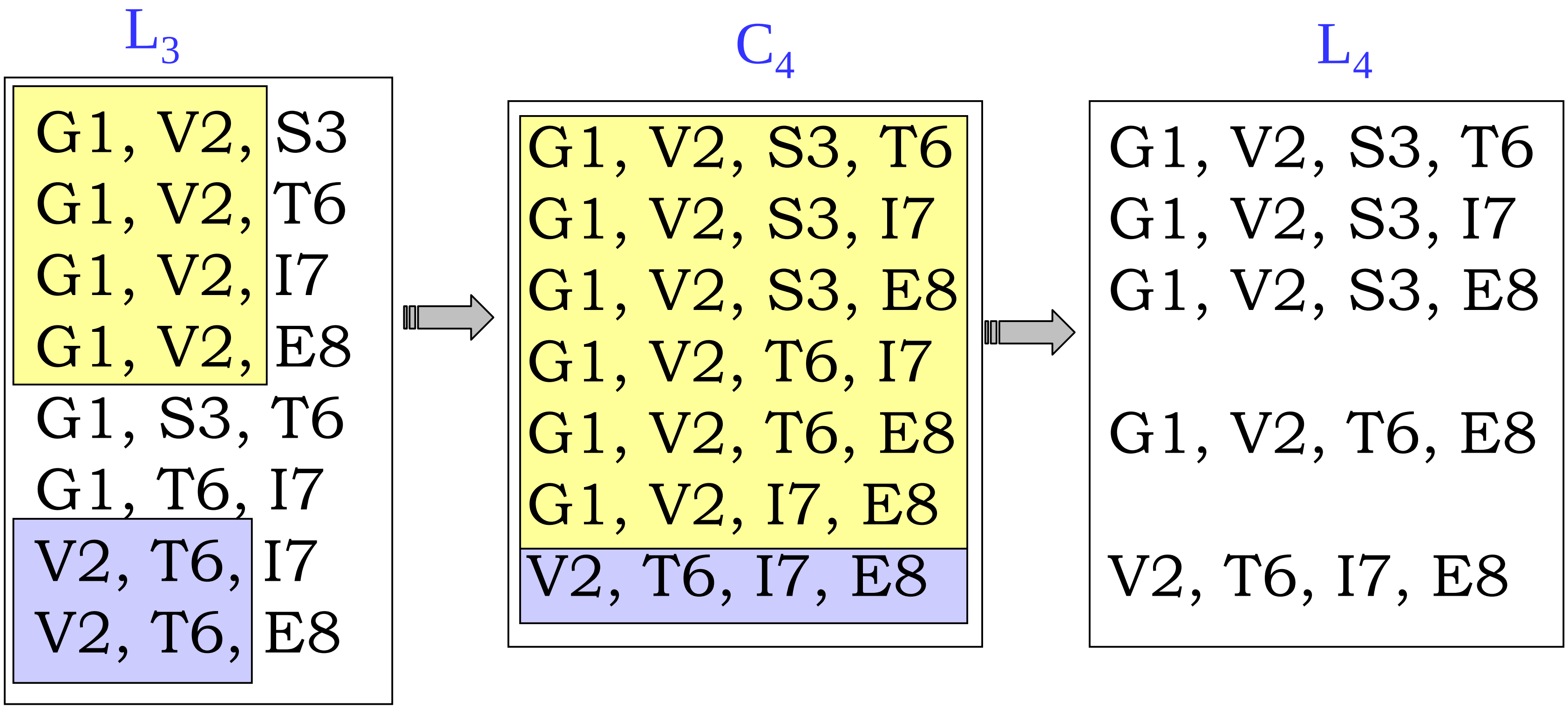
Algorithm **Pattern-Mining**

Input: Motif length m, support threshold T,
list of aligned motifs M.

Output: Dictionary L of frequent patterns.

```
$L_1$ := All frequent patterns of length 1
for i = 2 to m do
    $C_i$ := Candidates($L_{i-1}$)
    $L_i$ := Frequent candidates from $C_i$
    if ($|L_i| <= 1$) then
        return L as the union of all $L_j$ , j <= i.
```

Candidates Function



L_3

G1, V2, S3
G1, V2, T6
G1, V2, I7
G1, V2, E8
G1, S3, T6
G1, T6, I7
V2, T6, I7
V2, T6, E8

C_4

G1, V2, S3, T6
G1, V2, S3, I7
G1, V2, S3, E8
G1, V2, T6, I7
G1, V2, T6, E8
G1, V2, I7, E8
V2, T6, I7, E8

L_4

G1, V2, S3, T6
G1, V2, S3, I7
G1, V2, S3, E8

G1, V2, T6, E8

V2, T6, I7, E8

Motif Detection Algorithm

Algorithm Motif-Detection

Input : Motif length m, threshold score T, pattern dictionary L, and input protein sequence P[1..n].

Output : Information about motif(s) detected.

```
for each location i do
    S := MatchScore(P[i..i+m-1], L).
    if (S > T) then
        Report it as a possible motif
```

Experimental Results: **GYM 2.0**

Motif	*Protein Family*	*Number Tested*	*GYM = DE Agree*	*Number Annotated*	*GYM = Annot.*
HTH Motif (22)	Master	88	88 (100 %)	13	13
	Sigma	314	284 + 23 (98 %)	96	82
	Negates	93	86 (92 %)	0	0
	LysR	130	127 (98 %)	95	93
	AraC	68	57 (84 %)	41	34
	Rreg	116	99 (85 %)	57	46
	Total	675	653 + 23 (94 %)	289	255 (88 %)

Experiments

- Basic Implementation (Y. Gao)
- Improved implementation & comprehensive testing (K. Mathee, GN).
- Implementation for homeobox domain detection (X. Wang).
- Statistical methods to determine thresholds (C. Bu).
- Use of substitution matrix (C. Bu).
- Study of patterns causing errors (N. Xu).
- Negative training set (N. Xu).
- NN implementation & testing (J. Liu & X. He).
- HMM implementation & testing (J. Liu & X. He).