

CAP 5510: Introduction to Bioinformatics

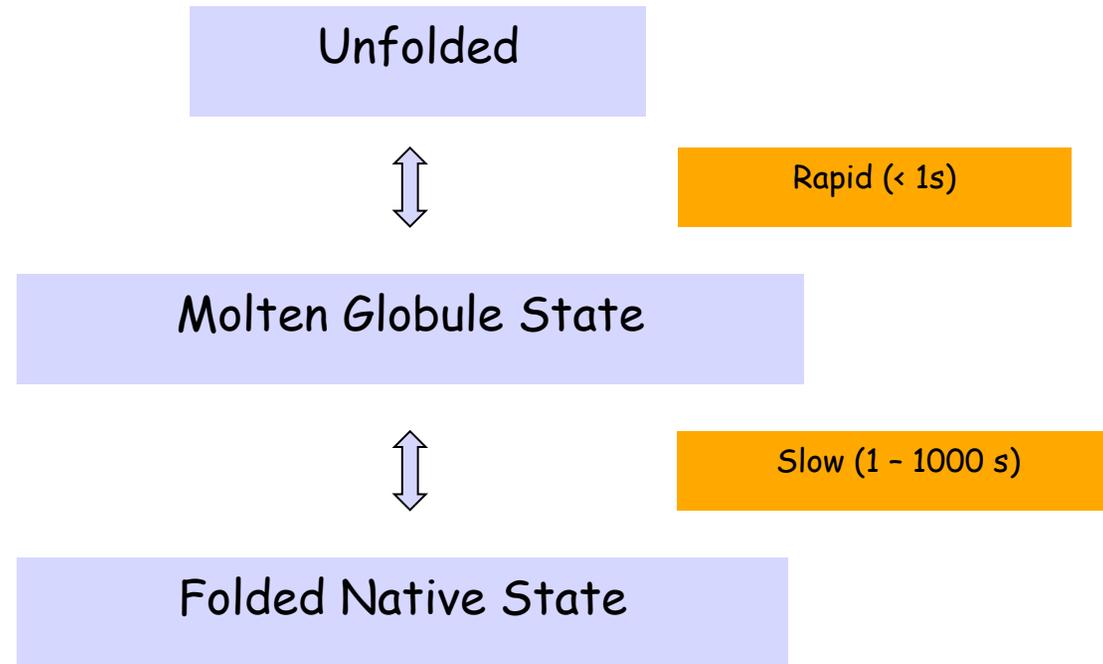
Giri Narasimhan

ECS 254; Phone: x3748

giri@cis.fiu.edu

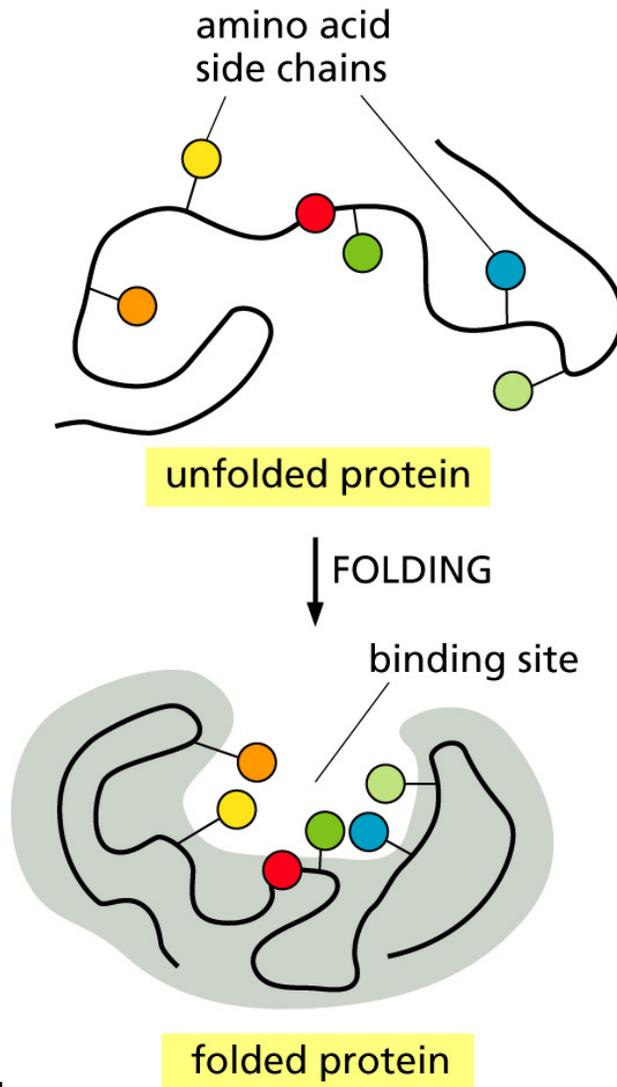
www.cis.fiu.edu/~giri/teach/BioinfS11.html

Protein Folding

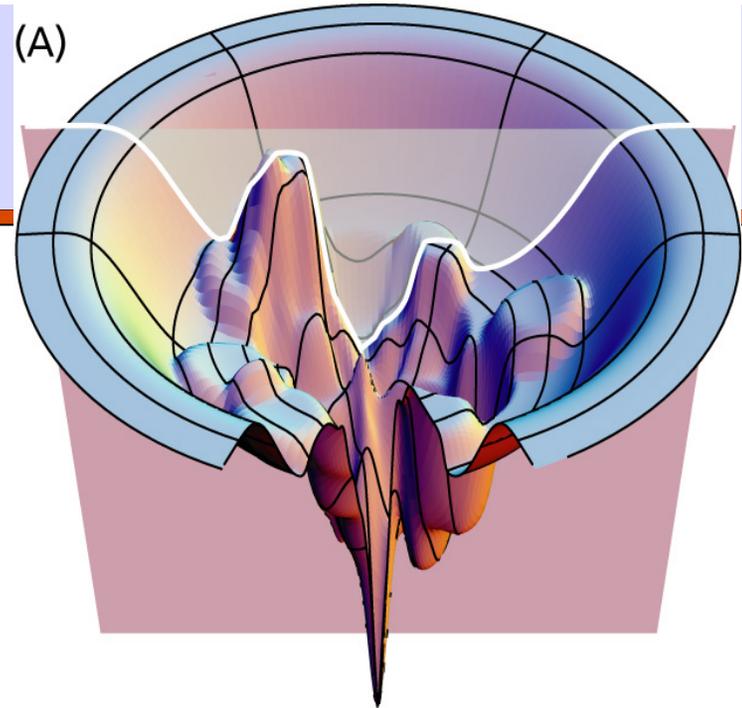


□ How to find minimum energy configuration?

Protein Folding



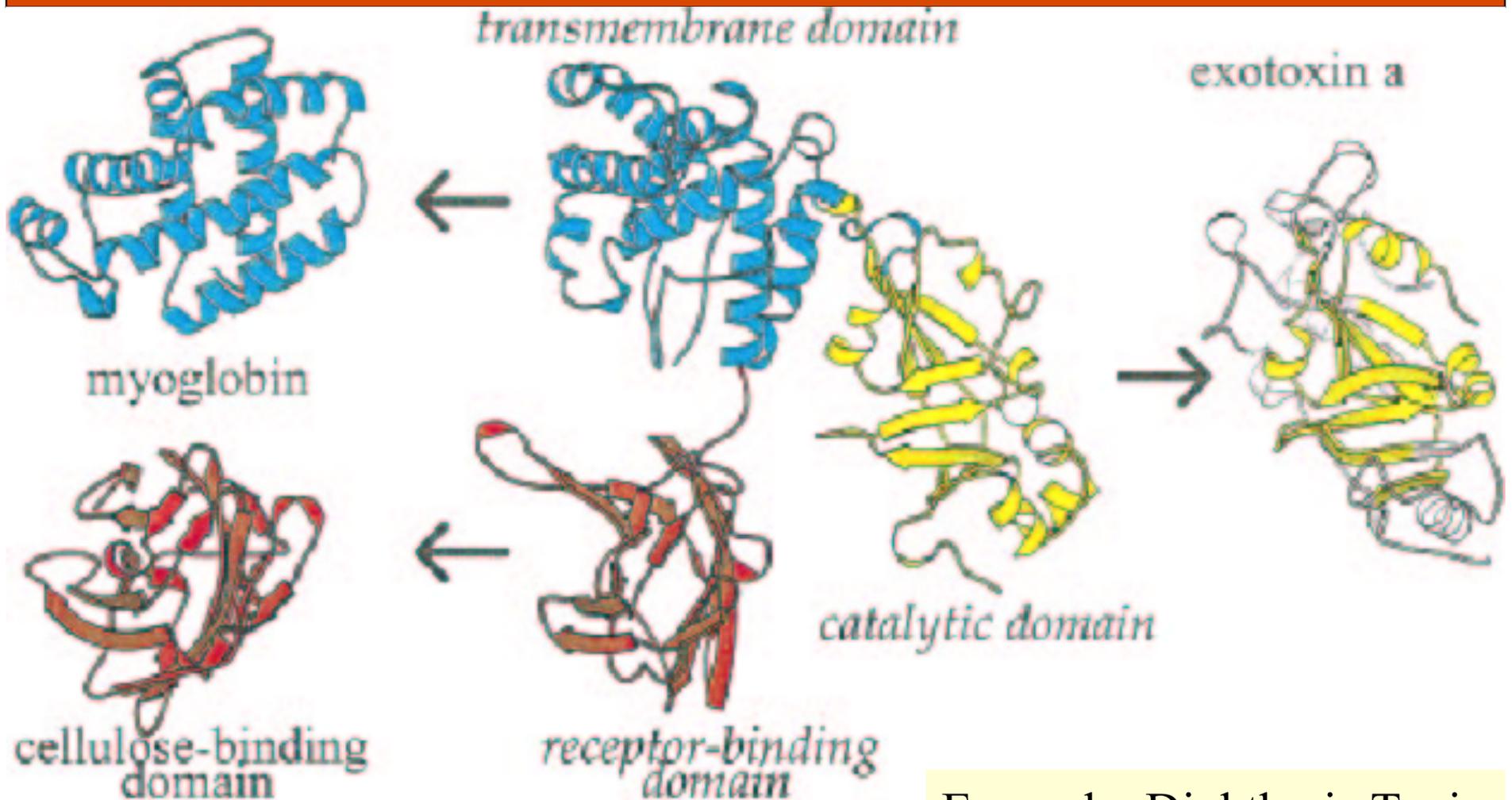
Energy Landscape



(B)



Modular Nature of Protein Structures



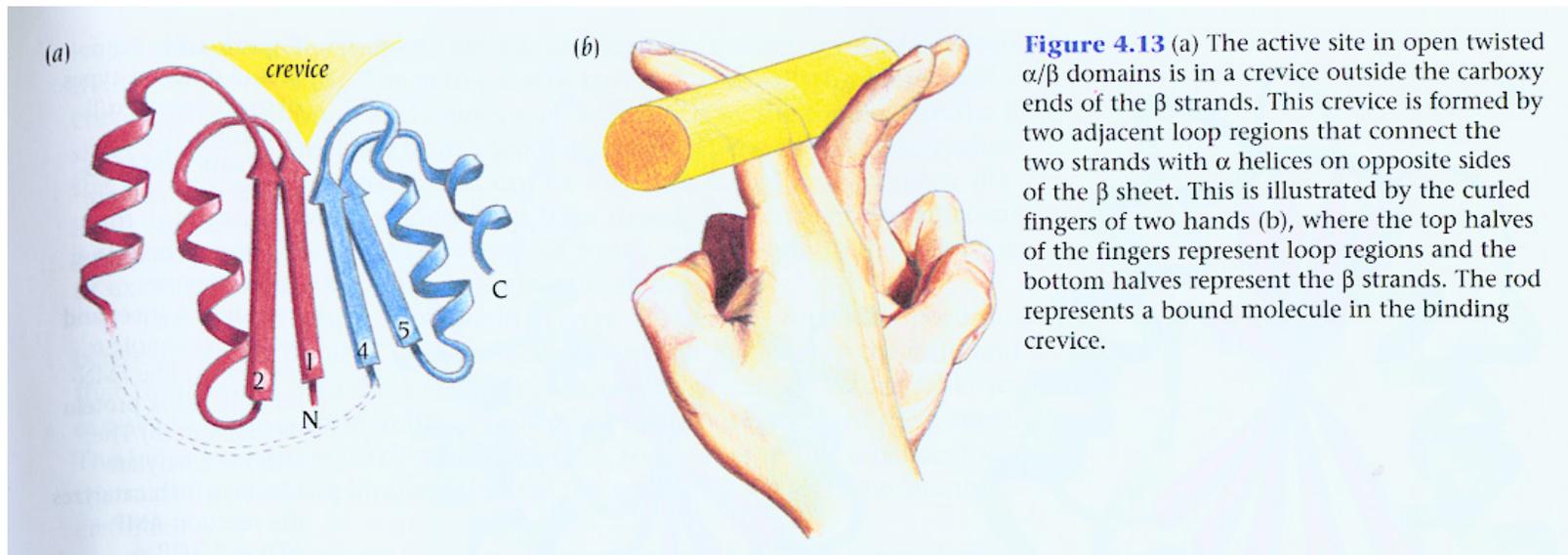
Example: Diphtheria Toxin

Protein Structures

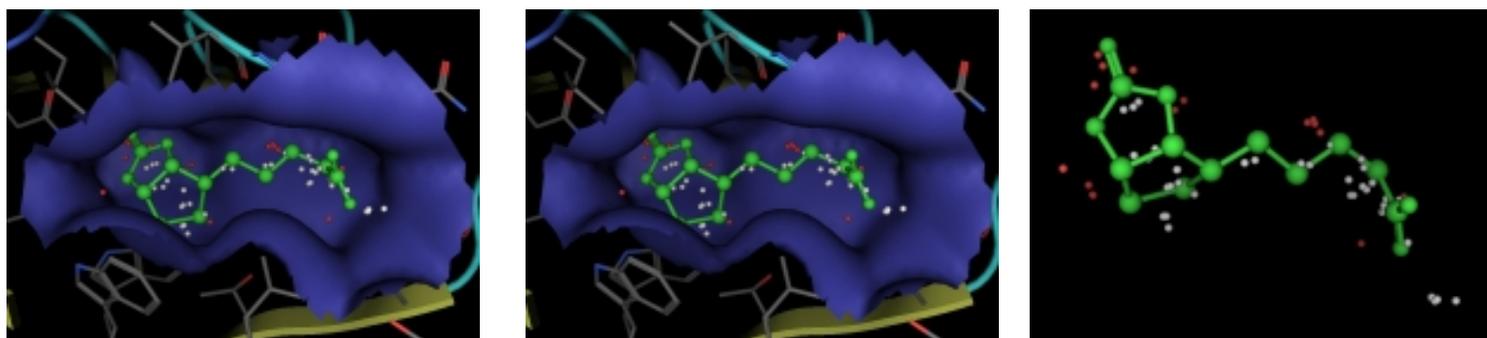
- ❑ Most proteins have a **hydrophobic core**.
- ❑ Within the core, specific **interactions** take place between amino acid side chains.
- ❑ Can an amino acid be replaced by some other amino acid?
 - Limited by space and available contacts with nearby amino acids
- ❑ Outside the core, proteins are composed of loops and structural elements in contact with water, solvent, other proteins and other structures.

Active Sites

Active sites in proteins are usually hydrophobic pockets/crevices/troughs that involve sidechain atoms.



Active Sites



Left PDB 3RTD (streptavidin) and the first site located by the MOE Site Finder. **Middle** 3RTD with complexed ligand (biotin). **Right** Biotin ligand overlaid with calculated alpha spheres of the first site.

Viewing Protein Structures

- SPDBV
- RASMOL
- CHIME

Structural Classification of Proteins

- Over 1000 protein families known
 - Sequence alignment, motif finding, block finding, similarity search
- **SCOP** (Structural Classification of Proteins)
 - Based on structural & 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

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
 - α class, β class, α/β class, $\alpha+\beta$ class
 - Architecture (A), orientation of sec. Str.
 - Topology (T), topological connections &
 - Homologous Superfamily (H), similar str and functions.

DALI/FSSP Database

- ❑ 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).

Structural Alignment

- What is structural alignment of proteins?
 - 3-d superimposition of the atoms as "best as possible", i.e., to minimize RMSD (root mean square deviation).
 - Can be done using **VAST** and **SARF**
- Structural similarity is common, even among proteins that do not share sequence similarity or evolutionary relationship.

Other databases & tools

- ❑ **MMDB** contains groups of structurally related proteins
- ❑ **SARF** structurally similar proteins using secondary structure elements
- ❑ **VAST** Structure Neighbors
- ❑ **SSAP** uses double dynamic programming to structurally align proteins

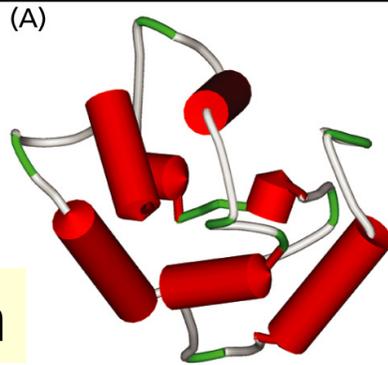
5 Fold Space classes



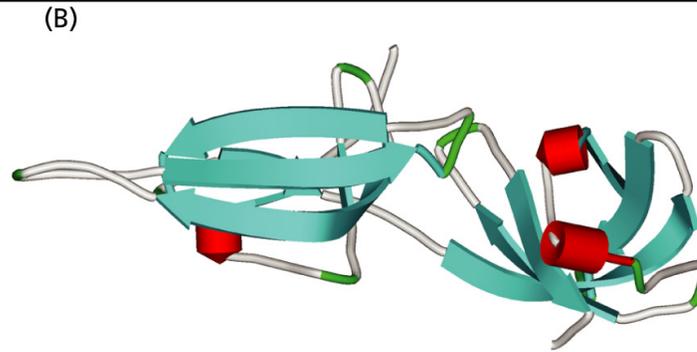
Attractor 1 can be characterized as alpha/beta, attractor 2 as all-beta, attractor 3 as all-alpha, attractor 5 as alpha-beta meander (1mli), and attractor 4 contains antiparallel beta-barrels e.g. OB-fold (1prtF).

Examples of protein classes

Parvalbumin



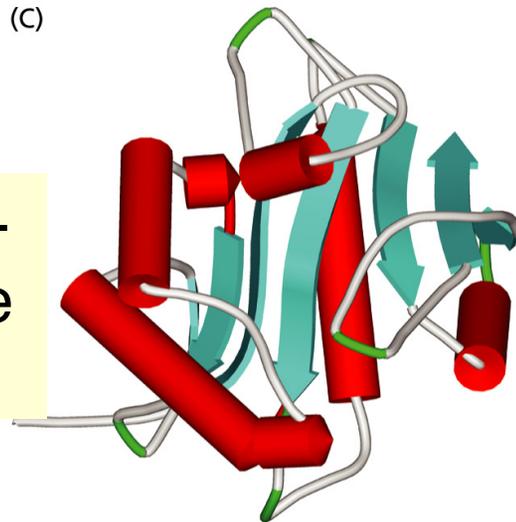
1B8C



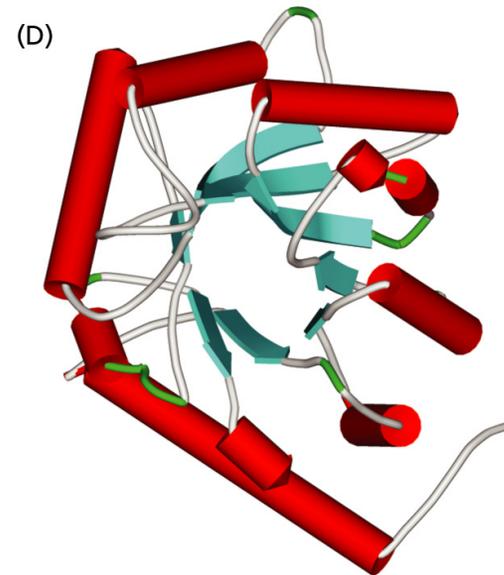
1BKB

Translation
Initiation
Factor 5A

Serotonin N-
acetyltransferase



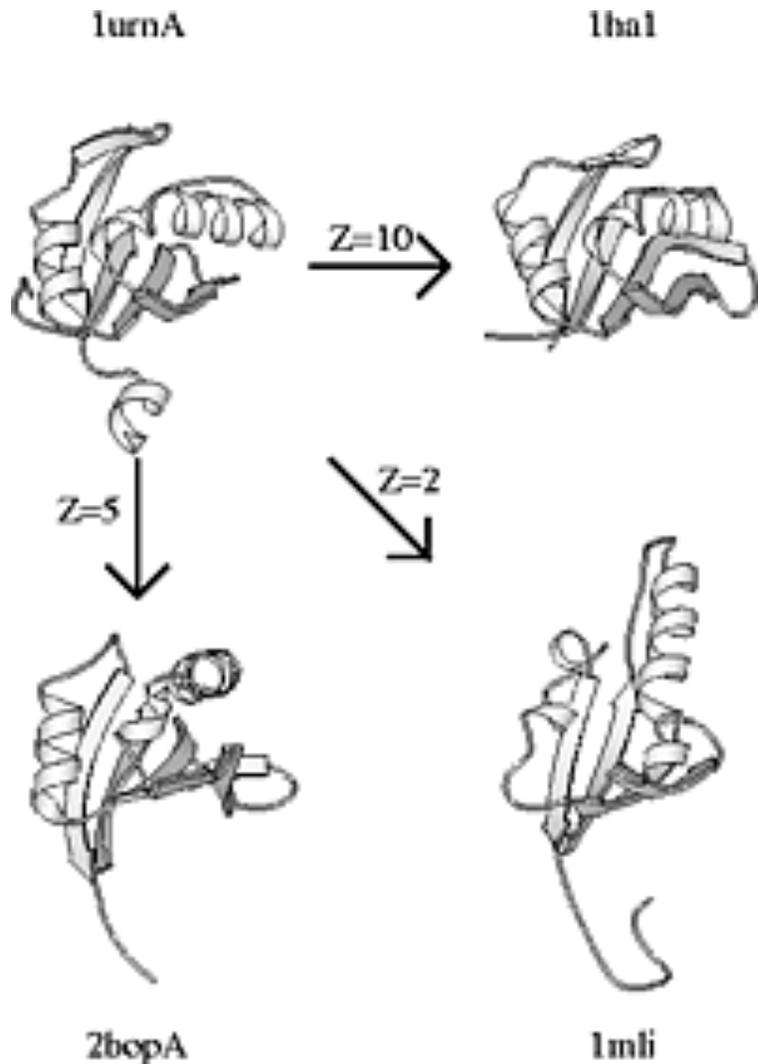
1CJW



1CT5

Hypothetical
protein from
yeast
(α/β alternating fold)

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

B

```

1urnA  --RPNHTIYINNLNEKI----KKDELKKSLSHAIFSRFG---QILDILV-SRS---LKM---
Z=10      *      *      *      *      *      *
1ha1    ahLTVKKIFVGGIKEDT-----EEHHLRDYFEOYG---KIEVIEI-MTDrgsGKK---
Z=5      *
2bopA   ----sCFALIS-GTANQ-----vKCYRFRVKKNHRHR-----YENCTTtWFT---Vadnga
Z=2
1mli    ---mLFHVKMTVKLPvdmdpakatgIkadeKELAQRlqregTWRHLWR-IAG-----

1urnA   ----RGQAFVIFKEV--SSATNALRSMQGFPPFYDKPMRIQYAKTDSDIIAKM-----
Z=10     **  ***  *      *      *
1ha1    ----RGFAFVTFDDH--DSVDKIVIQ-kYHTVNGHNCEVRKAL-----
Z=5      *      *      *      *      *      *
2bopA   erggQAQILITFGSP--SORODFLKHVPLPP---GMNISGF-----tASLDf-----
Z=2      *      *      **      *
1mli    ----HYANYSVFDVpsvEALHDTLMQLpLFPY---MDIEVD-----gLCRHps sihsddr
    
```

Frequent Fold Types



(141) 1hdcA:1
alpha/beta domain



(85) 1mfaA:3
immunoglobulin fold



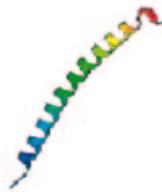
(63) 1ceo:2
TIM barrel



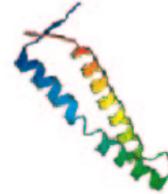
(43) 1bcfA:1
helical bundle



(36) 2pii:2
alpha/beta-meander



(33) 1vdfA:1
single helix



(27) 1grj:2
coiled coil



(25) 1bbt2:1
beta-meander



(19) 1tro:2
EF-hand



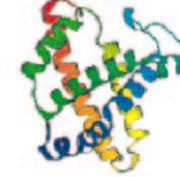
(18) 1oetC:3
HTH-motif



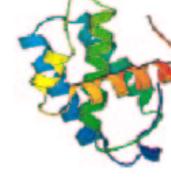
(18) 1prtF:1
OB-fold



(17) 3grs:2
FAD/NAD binding domain



(14) 1mbd:1
globin fold



(13) 1vin:3
cyclin fold



(13) 1aozA:15
blue copper protein



(13) 1lef:17
periplasmic binding protein



(12) 1eelA:3
lectin fold



(12) 1epaA:1
lipocalin fold



(12) 2arcA:4
beta-roll



(12) 2yhx:3
actin fold

Protein Structure Prediction

- ❑ **Holy Grail** of bioinformatics
- ❑ **Protein Structure Initiative** to determine a set of protein structures that span protein structure space sufficiently well. **WHY?**
 - Number of folds in natural proteins is limited. Thus a newly discovered proteins should be within modeling distance of some protein in set.
- ❑ **CASP**: Critical Assessment of techniques for structure prediction
 - To stimulate work in this difficult field

PSP Methods

- *homology*-based modeling
- methods based on *fold recognition*
 - Threading methods
- *ab initio* methods
 - From first principles
 - With the help of databases

ROSETTA

- ❑ Best method for PSP
- ❑ As proteins fold, a large number of partially folded, low-energy conformations are formed, and that local structures combine to form more global structures with minimum energy.
- ❑ Build a database of known structures (I-sites) of short sequences (3-15 residues).
- ❑ Monte Carlo simulation assembling possible substructures and computing energy

Threading Methods

☐ See p471, Mount

● http://www.bioinformaticsonline.org/links/ch_10_t_7.html

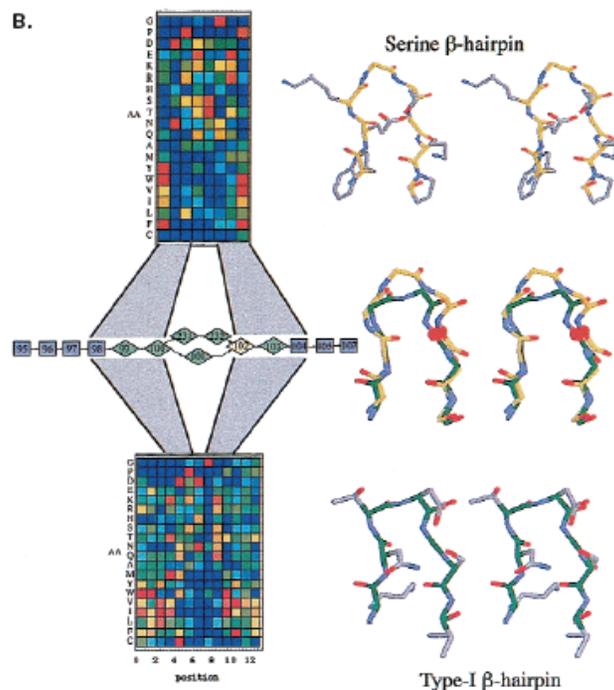


FIGURE 10.30. A hidden Markov model (discrete state-space model) of protein three-dimensional structure. (B) HMM called HMMSTR based on I-sites, 3- to 15-amino-acid patterns that are associated with three-dimensional structural features. The two matrices with colored squares represent alignment of sets of patterns that are found to be associated with a structure, in this case the hairpin turns shown on the right. Each column in the table corresponds to the amino acid variation found for one structural position in one of the turns. (Blue side chains) Conserved nonpolar residues; (green) conserved polar residues; (red) conserved proline; and (orange) conserved glycine. The two hairpins are aligned structurally in the middle structure on the right and the observed variation in the corresponding amino acid positions is represented by the HMM between the matrices on the left. The HMM represents an alignment of the two hairpin structural motifs in three-dimensional space and an alignment of the sequences. A short mismatch in the turn is represented by splitting the model into two branches. The shaped icons represent states, each of which represents a structure and a sequence position. Each state contains probability distributions about the sequence and structural attributes of a single position in the motif, including the probability of observing a particular amino acid, secondary structure, Φ - Ψ backbone angles, and structural context, e.g., location of β strand in a β sheet. Rectangles are predominantly β -strand states, and diamonds are predominantly turns. The color of the icon indicates a sequence preference as follows: (blue) hydrophobic; (green) polar; and (yellow) glycine. Numbers in icons are arbitrary identification numbers for the HMM states. There is a transition probability of moving from each state in the model to the next, as in HMMs that represent msas. This model is a small component of the main HMMSTR model that represents a merging of the entire I-sites library. Three different models, designated λ^p , λ^c , and λ^s , are included in HMMSTR, which differ in details as to how the alignment of the I-sites was obtained to design the branching patterns (topology) of the model and which structural data were used to train the model. HMMSTR may be used for a variety of different predictions, including secondary structure prediction, structural context prediction, and Φ - Ψ dihedral angle prediction. Predictions are made by aligning the model with a sequence, finding if there is a high-scoring alignment, and deciphering the highest-scoring path through the model. The HMMSTR program may be downloaded or used on a server that can be readily located by a Web search. (B, reprinted, with permission, from Bystroff et al. 2000 [©2000 Elsevier].)

Modeling Servers

- SwissMODEL
- 3DJigsaw
- CPHModel
- ESyPred3D
- Geno3D
- SDSC1
- Rosetta
- MolIDE
- SCWRL
- PSIPred
- MODELLER
- LOOPY