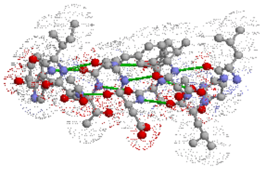
Biophysics 101:
Genomics & Computational Biology

Section 8: Protein Structure



Faisal Reza

Nov. 11th, 2003

B101.pdb from PS5 shown at left with:
• animated ball and stick model, colored CPK
• H-bonds on, colored green
• van der Waals radii on, also colored CPK

Based on the backbone and H-bond configuration shown, what secondary structure might this be?

Outline

- **Course Projects**
- **Biology/Chemistry of Protein Structure**
  - **Protein Assembly, Folding, Packing and Interaction**
  - **Primary, Secondary, Tertiary and Quaternary structures**
  - **Class, Fold, Topology**
- **CS/Math/Physics of Protein Structure**
  - **Experimental Determination and Analysis**
  - **Computational Determination and Analysis**
- **Proteomics**
- **Mass Spectrometry**

Course Projects

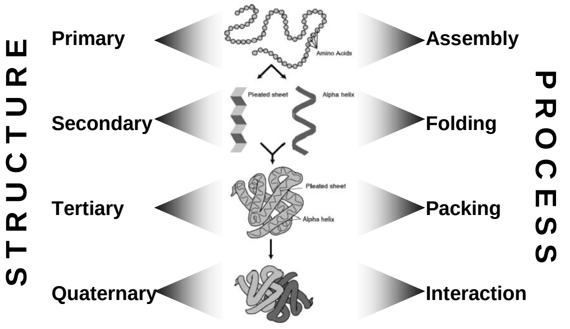
- **Videotaping authorization form**

- **Submission Parameters (via email)**
  - **when:** **December 2, 2003 12noon EST.**
    (9AM EST if presenting on December 2, 2003)
  - **where:** **bphys101@fas.harvard.edu**
  - **what:** **(1) written project (.doc, ~1000-3000 words)**
    (2) presentation slides (.ppt, 1-2 MB)

- **Presentation Parameters (in person)**
  - **when:** **December {2, 9, 16}, 2003 {12-2PM, 5:30-7:30PM} EST.**
  - **where:** **HMS Cannon Seminar Room for 12-2PM**
    Science Ctr. Lecture Hall A for 5:30-7:30PM
  - **what:** **(1) oral presentations (6 min/person + 2 min/person Q/A)**
    (2) grading rubric and further information:

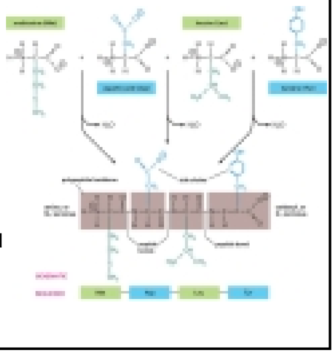
  http://www.courses.fas.harvard.edu/~bphys101/projects/index.html

Biology/Chemistry of Protein Structure



Protein Assembly

- **occurs at the ribosome**
- **involves dehydration synthesis and polymerization of amino acids attached to tRNA:**



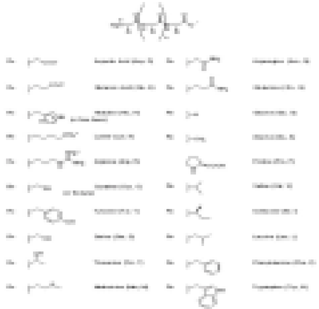
NH_3^{+}- {A + B → A-B + H_2O}$_n$-COO$^-$

- **thermodynamically unfavorable, with ΔE = +10kJ/mol, thus coupled to reactions that act as sources of free energy**
- **yields primary structure**

Primary Structure

primary structure of human insulin

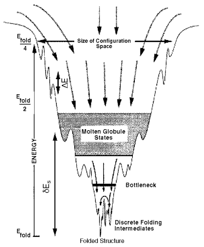
```
CHAIN 1: GIVEQ CCTSI CSLYQ LENYC N
CHAIN 2: FVNQH LCGSH LVEAL YLVCG ERGFF YTPKT
```



- **linear**
- **ordered**
- **1 dimensional**
- **sequence of amino acid polymer**
- **by convention, written from amino end to carboxyl end**
- **a perfectly linear amino acid polymer is neither functional nor energetically favorable → folding!**

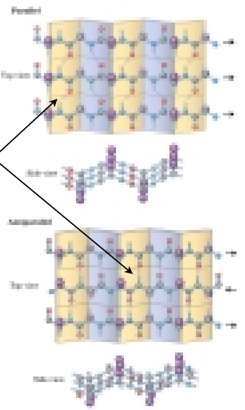
Protein Folding

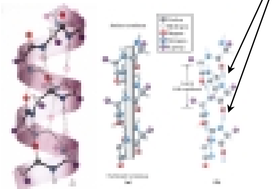
- occurs in the cytosol
- involves localized spatial interaction among primary structure elements, i.e. the amino acids
- may or may not involve chaperone proteins
- tumbles towards conformations that reduce ΔE (this process is thermodynamically favorable)
- yields secondary structure



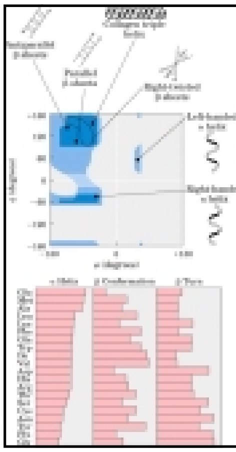
Secondary Structure

- non-linear
- 3 dimensional
- localized to regions of an amino acid chain
- formed and stabilized by hydrogen bonding, electrostatic and van der Waals interactions





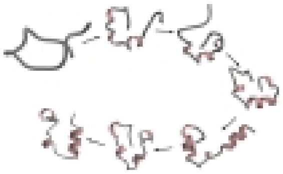
Ramachandran Plot



- Pauling built models based on the following principles, codified by Ramachandran:
  - (1) bond lengths and angles – should be similar to those found in individual amino acids and small peptides
  - (2) peptide bond – should be planer
  - (3) overlaps – not permitted, pairs of atoms no closer than sum of their covalent radii
  - (4) stabilization – have sterics that permit hydrogen bonding
- Two degrees of freedom:
  - (1) ϕ (phi) angle = rotation about N – Cα
  - (2) ψ (psi) angle = rotation about Cα – C
- A linear amino acid polymer with some folds is better but still not functional nor completely energetically favorable ➔ packing!

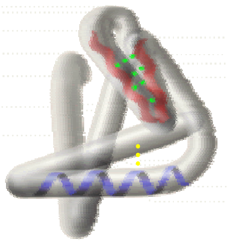
Protein Packing

- occurs in the cytosol (~60% bulk water, ~40% water of hydration)
- involves interaction between secondary structure elements and solvent
- may be promoted by chaperones, membrane proteins
- tumbles into molten globule states
- overall entropy loss is small enough so enthalpy determines sign of ΔE, which decreases (loss in entropy from packing counteracted by gain from desolvation and reorganization of water, i.e. hydrophobic effect)
- yields tertiary structure



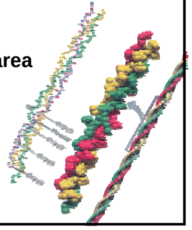
Tertiary Structure

- non-linear
- 3 dimensional
- global but restricted to the amino acid polymer
- formed and stabilized by hydrogen bonding, covalent (e.g. disulfide) bonding, hydrophobic packing toward core and hydrophilic exposure to solvent
- A globular amino acid polymer folded and compacted is somewhat functional (catalytic) and energetically favorable ➔ interaction!

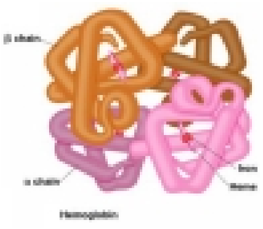


Protein Interaction

- occurs in the cytosol, in close proximity to other folded and packed proteins
- involves interaction among tertiary structure elements of separate polymer chains
- may be promoted by chaperones, membrane proteins, cytosolic and extracellular elements as well as the proteins' own propensities
- ΔE decreases further due to further desolvation and reduction of surface area
- globular proteins, e.g. hemoglobin, largely involved in catalytic roles
- fibrous proteins, e.g. collagen, largely involved in structural roles
- yields quaternary structure



Quaternary Structure



- non-linear
- 3 dimensional
- global, and across distinct amino acid polymers
- formed by hydrogen bonding, covalent bonding, hydrophobic packing and hydrophilic exposure
- favorable, functional structures occur frequently and have been categorized

Class/Motif

- class = secondary structure composition,

  e.g. all α, all β, segregated α+β, mixed α/β
- motif = small, specific combinations of secondary structure elements,

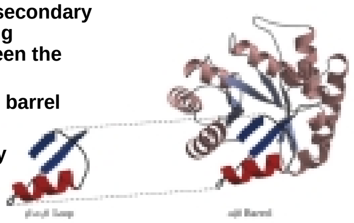
  e.g. β-α-β loop
- both subset of fold/architecture/domains

Fold/Architecture/Domains

- fold = architecture = the overall shape and orientation of the secondary structures, ignoring connectivity between the structures,

  e.g. α/β barrel, TIM barrel
- domain = the functional property of such a fold or architecture,

  e.g. binding, cleaving, spanning sites
- subset of topology/fold families/superfamilies



Topology/Fold families/Superfamilies

CLASS: α+β
FOLD: sandwich
FOLD FAMILY: flavodoxin

- topology = the overall shape and connectivity of the folds and domains
- fold families = categorization that takes into account topology and previous subsets as well as empirical/biological properties, e.g. flavodoxin
- superfamilies = in addition to fold families, includes evolutionary/ancestral properties

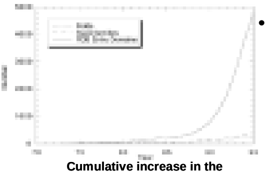
CS/Math/Physics of Protein Structure

- Experimental Determination and Analysis
- Computational Determination and Analysis

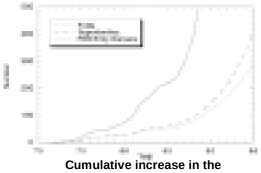
Experimental Determination and Analysis

- Repositories
  - Protein Data Bank
  - Molecular Modeling DataBase
- Resolution
  - X-Ray Crystallography
  - NMR Spectroscopy
  - Mass Spectroscopy (next week)
  - Fluorescence Resonance Energy Transfer

Protein Data Bank



Cumulative increase in the number of domains

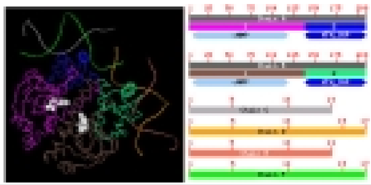


Cumulative increase in the number of folds and superfamilies

- Coordinates database
  RCSB Protein Data Bank (PDB)
  - has many structures, partly due to minor differences in structure resolution and annotation
  - has much fewer fold families, partly due to evolved pathways and mechanisms
  - .pdb = data from experiment, with missing parameters and multiple conformations

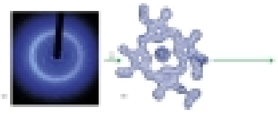
Molecular Modeling DataBase

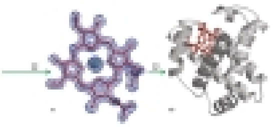
- Comparative database
  NCBI Molecular Modeling DataBase (MMDB)
  - subset of PDB, excludes theoretical structures, with native .asn format
  - .asn = single-coordinate per-atom molecules, explicit bonding and SS remarks
  - suited for computation, such as homology modeling and structure comparison



X-Ray Crystallography

- crystallize and immobilize single, perfect protein
- bombard with X-rays, record scattering diffraction patterns
- determine electron density map from scattering and phase via Fourier transform:



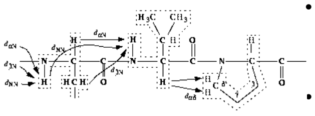


- use electron density and biochemical knowledge of the protein to refine and determine a model

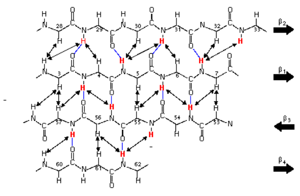
"**All crystallographic models are *not* equal.** ... The brightly colored stereo views of a protein model, which are in fact more akin to cartoons than to molecules, endow the model with a concreteness that exceeds the intentions of the thoughtful crystallographer. It is impossible for the crystallographer, with vivid recall of the massive labor that produced the model, to forget its shortcomings. It is all too easy for users of the model to be unaware of them. It is also all too easy for the user to be unaware that, through temperature factors, occupancies, undetected parts of the protein, and unexplained density, crystallography reveals more than a single molecular model shows.“

- Rhodes, “Crystallography Made Crystal Clear” p. 183.

NMR Spectroscopy



determining constraints

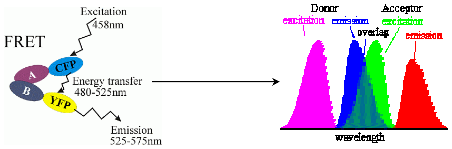


using constraints to determine secondary structure

- protein in aqueous solution, motile and tumbles/vibrates with thermal motion
- NMR detects chemical shifts of atomic nuclei with non-zero spin, shifts due to electronic environment nearby
- determine distances between specific pairs of atoms based on shifts, “constraints”
- use constraints and biochemical knowledge of the protein to determine an ensemble of models

Fluorescence Resonance Energy Transfer

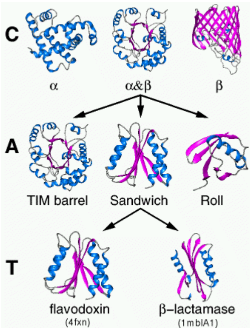
- FRET described as a “molecular ruler”
- segments of a protein are tagged with fluorophores
- energy transfer occurs when donor and acceptor interact, falls off as 1/d^6 where d is separation between donor and acceptor
- donor and acceptor must be within 50 Å, acceptor emission sensitive to distance change
- can determine pairs of side chains that are separated when unfolded and close when folded



Computational Determination and Analysis

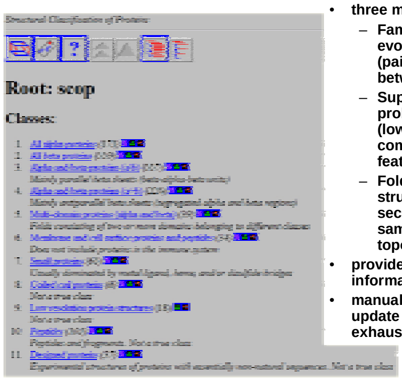
- Databases
  - CATH (Class, Architecture, Topology, Homologous superfamily)
  - SCOP (Structural Classification Of Proteins)
  - FSSP (Fold classification based on Structure-Structure alignment of Proteins)
- Prediction
  - Ab-initio, theoretical modeling, and conformation space search
  - Homology modeling and threading
  - Energy minimization, simulation and Monte Carlo
- Proteomics (next week)

CATH



- a combination of manual and automated hierarchical classification
- four major levels:
  - Class (C) – based on secondary structure content
  - Architecture (A) – based on gross orientation of secondary structures
  - Topology (T) – based on connections and numbers of secondary structures
  - Homologous superfamily (H) – based on structure/function evolutionary commonalities
- provides useful geometric information (e.g. architecture)
- partial automation may result in examples near fixed thresholds being assigned inaccurately

SCOP


Structural Classification of Proteins

Root: scop

Classes:

1. All alpha proteins (173)
2. All beta proteins (129)
3. Alpha and beta proteins (a/b) (117)
   Mainly parallel beta sheets (beta-alpha-beta units)
4. Alpha and beta proteins (a+b) (225)
   Mainly antiparallel beta sheets (segregated alpha and beta regions)
5. Multi-domain proteins (alpha and beta) (39)
   Folds consisting of two or more domains belonging to different classes
6. Membrane and cell surface proteins and peptides (34)
   Does not include proteins in the immune system
7. Small proteins (61)
   Usually dominated by metal ligand, heme, and/or disulfide bridges
8. Coiled coil proteins (6)
   Not a true class
9. Low resolution protein structures (13)
   Not a true class
10. Peptides (105)
    Peptides and fragments. Not a true class
11. Designed proteins (37)
    Experimental structures of proteins with essentially non-natural sequences. Not a true class

- a purely manual hierarchical classification
- three major levels:
  - Family – based on clear evolutionary relationship (pairwise residue identities between proteins are >30%)
  - Superfamily – based on probable evolutionary origin (low sequence identity but common structure/function features
  - Fold – based on major structural similarity (major secondary structures in same arrangement and topology
- provides detailed evolutionary information
- manual process influences update frequency and equally exhaustive examination

FSSP

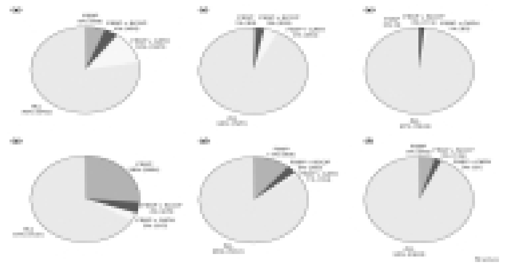
- a purely automated
- hierarchical classification
- three major levels:
  - representative set – 330 protein chains (less than 30% sequence identity)
  - clustering – based on structural alignment into fold families
  - convergence – cutting at a high statistical significance level increases the number of distinct families, gradually approaching one family per protein chain
- continually updated, presents data and lets user assess
- Without sufficient knowledge, user may not assess data appropriately

list of representative set

clustering dendogram

CATH vs. SCOP vs. FSSP

- approximately two-thirds of the protein chains in each database are common to all three databases



FSSP pairwise matches (Z-score ≥ 4.0) compared to CATH and SCOP matches at the fold level (a), homology level (b)

FSSP pairwise matches (Z-score ≥ 6.0) compared to CATH and SCOP matches at the fold level (c), homology level (d)

FSSP pairwise matches (Z-score ≥ 8.0) compared to CATH and SCOP matches at the fold level (e), homology level (f)

Ab-initio, theoretical modeling, and conformation space search

- Ab-initio = given amino acid primary structure, i.e. sequence, derive structure from first principles (e.g. treat amino acids as beads and derive possible structures by rotating through all possible ϕ, ψ angles using a “reliable” energy function, then optimize globally)
- Theoretical modeling = subset of ab-initio, given amino acid primary structure and knowledge about characteristic features, derive structure that has that structure and features (e.g. protein has an iron binding site ➔ possible heme substructure)
- Conformation space search = subset of ab-initio, but a stochastic search in which the sample space is reduced by initial conditions/assumptions (e.g. reduce sample space to conform to Ramachandran plot)

Homology modeling and threading

- Homology modeling = knowledge-based approach, given a sequence database, use multiple sequence alignment on this database to identify structurally conserved regions and construct structure backbone and loops based on these regions, restore side-chains and refine through energy minimization (apply to proteins that have high sequence similarity to those in the database)
- Threading = knowledge-based approach, given a structure database of interest (e.g. one that provides a limited set of possible structures per given sequence for fold recognition, one that provides a one structure per given limited set of possible sequences for inverse folding) use scoring functions and correlations from this database to derive structure that is in agreement (apply to proteins with moderate sequence similarity to those in the database)

Energy minimization, simulation and Monte Carlo

- Energy minimization = select an appropriate energy function and derive conformations that yield minimal energies based on this function
- Simulation = select appropriate molecular conditions and derive conformations that are suited to these molecular conditions
- Monte Carlo = subset of molecular simulation, but it is an iterated search through a Markov chain of conformations (many iterations → canonical distribution, P(particular conformation)~exp(-E/T)) proposed by N. Metropolis, in which a new conformation is generated from the current one by a small ``move'' and is accepted with a probability $P_{acc} = \min(1, \exp(-\Delta E/kT))$, which depends on the corresponding change in energy, ΔE, and on an external adjustable parameter, kT

Next Week

- Proteomics
- Mass Spectrometry

References

C. Branden, J. Tooze. “Introduction to Protein Structure.” Garland Science Publishing, 1999.

C. Chothia, T. Hubard, S. Brenner, H. Barns, A. Murzin. “Protein Folds in the All-β and ALL-α Classes.” Annu. Rev. Biophys. Biomol. Struct., 1997, 26:597-627.

G.M. Church. “Proteins 1: Structure and Interactions.” Biophysics 101: Computational Biology and Genomics, October 28, 2003.

C. Hadley, D.T. Jones. “A systematic comparison of protein structure classifications: SCOP, CATH and FSSP.” Structure, August 27, 1999, 7:1099-1112.

S. Komili. “Section 8: Protein Structure.” Biophysics 101: Computational Biology and Genomics, November 12, 2002.

D.L. Nelson, A.L. Lehninger, M.M. Cox. “Principles of Biochemistry, Third Edition.” Worth Publishing, May 2002.

.pdb animation created with PDB to MultiGif, http://www.dkfz-heidelberg.de/spec/pdb2mgif/expert.html