Protein Structure and Function
The largest interconnected group of protein families from the Swiss-Prot protein database (237 protein families; 21,727 sequences in total) is shown. Circles represent protein families. Lines show sequence similarities between families. Circles are colored according to the GeneOntology functional classes (where available). This is an example of two independent classifications, one based on sequence similarity (structural) and the other based on functional categorization.

http://www.nature.com/nrg/journal/v4/n7/fig_tab/nrg1113_F2.html
Protein Structure and Function

*Function is derived from the three-dimensional structure and the three-dimensional structure is specified by the amino acid sequence*

There are 20 common amino acids

You should know which amino acids fall into the following categories:

- hydrophobic (e.g. A, V, L, I, M, F, Y, W, P)
- polar, uncharged (e.g. S, T, N, Q, C, G)
- acidic (e.g. D, E)
- basic (e.g. K, R, H)
**HYDROPHILIC AMINO ACIDS**

**Basic amino acids**
- Lysine (Lys or K)
- Arginine (Arg or R)
- Histidine (His or H)

**Polar amino acids with uncharged R groups**
- Serine (Ser or S)
- Threonine (Thr or T)
- Cysteine (Cys or C)
- Glycine (Gly or G)
- Proline (Pro or P)

**Acidic amino acids**
- Aspartic acid (Asp or D)
- Glutamic acid (Glu or E)

**HYDROPHOBIC AMINO ACIDS**
- Alanine (Ala or A)
- Valine (Val or V)
- Isoleucine (Ile or I)
- Leucine (Leu or L)
- Methionine (Met or M)
- Phenylalanine (Phe or F)
- Tyrosine (Tyr or Y)
- Tryptophan (Trp or W)

**SPECIAL AMINO ACIDS**
- Cysteine (Cys or C)
- Glycine (Gly or G)
- Proline (Pro or P)

**FIGURE 3.2** The structures of the 20 common amino acids grouped into three categories: hydrophilic, hydrophobic, and special amino acids. The side chain determines the characteristic properties of each amino acid. Shown are the zwitterion forms, which exist at the pH of the cytosol. In parentheses are the three-letter and one-letter abbreviations for each amino acid.
Forces Influencing Protein Structure

Several different types of noncovalent interactions are important for protein structure

- **Hydrogen bonds**
  H bonds form between atoms in the peptide chain and also between peptide chain atoms and water

- **Hydrophobic interactions**
  Nonpolar side chains tend to cluster to minimize interactions with water

- **Electrostatic interactions**
  Charge-charge interactions occur between ionized side chains and between these side chains and water or salts

- **Van der Waals interactions**
  Dipole-induced dipole interactions – individually very weak but can add up to a significant stabilization energy
Hydrogen Bonds
When two atoms bearing partial negative charges share a partially positively charged hydrogen, the atoms are engaged in a hydrogen bond (H-bond). The correct 3-D structure of a protein is often dependent on an intricate network of H-bonds.
Hydrogen bonds in proteins can occur between a variety of atoms, involving:

1) atoms on two different amino acid sidechains
2) atoms on amino acid sidechains and water molecules at the protein surface
3) atoms on amino acid sidechains and protein backbone atoms
4) backbone atoms and water molecules at the protein surface
5) backbone atoms on two different amino acids

Bonds strengths can vary:
- O—H...:N (29 kJ/mol or 6.9 kcal/mol)
- O—H...:O (21 kJ/mol or 5.0 kcal/mol)
- N—H...:N (13 kJ/mol or 3.1 kcal/mol)
- N—H...:O (8 kJ/mol or 1.9 kcal/mol)

To put these energies in perspective, the energy of a C-C covalent bond is about 348 kJ/mol.
Hydrophobic interactions

In proteins, hydrophobic interactions arise as a consequence of the interaction of hydrophobic amino acids with the polar solvent, water.

The hydrophobic amino acids are gly, ala, val, leu, ile, met, pro, phe, trp. These aa's have hydrocarbon sidechains that, because of their non-polar chemistry, are forced into close association in an aqueous solvent.
Electrostatic Interactions

A. Ionic Bonds - Salt Bridges
Ionic bonds are formed as amino acids bearing opposite electrical charges are juxtaposed in the hydrophobic core of proteins. Ionic bonding in the interior is rare because most charged amino acids lie on the protein surface. Although rare, ionic bonds can be important to protein structure because they are potent electrostatic attractions that can approach the strength of covalent bonds.

B. Water Shells and Charged Surface Residues
Electrically charged amino acids, mostly found on protein surfaces, promote appropriate folding by interacting with the water solvent. Polar water molecules can form shells around charged surface residue sidechains, helping to stabilize and solubilize the protein.
Van der Waals force
This is a transient, weak electrical attraction of one atom for another. Van der Waals attractions exist because every atom has an electron cloud that can fluctuate, yielding a temporary electric dipole. The transient dipole in one atom can induce a complementary dipole in another atom, provided the two atoms are quite close. These short-lived, complementary dipoles provide a weak electrostatic attraction, the Van der Waals force. If the two electron clouds of adjacent atoms are too close, repulsive forces come into play because of the negatively-charged electrons.

The appropriate distance required for Van der Waals attractions differs from atom to atom, based on the size of each electron cloud, and is referred to as the Van der Waals radius. Van der Waals attractions, although transient and weak, can provide an important component of protein structure because of their sheer number. Most atoms of a protein are packed sufficiently close to others to be involved in transient Van der Waals attractions.
**FIGURE 2-17** The binding of a hypothetical pair of proteins by two ionic bonds, one hydrogen bond, and one large combination of hydrophobic and van der Waals interactions. The structural complementarity of the surfaces of the two molecules gives rise to this particular combination of weak bonds and hence to the specificity of binding between the molecules.
Covalent Bonds - Disulfide Bridges
Covalent bonds are the strongest chemical bonds contributing to protein structure. Covalent bonds arise when two atoms share electrons.

In addition to the covalent bonds that connect the atoms of a single amino acid and the covalent peptide bond that links amino acids in a protein chain, covalent bonds between cysteine side chains can be important determinants of protein structure. Cysteine is the sole amino acid whose side chain can form covalent bonds, yielding disulfide bridges with other cysteine side chains: --CH2-S-S-CH2-- .
Hierarchical Structure of Proteins

Four levels of structure are used to describe the shape of proteins

- **Primary structure**
  The linear arrangement or sequence of amino acids

- **Secondary structure**
  The localized organization of parts of a polypeptide chain (e.g., \( \alpha \) helix and \( \beta \) sheet)

- **Tertiary structure**
  The overall, three-dimensional arrangement of the polypeptide chain

- **Quaternary structure**
  The number and relative positions of subunits in a multimeric protein
(a) Primary structure
-Ala-Glu-Val-Thr-Asp-Pro-Gly-

(b) Secondary structure
\(\alpha\) Helix
\(\beta\) Sheet

(c) Tertiary structure

(d) Quaternary structure
(a) Primary structure
Amino end
--------------
Carboxyl end

(b) Secondary structure
Hydrogen bonds between amino acids at different locations in polypeptide chain
\( \alpha \) helix

Pleated sheet

(c) Tertiary structure
Heme

\( \beta \) polypeptide

(d) Quaternary structure
Heme group

\( \alpha \)
Fred Sanger

Sanger determined the complete amino acid sequence of the two polypeptide chains of bovine insulin, A and B, in 1952 and 1951, respectively. Prior to this it was widely assumed that proteins were somewhat amorphous. In determining these sequences, Sanger proved that proteins have a defined chemical composition. For this he received the Nobel Prize in Chemistry in 1958.

He then turned to sequencing DNA. In 1977 Sanger and colleagues introduced the "dideoxy" chain-termination method for sequencing DNA molecules. For this he received the Nobel Prize in Chemistry in 1980.
The Peptide Bond

[Chemical structure of amino acids and peptide bond]
Peptide bond rotation

- Double bond character (~40%) leads to restricted rotation
- The \textit{trans} form is favored 1000:1
- This is caused by side chain interactions
- The peptide bond preceding proline does not have double bond character
- The \textit{trans} form is favored 4:1
Rotation around bonds in a polypeptide chain

Note: rotation is only allowed about the N-C$_\alpha$ and C$_\alpha$-C=O bonds

Only structures which have “allowable” $\phi$ and $\Psi$ angles are stable

<table>
<thead>
<tr>
<th>Structure Type</th>
<th>Residues/ Turn</th>
<th>Rise (nm)</th>
<th>Number of Atoms in H-Bonded Ring</th>
<th>$\phi$ (°)</th>
<th>$\psi$ (°)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Antiparallel $\beta$ sheet</td>
<td>2.0</td>
<td>0.34</td>
<td>—$^a$</td>
<td>$-$139</td>
<td>+135</td>
</tr>
<tr>
<td>Parallel $\beta$ sheet</td>
<td>2.0</td>
<td>0.32</td>
<td>—$^a$</td>
<td>$-$119</td>
<td>+113</td>
</tr>
<tr>
<td>$3_{10}$ helix</td>
<td>3.0</td>
<td>0.20</td>
<td>10</td>
<td>$-$49</td>
<td>$-$26</td>
</tr>
<tr>
<td>$\alpha$ helix (3.613)</td>
<td>3.6</td>
<td>0.15</td>
<td>13</td>
<td>$-$57</td>
<td>$-$47</td>
</tr>
<tr>
<td>$\pi$ helix (4.416)$^b$</td>
<td>4.4</td>
<td>0.12</td>
<td>16</td>
<td>$-$57</td>
<td>$-$70</td>
</tr>
</tbody>
</table>

$^a$Bonding is between polypeptide chains.

$^b$Sterically permitted but not observed in protein.
$\alpha$-helix
**FIGURE 3-8 β sheets.** (a) A simple two-stranded β sheet with antiparallel β strands. A sheet is stabilized by hydrogen bonds (black dots) between the β strands. The planarity of the peptide bond forces a β sheet to be pleated; hence, this structure is also called a β pleated sheet, or simply a pleated sheet. (b) Side view of a β sheet showing how the R groups protrude above and below the plane of the sheet. (c) Model of binding site in class I MHC (major histocompatibility complex) molecules, which are involved in graft rejection. A sheet comprising eight antiparallel β strands (green) forms the bottom of the binding cleft, which is lined by a pair of α helices (blue). A disulfide bond is shown as two connected yellow spheres. The MHC binding cleft is large enough to bind a peptide 8–10 residues long. [Part (b) adapted from C. Branden and J. Tooze, 1991, *Introduction to Protein Structure*, Garland.]
FIGURE 6.12
The structure of silk fibroin. (a) A three-dimensional view of the stacked β sheets of fibroin, with the side chains shown in color. The region shown contains only alanine and glycine residues. (b) Interdigitation of alanine or serine side chains and glycine side chains in fibroin. The plane of the section is perpendicular to the folded sheets.
Both the $\alpha$-helix and $\beta$ sheet were proposed in 1951 by Linus Pauling.

Linus Pauling

Pauling also got two Nobel Prizes
(he was the only person ever awarded two unshared Nobel Prizes)
Ramachandran Plot

G.N. Ramachandran

Key:
- Allowed region
- Right-hand helix
- Left-hand helix
- Values of $n$
  1. Twisted β sheet — parallel or antiparallel
  2. $3_{10}$ helix
  3. $4_1$ helix
  4. 5-membered ring
  2. 2-fold ribbon

Antiparallel β sheet
Parallel β sheet
Collagen triple helix

$\beta_{\alpha}$
$\beta_{\psi}$

$\alpha_{\psi}$
$\pi$

Closed ring
2-fold ribbon
Flat ribbon

Values of $n$
- $n = 2$
- $n = 3$
- $n = 4$
- $n = 5$

ψ
0

Right-hand α helix
Left-hand α helix

+$180$
+$120$
+$60$
+$30$
+$0$
$-30$
$-60$
$-120$
$-180$
For example, repeating values of \( \phi \approx -57^\circ \) and \( \psi \approx -47^\circ \) give a right-handed helical fold (the alpha-helix). The structure of cytochrome C-256 shows many segments of helix and the Ramachandran plot shows a tight grouping of \( \phi, \psi \) angles near to \(-50, -50\).

Similarly, repetitive values in the region of \( \phi = -110 \) to \(-140 \) and \( \psi = +110 \) to \(+135 \) give extended chains with conformations that allow interactions between closely folded parallel segments (beta sheet structures). The structure of plastocyanin is composed mostly of beta sheets and the Ramachandran plot shows a broad range of values in the \(-110, +130 \) region.
Lovell et al., (Proteins: Structure, Function and Genetics 50: 437, 2003) looked at more than 500 protein structures and constructed the following Ramachandran plot.
Protein Structure Determinations

There are two principle techniques for high resolution structure

1) X-ray crystallography  ~85%
   Needs crystals
   Will work on large proteins

2) NMR  ~ 15%
   Provides structures in solutions
   Provides some measure of the dynamics
   Does not work (yet) for MW > ~40-50kDa
Crystal structures of proteins began to be solved in the late 1950’s. The first structure solved was that of myoglobin in 1957. This work was accomplished by Max Perutz and John Kendrew who received the Nobel Prize in Chemistry for this work in 1962.

A crystal has a repeating structure and the scattering of X-rays will give rise to constructive and destructive interference which results in a diffraction pattern.

Analysis of the diffraction pattern provides a map of the spatial distribution of electron densities within the crystal.

The structures of amino acids can then be fit to the electron density to yield the overall protein structure.

http://www.cs.cornell.edu/boom/2004sp/ProjectArch/AppofNeuralNetworkCrystallography/ProteinDiffraction.htm
Richard Ernst and Kurt Wüthrich were the principle developers of the multi-dimensional NMR methodologies used for structure determinations – they received the 1991 Nobel Prize in Chemistry for their contributions.

The key to their method was the Nuclear Overhauser Effect, discovered by Albert Overhauser while he was a postdoctoral fellow at the University of Illinois during 1951-1953.

Richard R. Ernst  Kurt Wüthrich  Albert W. Overhauser
The key to NMR protein structural determinations is the Nuclear Overhauser Effect (NOE) which is essentially interactions between nuclei – such as protons – which operate with a sixth power distance dependence.

Hence using appropriate methods (e.g., double irradiation) one can determine if two nuclei are close to each other. This approach allows the so-called “two-dimensional” NMR (2D-NMR) or 2D NOESY (Nuclear Overhauser Effect Spectroscopy).

Fig. 24. Scheme indicating the relations between an experimental 2D [\textsuperscript{1}H,\textsuperscript{1}H]-NOESY spectrum, a polypeptide with the chain ends indicated by N and C, sequence-specific assignments for two hydrogen atoms in the polypeptide chain indicated by circles, and the NOE upper distance constraint derived from the NOESY cross-peak connecting the chemical shift positions of the two assigned hydrogen atoms (see text).
The advent of recombinant protein methods allowed the facile introduction of stable isotopes such as carbon-13, nitrogen-15 and deuterium which led to hetero-nuclear methods – the so-called 3D triple resonance methods.
Advances in computation has reduced the time required to obtain a structure from days to minutes!

Here we see a good example of how the NMR structure can illustrate the dynamics of proteins

Fig. 1. NMR structure of the *Antennapedia* homeodomain [1]. A bundle of 20 superimposed conformers represents the polypeptide backbone. For the polypeptide segment 7–59 the tight fit of the bundle indicates that the structure is defined with high precision, whereas the two chain ends are disordered.
Another way to view the spatial distribution of the residues
The primary protein database for protein structure information is the **Protein Data Bank**, created sometime in the beginning of the 1970ties. Only few structures existed at that time.

One of the reasons for the large increase in available structures was that cloning techniques began which greatly increased the number and amount of proteins available for crystallization. Another important factor was the introduction of synchrotron radiation. Synchrotrons provide very high intensity X-rays, which may be used for collecting high quality X-ray diffraction data from crystals. The third factor was probably the introduction of personal computers, relatively cheap and with ever increasing power. Cheaper computers also meant new software, which also started to become user friendly. Then came the era of **structural genomics** – large consortia were formed with the aim to develop new technology for solving huge amounts of protein structures.

Source: The Protein Data Bank
104,034 as of Friday Sept 11, 2015 at 10 AM
Cryoelectron microscopy
What is cryo EM?

- EM = (Transmission) Electron Microscopy
- Cryo EM = technique where biological samples are preserved in vitreous ice and imaged by EM at cryogenic temperatures.
- EM reconstruction = 3D maps are generated by averaging over many EM images.
Growth of EM Method for Determining Structures of Macromolecular Assemblies

- coordinates, PDB
- maps, EMDB

- bacteriorhodopsin
- recA hexamer
- rhinovirus-receptor complex
- 70S ribosome
- acetylcholine receptor

Year:
- 1990
- 1995
- 2000
- 2005
Unified Data Resource for 3-Dimensional Electron Microscopy

EMDataBank is a unified global portal for deposition and retrieval of 3DEM density maps, atomic models, and associated metadata, as well as a resource for news, events, software tools, data standards, validation methods for the 3DEM community.

Announcing the 2015 EMDataBank Map Challenge

All members of the Scientific Community—at all levels of experience—are invited to participate as Challengers, and/or as Assessors.

EMPIAR: Raw Data Image Archive

EMPIAR (Electron Microscopy Pilot Image Archive; http://pdbe.org/empiar) is a public archive for raw image data (micrographs, particles, and tilt-series) related to EMDB entries. Important uses of EMPIAR data include validation and as test-data for methods development and teaching.
Small Angle Scattering (SAS)

Monochromatic beam

Radiation sources:
- X-ray generator ($\lambda=0.1 - 0.2$ nm)
- Storage ring ($\lambda=0.03 - 0.4$ nm)
- Neutron reactor ($\lambda=0.1 - 1$ nm)
**Rigid Body Modeling**

- SANS: multiphase
- SAXS: single phase

**Ab initio Dummy-Residue/Bead Modeling**

- SANS: multiphase
- SAXS: single phase

**Superposition of bead and rigid body models**

- SANS
- SAXS

**Rigid Body Modeling with Unknown Regions**

- C0, C1, C2, C3, C4

**Overall structural parameters**

- $R_g = 14.5 \text{ Å}$
- $D_{max} = 45 \text{ Å}$
- $MW = 14.3 \text{ kDa}$

**Data reduction**

- Solvent subtraction

**Guinier plot**

- $l(q)$ vs. $q (\text{Å}^{-1})$

**Comparison with known structures**

- $l(q)$ vs. $q (\text{Å}^{-1})$
Predicting Structures

Experimental structures are currently available for less than 1/1000th of the proteins for which sequence is known, so modeling has a major role to play in providing structural information for a wide range of biological problems.

In recent years, a number of algorithms for predicting secondary – and even tertiary – structures have appeared.
Predicting Structures

One of the first of these was based on work by P.Y. Chou and G.D. Fasman in the 1970’s.

The Chou-Fasman algorithm for secondary structure prediction is still widely used. It is based on the frequencies of particular amino acid residues in α–helices, β-sheets and turns. It is about 50-60% accurate.

http://www.imtech.res.in/raghava/betatpred/chou.html
A newer approach is the GOR method

- developed by Garnier, Osguthorpe & Robson
- build on Chou-Fasman $P_{ij}$ values
- evaluate each residue PLUS adjacent 8 N-terminal and 8 carboxyl-terminal residues
- sliding window of 17
- underpredicts β-strand regions
- GOR IV method accuracy ~64%
The newest methods use Neural network models
- machine learning approach
- provide training sets of structures (e.g. $\alpha$-helices, non $\alpha$-helices)
- computers are trained to recognize patterns in known secondary structures
- provide test set (proteins with known structures)
- accuracy ~ 70 –75%

Critical Assessment of Techniques for Protein Structure Prediction
http://predictioncenter.org
The Critical Assessment of protein Structure Prediction (CASP) experiments aim at establishing the current state of the art in protein structure prediction, identifying what progress has been made, and highlighting where future effort may be most productively focused.
FIGURE 3-5 Various graphic representations of the structure of Ras, a guanine nucleotide-binding protein. Guanosine diphosphate, the substrate that is bound, is shown as a blue space-filling figure in parts (a)-(d). (a) The Cα trace of Ras, which highlights the course of the backbone. Evident from this view is how the polypeptide is packed into the smallest possible volume. (b) Ball-and-stick model of Ras showing the location of all atoms. (c) A schematic diagram of Ras showing how β strands (arrows) and α helices (cylinders) are organized in the protein. Note the turns and loops connecting pairs of helices and strands. (d) The water-accessible surface of Ras. Painted on the surface are regions of positive charge (blue) and negative charge (red). Here we see that the surface of a protein is not smooth but has lumps, bumps, and crevices. The molecular basis for specific binding interactions lies in the uneven distribution of charge over the surface of the protein. [Adapted from L. Tong et al., 1991, J. Mol. Biol. 217:503; courtesy of S. Choe.]
Protein Domains

Analysis of protein structures typically begins with decomposition of the structure into more basic units called structural domains. The underlying goal is to reduce a complex protein structure to a set of simpler, yet structurally meaningful units, each of which can be analyzed independently.

The total number of domains recognized is currently around 1000. These range in size from 36 residues in E-selectin to 692 residues in lipoxygenase-1, but the majority, 90%, have less than 200 residues with an average of approximately 100 residues.
Beta-propeller domain

Characterized by 4-8 blade-shaped beta sheets arranged around a central axis. Each sheet typically has four antiparallel β-strands twisted so that the first and fourth sheets are almost perpendicular to each other.

Kringle domains

Characterized by large loops stabilized by 3 disulfide linkages. The name comes from the Scandinavian pastry.
β-barrels

α/β barrels

α/β horseshoe

Rossman Fold
The BAR domain is banana shaped and binds to membrane via its concave face. It is capable of sensing membrane curvature by binding preferentially to curved membranes. The BAR interaction with membranes is largely electrostatic and binds to negatively charged membranes.
The PH domain (Pleckstrin homology domain) has approximately 120 amino acids that occurs in a wide range of proteins involved in intracellular signaling or as constituents of the cytoskeleton.
Secondary Structure Motifs

A motif is built up from particular combinations of secondary structures

The coiled-coil motif is characterized by two or more helices wound around each other. Each helix has a *heptad repeat* sequence such as:

$$\text{HPPHPPPH}$$

with a leucine or other hydrophobic residue (H) at positions 1 and 4 (shown in red in the figure) and polar residues (P) in between, which forms a hydrophobic strip along the helix surface. Helices pair by interacting along these hydrophobic stripes.
The helix-loop-helix motif (b) occurs in many calcium binding proteins - oxygen containing groups of residues in the loop form a ring around a Ca\(^{2+}\) ion.

The zinc-finger motif (c) is found in many proteins that bind nucleic acids. A Zn\(^{2+}\) ion is held between a pair of \(\beta\) strands (green) and a single \(\alpha\) helix (blue) by a pair of cysteine and histidine residues.
The Greek-key motif - named for its similarity to the meander patterns found in ancient Greek vases – is very common structural feature.

A Greek-key is a series of 4 consecutive B strands.
Examples of proteins that contain different individual protein modules:

G – epidermal growth-factor-like module

K – the kringle domain (triple-looped disulfide cross-linked domains)

C – found in complement proteins

F1, F2 and F3 – first found in fibronectin

I – the immunoglobulin superfamily domain

E – a module homologous to the calcium binding domain

LB – a lectin module found in some cell surface proteins
Figure 1 | Arrangement of LRRK2 domains

The predicted size and number of repeat subunits for repeat domains are summarized to scale. aa, amino acids; COR, C-terminal of ROC.

Linear arrangement of predicted domains in LRRK2

<table>
<thead>
<tr>
<th>Domain</th>
<th>Size (aa x repeats)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Armadillo</td>
<td>609 (42 aa x 13)</td>
</tr>
<tr>
<td>ankyrin</td>
<td>227 (33 aa x 7)</td>
</tr>
<tr>
<td>LRR</td>
<td>337 (24 aa x 14)</td>
</tr>
<tr>
<td>Roc</td>
<td></td>
</tr>
<tr>
<td>COR</td>
<td></td>
</tr>
<tr>
<td>kinase</td>
<td></td>
</tr>
<tr>
<td>WD-40</td>
<td>352 (40 aa x 7)</td>
</tr>
</tbody>
</table>

Homology models of predicted repeat domains in LRRK2

- Armadillo
- Ankyrin
- LRR
- WD-40

Key mutations:
- A211V, E334K, R1067Q
- S1098C, S1228T, I1122V
- K544E, P755L, R753M, M712V
- T2358I, G2385R
Intrinsically disordered proteins are characterized by the lack of a stable structure. Even though it is commonly accepted that the functional role of a protein is highly related to its three-dimensional structure, an increasing number of proteins are being identified as disordered yet still biologically functional.

It has been proposed that the flexibility of disordered proteins facilitates the different conformational requirements for binding the modifying enzymes as well as their receptors. Intrinsic disorder is particularly enriched in proteins implicated in cell signaling, transcription and chromatin remodeling functions. [1][2][3]

A search of “intrinsically disordered proteins” on pubmed turned up 1940 hits; including 1546 since 2010.
Intrinsically unstructured proteins have been implicated in a number of diseases. Aggregation of misfolded proteins has been implicated in these diseases. The aggregation of the intrinsically unstructured protein α-Synuclein is thought to be responsible for many. The structural flexibility of this protein together with its susceptibility to modification in the cell leads to misfolding and aggregation.

The primary method to obtain information on disordered regions of a protein is NMR spectroscopy. The lack of electron density in X-ray crystallographic studies may also be a sign of disorder.
Integral Membrane Proteins

These proteins have one or more segments embedded in the phospholipid bilayer. The segments inserted into the bilayer have hydrophobic side chains that interact with the fatty acyl groups of the membrane lipids. Most of the integral membrane proteins examined so far use single or multiple α helices to span the bilayer.

Shown below are examples of a single α helical span and the very common seven membrane-spanning α helical motif.
Structure of the protein subunits in the photosynthetic reaction centre of Rhodopseudomonas viridis at 3Å resolution.

The Nobel Prize in Chemistry 1988

Johann Deisenhofer  Robert Huber  Hartmut Michel
Some last comments about quaternary structures

It is common to hear researchers discuss their proteins of interest and say something like: “My protein is a dimer” or “My protein is a tetramer”

For example, we all “know” that hemoglobin is a tetramer!

However – anytime we consider an oligomeric protein we MUST realize that the oligomerization state (e.g., monomer dimer, tetramer, etc) will depend upon the protein concentration and the dissociation constant of the oligomer.
Consider a dimeric protein.

If the proteins associate non-covalently there will always be a dimer/monomer equilibrium as shown below:

This process is described by a “dissociation constant” or $K_D$

If, for example, the $K_D$ for this dimer/monomer equilibrium was 1 micromolar it means that at a concentration of 1 micromolar the dimer is half-dissociated into monomers.
In quantitative terms:

\[ [D] \Leftrightarrow 2[M] \]

Where \([D]\) and \([M]\) correspond to dimer and monomer concentrations

\[ K_D = \frac{[M]^2}{[D]} \]

Where \(\alpha\) is the degree of dissociation (equal to 1 for complete dissociation and 0 for complete association) and \(C\) is the total protein concentration.
The solid circles correspond to a $K_D$ of 1 micromolar; the triangles correspond to a $K_D$ of 100 micromolar and the open circles correspond to a $K_D$ of 0.01 micromolar.
Back to hemoglobin:

The $K_D$ for the tetramer/dimer dissociation for oxygenated hemoglobin is around $1.2 \times 10^{-6}$ M.

But the $K_D$ for the case of deoxy hemoglobin is much lower, around $3.5 \times 10^{-11}$ M.

Hence the binding of oxygen to hemoglobin dramatically weakens the subunit interactions.

Clearly, the oligomerization state of hemoglobin will depend upon its state of oxygenation and its concentration.

At $10^{-8}$ M oxygenated hemoglobin will be largely dimeric while $10^{-8}$ M deoxyhemoglobin will be largely tetrameric.
Most of the information which determines the three-dimensional structure of a protein is carried in the amino acid sequence. This principle was first elucidated in classic experiments on ribonuclease by Christian Afinsen (Nobel Prize 1972).

a) depiction of conformational changes which occur when ribonuclease is heated above its denaturation temperature

b) the unfolding and refolding can be followed using several different physical methods including solution viscosity, optical rotation and UV absorption
A very common approach in protein folding studies is the use of chemical denaturants such as urea or guanidinium hydrochloride as shown on the right.

In addition to temperature and chemical denaturants, elevated pressure has also been shown to reversibly denature proteins as shown below.
From James et al., (2011) Analytical Biochemistry 410:70
In 1969 Cyrus Levinthal pointed out that so many conformations are available for a typical protein that the protein does not have sufficient time to reach its stable, final conformation by sampling all possible conformations.

This argument is known as “Levinthal’s paradox” - namely, consider a protein with 100 amino acids, assume only 2 conformational position per amino acid hence there are $2^{100}$ or $1.27 \times 10^{30}$ possibilities; allow $10^{-13}$ sec to test each possible conformation – this process would take $4 \times 10^9$ years!

This paradox led protein chemists to hypothesize that proteins fold by specific “folding pathways” which involve intermediate or partially folded states such as that shown below.
Models for protein folding:
(a) Framework model
(b) Hydrophobic collapse model
(c) Nucleation–condensation mechanism

(a) Formation of elements of secondary structure → Assembly of secondary structure
(b) Hydrophobic collapse → Growth of secondary structure → Folded conformation
(c) Nucleation–condensation → Hierarchical assembly

Unfolded state → Folding nucleus
Energy Surfaces and Protein Folding Conformations

Figure (a) above depicts a Levinthal-type energy landscape where $N$ is the native conformation and $D$ is the denatured state. In this scenario the chain searches for $N$ randomly.

Figure (b) above depicts an energy landscape which tends to “funnel” the protein towards the native state. Many folding paths are possible.
A more recent “brute-force” approach to protein folding involves “massively parallel computing” The “Folding@Home” project using the computing power from thousands of PCs

Local Structure Formation in Simulations of Two Small Proteins

Guha Jayachandran\textsuperscript{a,b}, V. Vishal\textsuperscript{b,c}, Angel E. Garcia\textsuperscript{d}, and Vijay S. Pande\textsuperscript{c}

\textsuperscript{a}Computer Science Department, Stanford University, Stanford, CA 94305, U.S.A.
\textsuperscript{b}Chemistry Department, Stanford University, Stanford, CA 94305, U.S.A.
\textsuperscript{c}Department of Physics, Applied Physics, and Astronomy, Rensselaer Polytechnic Institute, Troy, NY 12180, U.S.A.

Abstract

Massively parallel all-atom, explicit solvent molecular dynamics simulations were used to explore the formation and existence of local structure in two small alpha-helical proteins, the villin headpiece and the helical fragment B of protein A. We report on the existence of transient helices and combinations of helices in the unfolded ensemble, and on the order of formation of helices, which appears to largely agree with previous experimental results. Transient local structure is observed even in the absence of overall native structure. We also calculate sets of residue-residue pairs that are statistically predictive of the formation of given local structures in our simulations.

Keywords

molecular dynamics; simulation; protein folding; alpha helix; distributed computing

1. Introduction

Recent years have seen great advances in both experimental and computational methods for studying protein folding. Molecular dynamics simulation is among the techniques that have been at the vanguard of this wave of progress. In the years since Duan and Kollman’s (1998) landmark supercomputer trajectory for the villin headpiece, small peptides and proteins have been folded with molecular dynamics and initial comparisons with experiment made (Zagrovic et al., 2002; Garcia and Omchic, 2003; Snow et al., 2004; Rhee et al., 2004).

The use of model systems has been central to our progress in elucidating protein folding. Of particular utility have been proteins whose sizes and folding time scales are accessible to experiment and computation both. As noted, villin has been one of these systems. Zagrovic et al. (2002) reported folding the 36-residue protein in implicit solvent using massively parallel simulation. We have recently achieved this with explicit solvent (Jayachandran et al., 2006). Recent experiments have shed some light on the existence of structure in the denatured state of villin (Tang et al., 2004; Brewer et al., 2005). The B domain of protein A has been another important target for both experiment (Bai et al., 1997; Myers and Oas, 2001; Vu et al., 2004; Sato et al., 2004) and computation—ranging from weighted histogram analysis and use
Figure 1. Cartoon depiction of cotranslational folding of a polypeptide. The nascent polypeptide is shown assuming secondary structure as it emerges from the ribosome during the process of biosynthesis. The earliest intermediate, $I_1$, is not well-stabilized by extensive tertiary interactions and is in equilibrium with multiple conformations. The second intermediate shown, $I_2$, is the N-terminal domain; more extensive tertiary interactions will allow this intermediate to be more stable. The final intermediate, $I_3$, depicts the structure of the full-length polypeptide immediately prior to release from the ribosome with the C-terminal domain not yet fully packed. Chaperones and/or folding catalysts ($CH$ and $FC$) may interact with either the nascent intermediate structures or with the full-length product ($M^*$) following release from the ribosome. The final stages of folding from $M^*$ to native monomer, $M_n$, occur following release. Association of monomeric units into oligomeric structures, $O$, may occur posttranslationally as depicted or, as discussed in the text, may involve nascent polypeptides. The structure that was used to develop this cartoon was of one subunit of the bacterial luciferase 2 homodimer.

From: Fedorov and Baldwin, JBC, 1997 Vol. 272, 32715
Chaperone-mediated folding

(a) Many proteins fold into their proper three-dimensional structure with the assistance of Hsp70, a molecular chaperone that transiently binds to a nascent polypeptide as it emerges from a ribosome. Proper folding of some proteins also depends on the chaperonin TCP, a large barrel-shaped complex of Hsp60 units. (b) GroEL, the bacterial homolog of TCP, is a barrel-shaped complex of 14 identical 60,000-MW subunits arranged in two stacked rings. In the absence of ATP or presence of ADP, GroEL exists in a "tight" conformational state (left) that binds partially folded or misfolded proteins. Binding of ATP shifts GroEL to a more open, "relaxed" state (right), which releases the folded protein. [Part (b) from A. Rosemen et al., 1996, Cell 87:241. Courtesy of Helen Saibil.]
Misfolded Proteins and Disease

Diseases caused by lack of a particular functioning protein, due to its degradation as a consequence of misfolding, include cystic fibrosis (misfolded CFTR protein), Marfan syndrome (misfolded fibrillin), Fabry disease (misfolded alpha galactosidase), Gaucher’s disease (misfolded beta glucocerebrosidase) and retinitis pigmentosa 3 (misfolded rhodopsin).

Many protein misfolding diseases are characterized not by disappearance of a protein but by its deposition in insoluble aggregates within the cell. Diseases caused by protein aggregation include Alzheimer’s disease (deposits of amyloid beta and tau), Type II diabetes (deposits of amylin), Parkinson’s disease (deposits of alpha synuclein), Huntington’s disease, and the spongiform encephalopathies such as Creutzfeldt-Jakob disease (deposits of prion protein). Hereditary transthyretin amyloidosis is caused by the deposition of transthyretin in various tissues, eg the heart (leading to congestive heart failure), or the nerves (leading to peripheral neuropathy).
The diagnostic feature common to the protein aggregation diseases is the deposition of insoluble protein aggregates called amyloid fibrils, hence the generic term amyloidosis. The fibrils may themselves be assembled into discrete bodies, termed plaques.

An interesting website which discusses protein folding and protein folding diseases is: http://www.nature.com/horizon/proteinfolding/background.html
**Prions**

*Prions* are pathogenic proteins which cause fatal degenerative diseases in mammals, such as “mad cow disease” and scrapie. Prions are also believed to be responsible for the human diseases kuru and Creutzfeld-Jacob disease (for an excellent review read “Prions” by S. Prusiner (1998) PNAS 95:13363).

The prion protein – PrP comes in various forms such as PrPc, the normal cellular form, and PrPsc, the scrapie form which is a conformational variant.

The two forms are thought to differ only in their secondary structure with PrPc (a) dominated by $\alpha$-helical elements and PrPsc (b) having both $\alpha$-helices and $\beta$-strands. It is believed that PrPsc can bind to PrPc and cause it to adopt the PrPsc conformation.
Post-translational Modification of Proteins

Nearly every protein in a cell is chemically altered after its synthesis on a ribosome. These modifications can be categorized as either (1) chemical modifications or (2) processing.

Chemical modifications include:
(1) acetylation of N-terminal residues

(2) phosphorylation of serine, threonine and tyrosine residues (very common regulatory mechanism)

(3) glycosylation of asparagine, serine and threonine residues

(4) attachment of lipids

Other less common modifications are shown above.
N & O glycoproteins

Examples:

(a) O-linked

(b) N-linked

Examples:
Figure 1. The antigens corresponding to the A, B and O blood groups. The A and B antigens differ from the H antigen by the addition of different monosaccharides to the H antigen (corresponding to blood group O). Remarkably, the A and B antigens differ only by the replacement of an acetamido group (\(-\text{NHCOC}_{3}\)) in A with a hydroxyl group (\(-\text{OH}\)) in B.
Anchoring of Proteins to Membranes
Protein Processing

Proteolytic cleavage is the most common form of protein processing

Some examples include enzymes involved in blood coagulation and digestion

Also, such processing generates active peptide hormones such as insulin and EGF

The processing of preproinsulin into proinsulin and finally into insulin is depicted in the figure to the right
Self-Splicing

Some bacteria and lower eukaryotes have a self-splicing protein-processing mechanism.

A section of a protein – called an intein – is removed and a new peptide bond is formed between the two segments.

The process is autocatalytic, i.e., it does not require additional enzymes.
**In Vivo Protein Turnover**

Proteins undergo continuous biosynthesis and degradation.

An average person will synthesize ~ 400 grams of protein per day and ~ 400 grams of protein will be broken down each day.

<table>
<thead>
<tr>
<th>Half-life (hours)</th>
<th>Nucleus</th>
<th>Cytosol</th>
<th>Mitochondria</th>
<th>Endoplasmic Reticulum and Plasma Membrane</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;2</td>
<td>Oncogene products</td>
<td>Ornithine decarboxylase, tyrosine aminotransferase, protein kinase C</td>
<td>δ-Aminolevulinic acid synthetase</td>
<td>HMG-CoA reductase</td>
</tr>
<tr>
<td>2–8</td>
<td>—</td>
<td>Tryptophan oxygenase, cAMP-dependent protein kinase</td>
<td>—</td>
<td>γ-Glutamyl transferase</td>
</tr>
<tr>
<td>9–40</td>
<td>Ubiquitin</td>
<td>Calmodulin, glucokinase</td>
<td>Acetyl-CoA carboxylase, alanine aminotransferase</td>
<td>LDL receptor, cytochrome P450</td>
</tr>
<tr>
<td>41–200</td>
<td>Histone H1</td>
<td>Lactate dehydrogenase, aldolase, dihydrofolate reductase, phytochrome P670</td>
<td>Cytochrome oxidase, pyruvate carboxylase, cytochrome c</td>
<td>Cytochrome b&lt;sub&gt;5&lt;/sub&gt;, cyt b&lt;sub&gt;5&lt;/sub&gt; reductase</td>
</tr>
<tr>
<td>&gt;200</td>
<td>Histones H2A, H2B, H3, H4</td>
<td>Hemoglobin, glycogen phosphorylase</td>
<td>—</td>
<td>Acetylcholine receptor</td>
</tr>
</tbody>
</table>
Protein Degradation Pathways

Protein degradation in cells is compartmentalized either in degradative organelles such as lysosomes or in macromolecular structures known as proteasomes.

Proteasomes, found in both prokaryotic and eukaryotic cells, are large oligomeric structures enclosing a central cavity where proteolysis takes place.

Below is the structure of the proteosome from an archaebacterium which consists of four stacked rings with $\alpha_7\alpha_7\beta_7\beta_7$ subunit organization. The $\alpha$ rings unfold the target protein and transport it to the central cavity - the $\beta$ subunits possess the proteolytic activity. The products of the degradation are oligopeptides 7 to 9 residues long.
Ubiquitination is the most common mechanism to label a protein for proteasome degradation

Ubiquitin is a small (76 residue) protein found in all eukaryotic cells. It undergoes an ATP dependent reaction with proteins which condenses the C-terminal glycine residues of ubiquitin with lysine amino groups on the target protein.

The proteasome shown here is the eukaryotic 26S proteasome which has two additional substructures known as the 19S caps.
Mass Spectrometry: What is it and What Does it mean to the Biologist?

• In the last 20 years two Nobel Prizes have been won for work on mass spectroscopy. 1989 Hans Dehmelt & Wolfgang Paul. “Development of the Ion Trap”. 2002 John Fenn, Koichi Tanaka & Richard Smalley. “Soft Ionization Techniques including Electro-spray”.

• Every year more than 3500 peer reviewed journal articles are published in this field. Most papers are published in the field of Biological Applications.

• Proteomics represents a new branch of science in which MS plays a pivotal role.
MALDI
Matirx Assisted Laser Desorption Ionization

The basic idea is that proteins (or protein digests) are embedded into a sample grid and then blasted from the grid matrix by a laser which ionizes and displaces them.

The resultant ions are then accelerated towards an electrode and impact a detector. The charge-to-mass ratio of the ions will determine their “time of flight” which can then be used to identify the molecule.
One approach is to digest the target protein with proteolytic enzymes and then analyze the fragments using MS. The fragments can then be immediately injected into another MS (tandem MS) for further analysis.
The technique has progressed to where intact proteins can be directly analyzed.

Horse Heart Myoglobin Protein Charge Envelope

Maximum Entropy Deconvolution

Theoretical 16951.499
Measured 16951.512

Courtesy of MicroMass
Spectroscopic Properties of Proteins

- All amino acids absorb in the infra-red
- Only Phe, Tyr and Trp absorb ultraviolet (UV)
- Absorbance at 280nm is often used to quantify protein concentrations
Green Fluorescent Protein

**Aequorea victoria** jellyfish

Osamu Shimomura

70 mgs of purified GFP were obtained. The 30,000 jellyfish weighed about 1.5 tons.

The outer ring of the jellyfish had to be isolated. Initially scissors were used but then a “ring-cutting” machine was built.
Secondary/3-D structure of the GFP

Beta-barrel isolates the intrinsic chromophore from the solvent

- Any proteolysis destroys fluorescence
- The fluorescence is pH dependent
- The emission is highly dependent on molecular interactions of the probe with the immediate surroundings
Ryuzo Yanagimachi
Another increasingly popular fluorescent protein is DsRed - originally isolated from the Indo-Pacific sea anemone relative *Discosoma* species.
A number of new GFP proteins have been made using site directed mutagenesis to alter the amino acids near the chromophore and thus alter the absorption and fluorescence properties. These include EGFP (for enhanced GFP), YFP, CFP and others.

Note: HcRed is a tetrameric protein from Anthozoa coral Heteractis crispa
The 2004 palette of nonoligomerizing fluorescent proteins

- **GFP-derived**
  - Exc.: 380 / 433/452
  - Em.: 440 / 475/505
  - High QY (~0.7), good FRET acceptor; acid-quenched, usable as exocytosis indicator

- **mRFP1-derived**
  - Exc.: 540
  - Em.: 553
  - Highest overall brightness ($\varepsilon \times$ QY), but twice the MW
  - Closest successor to mRFP1; higher $\varepsilon$, faster maturing, several-fold more photostable
  - Easily and reversibly photoisomerizable by 470 nm illumination
  - Longest emission wavelength, largest Stokes’ shift, quite photostable

**Evolved by SHM**
- Exc.: 596 / 605 / 590 nm
- Em.: 625 / 636 / 648 nm

Nathan Shaner, Lei Wang, Paul Steinbach
2008 Nobel Prize in Chemistry

Osamu Shimomura

Martin Chalfie

Roger Y. Tsien