Fundamentals of Protein Structure

I690/B680: Structural Bioinformatics

January 11, 2006
The Logic of Biological Phenomena

Facts

• molecules are lifeless
• in appropriate complexity and number, molecules compose living things
• the organisms consist of cells
• the cells consist of (bio)molecules that must conform to the physical and chemical principles that govern all matter

What makes living things distinct?

• they can grow
• they can move
• they can replicate themselves
• they can respond to stimuli
• they can perform metabolism

[Cross-Section of an Animal Cell]
More Formally

- **living organisms are complex and highly organized**
  - organism → cell → subcellular structures (organelles) → polymeric molecules (macromolecules) → building blocks (e.g. sugars, amino acids)
  - 3-D structure of a molecule aka its *conformation*

- **biological structures serve functional purposes**
  - biological *purpose* can be given for each component

- **living systems are actively involved in energy transformations**
  - extract energy from the environment (ultimate source is the sun)
  - organisms capture sun’s energy (from photosynthesis of metabolism of food)
  - the living state is characterized by the flow of energy through the organism
  - energy is spent to maintain a steady-state (which is very dynamic)

- **living systems can self-replicate**
  - e.g. simple division by bacteria, sexual reproduction by plants or animals
  - it is *high-fidelity* reproduction, but *imperfect* (enables evolution)
Biomolecules

• **Some facts**
  – hydrogen (H), oxygen (O), carbon (C), and nitrogen (N) constitute >99% of human body
  – most of the H and O occur as $\text{H}_2\text{O}$

• **Why are H, O, C, N so suitable to chemistry of life?**
  – they can form covalent bonds by electron-pair sharing!
  – also, H, C, N, and O are among the lightest elements of the periodic table capable of forming covalent bonds $\Rightarrow$ they form the strongest bonds
  – phosphorus (P) and sulfur (S) are also important in biomolecules

• **So, what are biomolecules?**
  – all biomolecules contain carbon (a very versatile atom)
  – necessary for the existence of all known forms of life
  – C can share 4 electrons (can form 4 bonds); N has 3, O has 2 and H has 1 unpaired electron(s)
How are Biomolecules Produced?

- major precursors for the formation of biomolecules are water, carbon dioxide, and 3 inorganic nitrogen compounds (amonium $\text{NH}_4^+$; nitrate $\text{NO}_3^-$; dinitrogen $\text{N}_2$).
- metabolic processes transform these inorganic precursors into biomolecules.
Precursors → Cell

The inorganic precursors:
(18–64 daltons)
Carbon dioxide, Water, Ammonia,
Nitrogen(N₂), Nitrate(NO₃⁻)

Metabolites:
(50–250 daltons)
Pyrurate, Citrate, Succinate,
Glyceraldehyde-3-phosphate,
Fructose-1,6-bisphosphate,
5-Phosphoglyceraldehyde

Building blocks:
(100–350 daltons)
Amino acids, Nucleotides,
Monosaccharides, Fatty acids,
Glycerol

Macromolecules:
(10⁵–10⁶ daltons)
Proteins, Nucleic acids,
Polysaccharides, Lipids

Supramolecular complexes:
(10⁶–10⁸ daltons)
Ribosomes, Cytoskeleton,
Multi-enzyme complexes

Organelles:
Nucleus, Mitochondria,
Chloroplast, Endoplasmic reticulum,
Golgi apparatus, Vacuole

Biochemistry by Reginald H. Garrett and Charles M. Grisham
Weak Forces and Interactions

- **Van der Waals forces**
  - result of induced electrical interactions between closely approaching atoms or molecules as their negatively-charged electron clouds fluctuate instantaneously in time; may be *attractive* – between positively charged nuclei and the electrons of nearby atoms- dipole-dipole interactions, dipole induced dipole interactions- or *repulsive* when two atoms approach too close to one another

- **hydrogen bonds**
  - form between hydrogen atom covalently bonded to an electronegative atom (O, N) and a second electronegative atom that serves as the hydrogen bond acceptor; stronger than Van der Waals forces; highly directional

- **ionic interactions**
  - result of attractive forces between oppositely charged polar functions, such as negative carboxyl groups and positive amino groups

- **hydrophobic interactions**
  - exist due to the strong tendency of water to exclude non-polar groups or molecules; water molecules prefer the stronger interactions that they share with one another, compared with their interaction with non-polar interactions
Weak Forces and Interactions

1. weak forces are non-covalent bonds
2. they are 1-3 orders of magnitude weaker than covalent bonds
3. they are several times greater than dissociating tendency due to thermal motion of molecules (at 25°C)

<table>
<thead>
<tr>
<th>Type of bonding</th>
<th>Strength (kJ/mol)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Van der Waals interactions</td>
<td>0.4-4.0</td>
</tr>
<tr>
<td>Hydrogen bonds</td>
<td>12-30</td>
</tr>
<tr>
<td>Ionic interactions</td>
<td>20</td>
</tr>
<tr>
<td>Hydrophobic interactions</td>
<td>&lt;40</td>
</tr>
<tr>
<td>The average kinetic energy of molecules</td>
<td>2.5</td>
</tr>
<tr>
<td>at 25°C</td>
<td></td>
</tr>
<tr>
<td>Covalent bond (O₂)</td>
<td>402</td>
</tr>
<tr>
<td>Covalent bond (N₂)</td>
<td>946</td>
</tr>
</tbody>
</table>

A commonly-used example of a polar compound is water (H₂O). The electrons of water's hydrogen atoms are strongly attracted to the oxygen atom, and are actually closer to oxygen's nucleus than to the hydrogen nuclei; thus, water has a relatively strong negative charge in the middle (red shade), and a positive charge at the ends (blue shade). (source: Wikipedia)
Why are Weak Forces Important?

1. They are *reversible*! Rigid molecules would not facilitate cellular activities.

2. Biomolecular recognition is performed via interplay of complementary molecules. So, **biological function** is achieved through mechanisms based on structural complementarity and weak chemical interactions.

3. If a sufficient number of weak bonds can be formed (as in macromolecules complementary in structure to one another) larger structures assemble spontaneously.

4. A consequence: weak forces restrict organisms to a narrow range of environmental conditions (temperature, ionic strength, relative acidity).

5. The loss of structural order is called *denaturation*. It is accompanied by loss of function.
Protein

- biomolecule, macromolecule
  - more than 50% of the dry weight of cells → proteins
- polymer of amino acids connected into linear chains
- strings of symbols
- machinery of life
  - play central role in the structure and function of cells
  - regulate and execute many biological functions

a) amino acid       b) peptide bond

Introduction to Protein Structure by Branden and Tooze
Zwitterion

\[
\begin{align*}
\text{H}_3\text{N}^+ & \quad \text{OH}^{-} \quad \text{pH} \ 2.0 \\
\text{R}^- & \quad \text{H}_2\text{O} \\
\text{R} & \quad \text{OH}^{-} \quad \text{pH} \ 6.0 \\
\text{H}_3\text{N}^- & \quad \text{O}^+ \\
\text{R} & \quad \text{OH}^{-} \\
\end{align*}
\]
Peptide Bonds

- they are planar, nearly
- they are very strong

a) Pauling’s theoretical model (1951)
b) experimentally determined bond lengths
Amino Acids

(a) Hydrophobic amino acids

A Ala, Alanine
V Val, Valine
F Phe, Phenylalanine
P Pro, Proline
I Ile, Isoleucine
L Leu, Leucine
M Met, Methionine
Amino Acids

D  Asp, Aspartic acid
E  Glu, Glutamic acid
K  Lys, Lysine
R  Arg, Arginine
Amino Acids

- Ser, Serine
- Thr, Threonine
- Tyr, Tyrosine
- His, Histidine
- Cys, Cysteine
- Asn, Asparagine
- Gln, Glutamine
- Trp, Tryptophan
Amino Acids

Aliphatic side chains: A, V, L, I
Hydroxyl-containing residues: S, T
Acidic residues: D, E
Amide-containing residues: N, Q
Sulfur-containing residues: C, M
Basic residues: K, R
Aromatic residues: F, Y, W
Other: P, H, G
Ramachandran Plot

1. **α region** – corresponds to the residues found typically in the alpha helices

2. **β region** – corresponds to the residues found typically in the beta sheets

3. **L region** – corresponds to the residues typically found in the left-handed helices

- most angles are not allowed because of steric collisions
  - between atoms of the same residue
  - between atoms of the neighboring residues
- the allowed combinations can be calculated
Ramachandran Plot

a) observed Ramachandran angles for all residues except glycine
b) observed Ramachandran angles for glycine

Introduction to Protein Structure by Branden and Tooze
Ramachandran Angles

- Almost completely describe overall fold of a protein

- Why?
  - Bond *lengths* are approximately fixed
  - Bond *angles* are approximately fixed
  - Only Ramachandran (torsion, dihedral) angles are variable
  - Only 2 variables per amino acid (for the backbone)
Peptide Bond Revisited

• Usually *trans*
  – with $\omega \approx 180^\circ \pm 6^\circ$ rms°

• Occasionally *cis*
  – with $\omega \approx 0^\circ$:
  – ~ 1/4 of prolines
  – very infrequently glycines
  – almost never other amino acids
Levels of Protein Structure

Native state (conformation) – conformation at which protein shows its activity
Helical Structure

a) idealized diagram
b) the same as a) but with approximate positions for main-chain atoms
c) schematic diagram of an alpha helix
d) a ball and stick model
Helical Structure
Other Helical Conformations

- **3\textsubscript{10} helix**
  - found in proteins when a regular helix is distorted by the presence of unfavorable residues, near turn regions or in short helices
  - hydrogen bonds between \(i\) and \(i+3\) (instead of \(i+4\))
  - 3 residues per turn and 10 backbone atoms between donor and acceptor atom
  - tighter and narrower
- **\(\pi\) helix**
  - more loosely coiled
  - hydrogen bonds between \(i\) and \(i+5\) (instead of \(i+4\))
  - 4.4 residues per turn
  - can be very long
- **Poly(Pro) helices**
  - all \(cis\) with 3.3 residues per turn, right-handed (type I)
  - all \(trans\) with 3 residues per turn, left-handed (type II)
- **Poly(Gly) chains**
  - type I: beta conformation
  - type II: similar to poly(Pro) helix with 3 residues per turn
Polyproline II Helices (PPII)

- shorter than regular helices (4-5 residues)
- longer (physically) than regular helices (rise per turn is twice that of regular helices – 9.3Å)
- seem to be stabilized by main chain – water hydrogen bonds
- found mostly on protein surface
- preference for hydrophilic residues and proline
- Gln, Ser, Arg and Ala are found in PPII regions, Gly is rare
- involved in protein-protein interactions
- important roles in signal transduction, transcription, cell motility, etc.
- contain sequence motifs, e.g. PXXP
- dominant element of secondary structure in unfolded proteins (Horng and Raines, Prot. Sci. 2006)
Anti-parallel $\beta$-sheet
Anti-parallel $\beta$-sheet

Typical length: 5-10 residues
Parallel $\beta$-sheet
Mixed $\beta$-sheet

Only about 20% of sheets are of mixed type.

Almost all sheets have twisted strands with fixed handedness (right-handed twist)
(Hairpin) Loop

a) histogram showing the frequency of hairpin loops of different lengths in 62 proteins

b) the two most frequently occurring two-residue hairpin loops
Super Secondary Structure

Helix-loop-helix motif

- helix-loop-helix motif with Ca^{2+} atom attached
Domains

Four Helix Bundle

- hydrophobic residues tend to be on the inside
- polar residues tend to be on the outside of proteins

a) four helix bundle – red cylinders are helices while green parts are loops
b) projection from above
Disulphide Bonds (Bridges)

- Covalent bond between two cysteins (i.e. their sulfur atoms)
- Require oxidative environment
- Present in extracellular proteins
- Stabilize proteins
- "Create" so-called long-range interactions
Levinthal’s Calculation

- Cyrus Levinthal 1968

- Q: do proteins explore all possible conformations before they adopt a specific 3-D structure?

- A: let’s consider a simplified problem
  - each residue can adopt one of the three discrete groups from the Ramachandran plot (alpha, beta, loop)
  - a switch between conformations can be done in $10^{-12}$ seconds
  - then, a protein with 150 residues would need to explore $3^{150}$ possible states, which is $10^{71}$
  - at the rate of $10^{-12}$ a protein would need $\sim 10^{50}$ years

- we know that protein folds between 0.1s and 1000s
Protein Folding Problem

- How do proteins fold into a specific 3-D structure?
- How does the primary structure of a protein determine its secondary and tertiary structure?

-----

- there are two conditions a protein needs to meet
  - there must be a single, stable, folded conformation (thermodynamic condition)
  - a protein must fold on an appropriate time scale (kinetic condition)

- thus, only a small amount of conformational space is explored
- also, there must exist a specific folding pathway
- the paradox how proteins quickly fold into specific 3-D conformations is called a protein folding problem