Protein Function and Folding

I690/B680: Structural Bioinformatics

January 18, 2006
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.
Folding and Flexibility

- the process by which a polypeptide chain acquires its correct 3D structure to achieve biologically active native state is called **protein folding**
- many protein chains spontaneously fold into the native state, others require the assistance of enzymes or other proteins called chaperones

- a protein in its native state is not static
- secondary structural elements of the domains as well as the entire domains continually undergo small movements in space
- either fluctuations of individual atoms or collective motions of groups of atoms
- functional activities of many proteins depend upon large conformational changes triggered by ligand binding
Globular Proteins are only Marginally Stable

- slight changes in pH or temperature can convert a solution of biologically active proteins in their native state to a biologically inactive denatured state
Folded vs. Denatured State

- there are two major contributors to the energy difference between the folded and the denatured state
  - enthalpy
  - entropy

- **Enthalpy**
  - derives from the energy of the non-covalent interactions within the polypeptide chain (H-bonds, ionic bonds, hydrophobic interactions)
  - the covalent bonds within and between the amino acid residues are the same in the native and denatured states, with the exceptions of disulphide bonds

- **Entropy**
  - derives from the second law of thermodynamics which states that energy is required to create order
  - in the absence of other forces, it would be energetically favorable for a protein to remain in the disordered denatured state
Proteins are Marginally Stable

- the total energy difference between the native and the denatured state is 5-15 kcal/mol, which is called the free energy difference.
- free energy difference is small, but the problem is that this is the difference between two very large numbers (enthalpy difference and entropy difference).
  - this is a severe problem in predicting possible native state using molecular dynamics.
- the marginal stability of the native state over the denatured state is biologically important.
- living cells need globular proteins in correct quantities at appropriate times.
- it is important to degrade them quickly as it is important to synthesize them quickly.
Kinetic Factors

- High resolution x-ray structures of several hundred proteins have shown that in each case the specific sequence of a polypeptide chain appears to yield only a single, compact, biologically active fold in the native state.
- NMR experiments show that the same fold prevails in solution too.
- Proteins cannot search all possible conformations (Levinthal’s “paradox”).
- Thus, to occur on the short time scale, the folding process must be directed in some way through a kinetic pathway of unstable intermediates to escape sampling a large number of irrelevant conformations.
Kinetic Factors

- folding mechanism is difficult to examine experimentally since possible intermediates have short lifetime
- if kinetic factors are important for the folding process it is possible that the observed folded conformation is not the one with the lowest free energy, but rather the most stable of those conformations that are kinetically accessible
  - protein might be kinetically trapped in a local low energy state with high energy barrier that prevents it from reaching the global energy minimum
  - global energy minimum state may have a different fold
  - how can this affect structure prediction based on molecular simulations?
- how a living cell can prevent the folding pathway from becoming blocked at an intermediate stage?
  - aggregation of the intermediates through exposed hydrophobic groups
  - formation of incorrect disulphide bonds
  - isomerization of prolines
Folding Intermediates

- molten globule state
  - first observable state in the folding pathway
  - collapse of the flexible disordered state into partially organized folded state
Molten Globule and Folded State

First Step
- occurs in a few milliseconds and is hard to observe experimentally
- has most of the secondary structure of the native state
- in some cases, has native-like positions of helices and strands
- less compact than the native structure and the proper packing interactions in the interior of the protein have not been formed
- should be seen as an ensemble of structures

Second Step
- can last up to a second or more
- persistent native-like elements of secondary structure begin to develop
- forming of subdomains
- still not in a single form (proper hydrophobic packing is not present and surface loops are not fixed)
Folding Process

- unfolded state, U
  - ensemble of conformationally different molecules

- molten globule, M
  - ensemble of structurally related molecules which are rapidly interconverting and which slowly change into a single conformation

- the folded state, F
  - a molecule must go through the high energy transition state T
Burying Hydrophobic Side Chains

- key event and a main mystery of protein folding

**Secondary structure formation** cannot be the driving force of folding
- there is very little change in free energy by forming the internal H-bonds characteristic for helices and sheets
- in the unfolded state, equally stable H-bonds can be formed with water

**Hydrophobic effect**
- there is a large free energy change by bringing hydrophobic side chains out of contact with water and into the contact with each other
- vastly reduces the number of conformations to be searched
- buried residues will have to make H-bonds in secondary structure elements
- secondary structure formation is consequence of hydrophobic effect
Hierarchical Building Block Folding Model

- there is a major (not necessarily unique) folding pathway that most proteins follow
- local neighborhoods interact and create folding hydrophobic units
- then, domains and entire proteins are created
- however, not all local neighborhoods show propensities towards one preferred conformations
Folding Pathways

- both single and multiple folding pathways have been observed
- folding of the lysozyme involves parallel pathways and distinct folding domains
Folding Pathways

Lysozyme has multiple folding pathways.

Barnase has a single folding pathway.
Folding Funnel

E represents the energy of the system,
Q is defined as the proportion of native contacts formed,
P is a measure of the available conformational space

Three pathways are shown corresponding to (yellow) fast folding, (green) slow folding pathway that crosses the high energy barrier, and (red) slow folding pathway which returns to a less folded state before following the pathway for fast folding.

More Folding Funnels
Forming Disulphide Bridges

- In eukaryotic cells, disulphide bond formation occurs in the endoplasmic reticulum before proteins are exported to the cell surface.
- Enzyme PDI catalyzes disulphide exchange to remove intermediates with incorrectly formed disulphide bridges.
- Proteins with disulphide bonds are not found in cytosol, but are located in the plasma membrane or are secreted.
Proline Isomerization

- *cis-trans* isomerization of proline peptides is intrinsically slow process

- *in vitro* it is a rate limiting step in folding for those molecules that have been trapped in the folding intermediate with the wrong isomer

- peptidil prolyl isomerase (cyclophilin) catalyzes the process *in vivo* (both in prokaryotes and eukaryotes)
Molecular Chaperones

• before they attain native conformation proteins may expose their hydrophobic patches to the solvent
• isolated purified proteins can thus aggregate *in vitro* even at low protein concentrations
• inside cells, at much higher concentrations of many proteins, aggregation can easily occur
• this is prevented by molecular chaperones
  – ubiquitous and abundant families of proteins that assist the folding of both nascent polypeptides still attached to ribosomes and released complete polypeptide chains

• some chaperones bind together into chaperonins and then bind unfolded and incorrectly folded proteins, but not native proteins
There is sufficient information contained in the protein sequence to guarantee correct folding from any of a large number of unfolded states.
Thermodynamic Hypothesis

- native conformation of a protein is adopted spontaneously i.e.

  amino acid sequence $\rightarrow$ 3-D structure

Anfinsen’s demonstration of this fundamental property of proteins opened the problem to a massive amount of experimental and theoretical effort.

His summary of the experiments was presented as a Nobel Prize Lecture and published in:

Fischer’s Experiment

Hermann Emil Fischer – 1894

An enzyme and a substrate have to fit each other like a lock and key in order to exert chemical effect on each other

lock-and-key theory

later, lock-and-key paradigm was expanded to contain so-called induced fit theory

“\textit{The examination of the synthetic glucosides has shown that the action of the enzymes depends to a large extent on the geometrical structure of the molecule to be attacked, that the two must match like lock and key.}” H. E. Fischer in his Nobel Lecture
Sequence-Structure-Function Paradigm

Standard protein structure/function paradigm (Fischer, 1894, Anfinsen 1973)

Amino Acid Sequence

EKKIRVAINGFGRIGRNFLRCHQGRQNTLDDLVIINDSGGVRKQASHLLKYDSTLGFAAD

> 1NLG:_ NADP-LINKED GLYCERALDEHYDE-3-PHOSPHATE

VKIVDSDHSVGDGQIKIVSSRDPLQWPWKMIDLVIEGTGVFIDKVGAGKHIQAGASK
VLITAPAKDIPTFVVGNEQGYKHEYPISNASCTTNCLAPFVKVEQKFGIVKMT
TTHSYTGQKRTLASHRDLRRARAALNIVPTTTGAATAKAVSLPQSKGLNLGIALRVTPT
PTVSVDLVQKVEKTKFAEEVNAAFREAANGPMKVLHVEDAPLSIFDKCTDQSTSID
SLTMVGDDMVKVAWYDNEWQYSGRVVLAEVTAKKWVA

3-D Structure

Classification: Gene Transfer

EC Number: 1.2.1.13

Protein Function

Dominant view: 3-D structure is **prerequisite** for protein function
Calcineurin-Calmodulin Counter Example

**Calcineurin:**
- calcium-dependent phosphatase
- regulated by calmodulin (calcium-binding protein)
- induces conformational change of calmodulin upon binding
- may be involved in human heart failure when calcium concentration is chronically increased
- “disorder” is important for the binding mechanism

**Intrinsically disordered proteins (natively unfolded/unstructured proteins):**
- do not have stable 3D conformation under physiological conditions
- abundant in nature
Can Proteins Misfold?

- the lack of function is not always the worst-case scenario
- misfolding can lead to diseases
Open Problems in Protein Bioinformatics

The Ten Most Wanted Solutions in Protein Bioinformatics
by Anna Tramontano
<table>
<thead>
<tr>
<th>1. Protein Sequence Alignment</th>
<th>6. Functional Site Identification</th>
</tr>
</thead>
<tbody>
<tr>
<td>2. Predicting Protein Features from Sequence</td>
<td>7. Protein-Protein Interactions</td>
</tr>
<tr>
<td>3. Function Prediction</td>
<td>8. Protein-Small Molecule Interactions</td>
</tr>
<tr>
<td>4. Structure Prediction</td>
<td>9. Protein Design</td>
</tr>
<tr>
<td>5. Membrane Proteins</td>
<td>10. Protein Engineering</td>
</tr>
</tbody>
</table>
Problem 1. Protein Sequence Alignment

- amino acid sequences have been evolutionarily selected for their favorable thermodynamic, kinetic and functional properties

- when variations that do not impair essential function occur in replicating (germinal) cells they are transmitted to the progeny and generate diversity in the population
- variations that do impair essential functions disappear

- if the variation confers selective advantage it can become the most frequent variant in the population

- if the function performed by a protein has to be conserved and is brought about by specific residues and their relative position in the 3-D structure, then residues responsible for function and structure must be conserved!!!
Evolution-Based Inference of Protein Function

• if we can identify an evolutionary relationship between two proteins between species and find conserved residues, these residues are candidates for involvement in functional mechanisms

• two groups of conserved amino acids
  – those that are conserved because of their structural role
  – those that are conserved because of their functional role

• similar amino acids can more easily replace each other in a structural role; example: catalysis requires specific atoms

• Homology detection (or protein sequence alignment) problem
  – given proteins $p_1$ and $p_2$, what is the probability they are homologous
  – given homologous proteins $p_1$ and $p_2$ identify all pairs of amino acids that derive from the same amino acid of the common ancestor
Orthology vs. Paralogy

(a) Speciation and subsequent divergence

(b) Speciation and subsequent divergence

Gene duplication

Selective pressure to maintain function $F_a'$

Selective pressure to maintain function $F_a'$
Detecting Remote Homology

- duplication and subsequent divergence
- mixing and matching of domains
- detecting very distant homologous relationships is important
  - enlarges the number of proteins for which some functional inference can be made
  - makes easier detection of functional residues
  - detection of distant relationships may shed new light to the process of evolution between organisms
Achieved vs. Non-Achieved

• pairwise sequence alignment is a solved problem
  – Needleman-Wunsch algorithm for global alignment
  – Smith-Waterman algorithm for local alignment
  – BLAST and FASTA heuristics

• multiple sequence alignment is NOT a solved problem
  – dynamic programming – unacceptable
  – progressive alignment: Feng-Doolittle and ClustalW algorithms

• what is a good scoring system?
• sequence profiles and hidden Markov models

• database searching (BLAST and FASTA, again)

• how can structure be incorporated into sequence alignment?
Problem 2.
Predicting Protein Features from Sequence

• features to be predicted: secondary structure elements, post-translational modification sites, cellular compartments, functional sites

• task: given a training set composed of proteins that share a given property, infer the rules important for function

• How can function be deduced?
  – by the presence of a particular sequence pattern (deterministic)
  – by estimating probability that the given sequence belongs to the set of positive examples (stochastic)

• if only positive set is used ⇒ conservation problem
• if both positive and negative sets are used ⇒ classification problem
Deterministic Patterns

Example #1:
• NS3 protease in hepatitis C virus
• contains serine, histidine and aspartic acid at key positions
• Pattern over many similar proteins: [DE] S G [GS]

Example #2:
• regions with no definite constraints can be included
• D X(1, 4) [LI] X [DE]
  – aspartic acid; 1-4 unconstrained residues; leucine or isoleucine; unconstrained residue; aspartic or glutamic acid

Example #3:
• NS3 protease: 1d xp
• GLGNGLGRLA
Stochastic Patterns

Calmodulin Binding Motif (IQ Motif)

Tyrosine Phosphorylation Sites
Predicting Domain Boundaries

- there is no consensus on a definition of a domain
  - say, a globular, compact regions of a protein structure with relatively more contacts within themselves than with the rest of the structure
- precise domain boundaries are difficult to define even when the structure is present ⇒ manual inspection is required
- thus, hard to obtain clean set of examples for informatics methods

The structure of the elongation factor-1 from *Sulfolobus solfataricus*, a protein involved in RNA translation. Three domains are connected by long amino acid stretches.
Predicting Domain Boundaries

- domains are not necessarily contiguous

- Some ideas:
  - SnapDRAGON: produces several hundred putative 3D models and detects domains by averaging prediction results
  - DomSSEA: predicts secondary structure of the target protein and maps predicted sequence of helices and sheets on the known domains

A discontinuous domain on the RNA 3’-terminal phosphate cyclase from yeast.