2. Genomes: From Sequence to Structure

a. A fundamental paradigm of protein science: amino acid sequence encodes the three-dimensional structure, which in turn specifies the function

What does a protein do with its one-dimensional information? A fundamental paradigm of modern science is that the one-dimensional amino acid sequence encodes the protein’s three-dimensional structure, whether stretched out, or balled up into some particular structure. Physically, this feat is achieved by means of the many interactions specified in its sequence, which guide the chain to fold – in many cases automatically – to its functional native state. The mapping between sequence and its native structure is often referred to as the folding code, or the second genetic code, in analogy to the (first) genetic code (Table I.1.2). As described above, the genetic code is a dictionary for translating the four-letter alphabet of nucleotides in DNA and RNA into the twenty-letter code of proteins. The folding code is different. The folding code is not a simple conversion of one-dimensional information in one alphabet into one-dimensional information in another alphabet, like the genetic code. Rather, it is an expression of the physical forces that cause a given one-dimensional sequence of amino acids to coil up or collapse into a particular three-dimensional structure.

According to the fundamental paradigm of protein science, the amino acid sequence specifies the three-dimensional structure, which in turn determines how the protein functions:

SEQUENCE $\rightarrow$ STRUCTURE $\rightarrow$ FUNCTION

Thus, understanding how the three-dimensional structure is encoded in the amino acid sequence is an important step towards understanding the encoding of biological function in biomolecules and comprehension of complex organisms from their genome information.

b. There has recently been an exponential growth in sequence, structure and expression level information from experiments

In the last few decades, there has been a spectacular explosion in the number of experimentally determined sequences and structures of biomolecules. Progress in this field is proceeding at breathtaking speed. Compare the time (~ years) it took to crystallize and determine the X-ray structure of the first protein (myoglobin) by Kendrew about four decades ago, and the fact that currently 2-3 new structures are determined on a daily basis, - despite the fact that structure characterization is generally impeded by the details of crystallization, data collection and analysis. Sequence characterization, or genome sequencing,
is significantly faster. Figure 1.1. shows the number of nucleotide sequences, protein sequences, and protein structures accumulated over the last 25 years \cite{Rost1998}. An important aspect of these studies is that the data have been systematically collected in databases. The Protein Data Bank (PDB) is the database of known structures of proteins and their complexes \cite{Berman2000}. Atomic coordinates down to nanometer resolution are available in the PDB for > 21,000 structures. The nucleotide sequence database (EMBL) \cite{Stoesser1997} is about one order of magnitude larger than the amino acid sequence database (SWISS-PROT) \cite{Bairoch1998}. Knowledge of nucleotide sequence does not necessarily provide information on amino acid sequence, in spite of the well-established genetic code (Table 1.1.2). The discrepancy is due to the fact that the protein-encoding sequences, also called exons, comprise only a few per cent of the genome and an average of about 5% of each gene \cite{Lander2001}. Only 1.1% of the human genome is spanned by exons, 24% by introns (nucleotide sequences between the exons forming a given gene), the remaining 75% is intergenic \cite{Venter2001}.

On the other hand, not all of the known amino acid sequences can be readily expressed and crystallized, and hence another decrease of about one order of magnitude in the number of structures deposited in the PDB. Two key developments in X-ray crystallographic structure determination have been the high-flux synchrotron source at Argonne National Laboratories, and the anomalous diffraction method of analysis or MAD phasing of selenomethionyl proteins \cite{Hendrickson1991}. These, combined together, have greatly sped up crystal structure production. The output from X-ray experiments is now > 2000 new structures per year (Figure 1.2.2), although the fraction of ‘new folds’ is decreasing (Figure 1.2.3). Currently, the number of distinct folds (3-dimensional structures at a coarse-grained scale, also referred to as architecture)
amounts to less than 800 (http://scop.mrc-lmb.cam.ac.uk/scop/count.html). In fact, sequence tends to diverge more rapidly than structure, and protein folds can accommodate surprisingly large sequence alterations by structural rearrangements on a local scale [Chothia & Gerstein 1997 ID: 308] (see Figure I.2.4). Many proteins encoded by newly sequenced genes are found to fold into known architectures, when their 3-D structures are subsequently determined. It is anticipated that overall 1500 folds may exist [Gaasterland 1998 ID: 326][Kim 1998 ID: 325][Pennisi 1998 ID: 324] although this number has been widely debated [Rost 1998 ID: 331][Zhang C.O. & Delisi 1998 ID: 1150].
Figure I.2.4. Fraction of residues superimposable in the core of the structures as a function of sequence identity percentage. The majority of core residues are structurally superimposable, for example, for two proteins sharing > 30% sequence identity. The figure is redrawn ([Durbin, Eddy, et al. 1998 ID: 305], Figure 6.2) from data reported by Chothia and Lesk [Chothia & Lesk 1986 ID: 310]. ‘Other’ refers to two dehydrofolate reductases, two lysozymes, and the pairs plastocyanin/azurin and papain/actinidin.

Scop Classification Statistics

SCOP: Structural Classification of Proteins. 1.63 release
18946 PDB Entries (1 March 2003). 49497 Domains. 28 Literature References (excluding nucleic acids and theoretical models)

<table>
<thead>
<tr>
<th>Class</th>
<th>Number of folds</th>
<th>Number of superfamilies</th>
<th>Number of families</th>
</tr>
</thead>
<tbody>
<tr>
<td>All alpha proteins</td>
<td>171</td>
<td>286</td>
<td>457</td>
</tr>
<tr>
<td>All beta proteins</td>
<td>119</td>
<td>234</td>
<td>418</td>
</tr>
<tr>
<td>Alpha and beta proteins (a/b)</td>
<td>117</td>
<td>192</td>
<td>501</td>
</tr>
<tr>
<td>Alpha and beta proteins (a+b)</td>
<td>224</td>
<td>330</td>
<td>532</td>
</tr>
<tr>
<td>Multi-domain proteins</td>
<td>39</td>
<td>39</td>
<td>50</td>
</tr>
<tr>
<td>Membrane and cell surface proteins</td>
<td>34</td>
<td>64</td>
<td>71</td>
</tr>
<tr>
<td>Small proteins</td>
<td>61</td>
<td>87</td>
<td>135</td>
</tr>
<tr>
<td>Total</td>
<td>765</td>
<td>1232</td>
<td>2164</td>
</tr>
</tbody>
</table>
Protein dynamics

- Folding/unfolding dynamics
  - Passage over one or more energy barriers
  - Transitions between infinitely many conformations
- Fluctuations near the folded state
  - Local conformational changes
  - Fluctuations near a global minimum

Folding/unfolding energy landscapes

- Thermodynamic equilibrium $\rightarrow$ native state has lowest free energy

Knowledge-based studies

Exploiting PDB structures
Protein folding problem:
“Predicting 3-dimensional structure from amino acid sequence”

- A unique folded structure (native conformation, native fold) is assumed by a given sequence, although infinitely many conformations can be accessed.
- Which? (Protein folding problem)
- How, why? (Folding kinetics)

Basic postulate: Thermodynamic equilibrium
⇒ Global energy minimum

CASP (Critical Assessment of Structure Prediction)

Homology/comparative modeling

Three computational methods:
- Homology modeling
- Threading
- Ab initio simulations
SWISS-MODEL

An Automated Comparative Protein Modelling Server accessible via the ExPASy web server (by Peitsch et al.)

STEPS:
1. Search for suitable templates (from ExNRL-3D, using BLAST)
2. Check sequence identity with target
3. Generate ProModII jobs
4. Generate models (ProModII) using known 3-d templates
5. Energy minimization with Gromos96

http://swissmodel.expasy.org/SWISS-MODEL.html

Predict Protein (Rost)

http://www.embl-heidelberg.de/predictprotein/

Structural Homology

- Dali Server (Sander-Holm)
  http://www2.ebi.ac.uk/dali/


Threading (Fold recognition)

Loopp (Elber)
Threader (Jones)
Ab initio simulations

- Protarch (Scheraga’s group)
- Rosetta (Baker’s lab)
- Touchstone (Skolnick)

Need for Low Resolution Approaches

Coarse-grained Models

with

Empirical Force Fields

are the most tractable - if not the only possible – computational tools for investigating large systems, and complex biological processes

Virtual bond model

1. Single interaction site per residue, identified by the α- or β-carbon
2. Need for empirical potentials for inter-residue interactions
Topological and Energetic Factors: What determines the transition state ensemble, and folding intermediates?

Can we use such simplified approaches for estimating amyloidogenic intermediates?

Gaussian Network Model
A single parameter potential is sufficient to reproduce the slow dynamics in good detail.

Detailed specific potentials

\[ V_{\text{det}} = (\gamma/2) \Delta R^T \Gamma \Delta R \]

Approximate uniform potential

\[ V_{\text{tot}} = \frac{\gamma}{2} \left( \sum_{k=1}^{N} \left( \Delta R_{1k} \right)^2 + \sum_{k=1}^{N-1} \left( \Delta R_{k,k+1} \right)^2 \right) = \frac{\gamma}{2} \left( \sum_{k=1}^{N} \left( \Delta R_{1k} \right)^2 + \sum_{k=1}^{N-1} \left( \Delta R_{k,k+1} \right)^2 \right) \]

Kirchhoff matrix of contacts

\[ \Gamma = \begin{cases} 1 & \text{if } \Gamma_{ik} < \text{cut} \\ 0 & \text{if } \Gamma_{ik} > \text{cut} \\ -\sum_k \Gamma_{ik} & \end{cases} \]

Comparison with X-ray Temperature Factor

Debye-Waller factors:

\[ B_k = 8 \pi^2 \langle \Delta R_k \rangle^2 \]

Comparison with H/D Exchange – NMR data

\[ \Delta S_i = k \ln W(\Delta R_i) = -\gamma (\Delta R_i)^2 / (2T G_{ii}^{-1}) \]

V. Folding Kinetics and Folding Intermediates

2.3.1. Mechanism of folding

2.3.1.1. All-or-none or sequential transition? Single or multiple pathways?

A crucial question is whether folding proceeds by an all-or-none process from the unfolded state to the native form, or follows some definite pathway involving formation of intermediates. The former process holds for the heat-induced denaturation of small, single domain proteins, - although the same proteins can exhibit multiphase folding kinetics under suitably adjusted (pH, temperature) conditions. A classical example is the folding of chymotrypsin inhibitor 2 (CI2). However, in the case of multiple domain proteins, or proteins consisting of structural elements capable of undergoing independent folding kinetics - also referred to as autonomous folding units - the formation of local structures and their gradual consolidation into larger structural elements seems likely. In this case, the passage between the unfolded (U), intermediate (I₁, I₂, etc.) and native (N) states can involve an off-pathway

\[
\text{U} \xleftrightarrow{\text{I}} \text{N} \quad \text{(V.I)}
\]

or on-pathway intermediate (I),

\[
\text{U} \xleftrightarrow{\text{I}} \xleftrightarrow{\text{N}} \quad \text{(V.II)}
\]

or multiple intermediates

\[
\text{U} \xleftrightarrow{\text{I}_1 \ldots} \text{I}_2 \xleftrightarrow{\ldots} \text{N} \quad \text{(V.III)}
\]

in a sequential transition scheme [Baldwin, 1996 #199]. We note that intermediate accumulation may be non-productive (scheme V.I), or act as a kinetic trap (V.II or V.III). Empirical methods are usually employed, i.e. parameters best fitting the experimental data are determined, for assigning rate constants to these steps. This type of
phenomenological approach in which the unfolded, intermediate and native structures are each considered as *macrostates* devoid of statistical properties, is the *classical chemical kinetics approach* [Baldwin, 1996 #1361; Dill, 1995 #176].

Inasmuch as the protein in the unfolded state is actually an *ensemble* of conformations, it is natural to conceive the occurrence of *multiple pathways* of folding starting from different conformations. Some of the conformations visited during the passage from unfolded to folded state are energetically more favorable and therefore appear more frequently; or some local structural arrangements, such as native-like α-helices or β-strands, could be more persistent once formed. This may reconcile the occurrence of multiple folding pathways with the observation of some well-defined structures during the course of folding.

The description of folding mechanism, as a parallel flow process from an ensemble of chain molecules, is the *New View* of protein folding kinetics [Baldwin, 1994 #221; Baldwin, 1995 #222; Dill, 1997 #233]. It relies on statistical mechanics and energy landscape concepts strongly based upon polymer approaches [Flory, 1969 #13; Chan, 1991 #232]. It was put forward first by theorists, based on statistical mechanical theories and Monte Carlo simulations of folding [Bryngelson, 1987 #234; Sali, 1994 #148; Wolynes, 1995 #198; Dill, 1997 #233].

**2.3.1.2. An overview of proposed mechanisms**

Our view of folding mechanism has undergone a considerable change, since the original argument of Levinthal [Levinthal, 1968 #193] against the possibility of finding the native fold via *random search* - amongst infinitely many conformations accessible in the unfolded state, within the experimentally observed short folding times. This astronomically large combinatorial problem has been referred to as 'finding a needle in a haystack' [Dill, 1993 #20]. Levinthal suggested that intermediate structures form along well-defined pathways that are *kinetically driven*, i.e. there would be an evolution only towards those states that can be reached by passing over *lower energy barriers*. This picture is not compatible with a *thermodynamic equilibrium* between folded and unfolded states in which (i) the depth of energy wells matters, rather than the height of the energy barriers, (ii) folding is path-independent;
everything is controlled by free energy differences which are state functions, and therefore depends on the initial and final states, only. In conformity with these arguments, a hierarchical folding mechanism consisting of three was suggested by Ptitsyn [Ptitsyn, 1973 #192]: formation of secondary structures, merging of these into a compact state with a native-like folding pattern (molten globule), and finally transition of this intermediate into the rigid native structure. The framework model defined by Kim and Baldwin [Kim, 1982 #21] is based on the same assumption that intermediate structures form, first on a local scale, than at a more global scale, along the folding pathway. A schematic view of molten globule state is given in Figure 2.3.1.

![Figure 2.3.1. Schematic view of the transition between native (N) and molten globule (MG) states of a protein. The overall architecture of the MG state bears close resemblance to that of the state N. Most of the secondary structural elements are conserved, but these are usually less tightly packed. The MG state thus enjoys an increased conformational flexibility compared to the native state. Sidechain interactions are loose and relatively non-specific compared to those in native structures. (Adapted from [Ptitsyn, 1992 #10]; see also Figure 6 [Ptitsyn, 1995 #12]).](image)

It is clear that not all proteins fold through this mechanism; some undergo a hydrophobic collapse succeeded by optimization of intramolecular interactions (most likely via hydrogen bonded secondary structures). A 'hydrophobic zipping' mechanism has also been proposed in which hydrophobic collapse is concurrent with the growth of helices and sheets [Dill, 1993 #215]. See Figure 2.3.1. In this model, first a pair of hydrophobic (H) residues that are sufficiently close along the sequence make a contact. These residues are designated as a and a’ in Figure 2.3.1. The entropy loss accompanying the association of the H residues is assumed to be more than counterbalanced by the favorable enthalpic effect, such that the overall Gibbs free
energy decreases upon H-H (or a – a’) contact. This contact can bring other H residues into close proximity (b and b’, in the figure), then they too form a contact. The new contact is even more favorable than the first, because of the smaller entropy loss incurred at this second step. This is simply due to the fact that the conformational space has already been reduced by the first contact, and the additional entropy loss is relatively small. This process, hydrophobic zipping, continues with minimal loss of configurational entropy at each step.

**Figure 2.3.2.** Schematic view of the hydrophobic zipper model for protein folding. The hydrophobic (H) residues along the protein chain are indicated by filled circles. H-H contacts drive the zipping of helices and sheets, provided that the chain has an appropriate sequence of H residues. (Reproduced from [Dill, 1993 #215], Figure 1)

Another mechanism proposed for the two-state cooperative folding of small proteins is the *nucleation-condensation mechanism* [Abkevich, 1994 #225; Fersht, 1995 #229; Fersht, 1997 #223]. According to this mechanism, the nucleus comprises a set of residues, usually on an α-helix, stabilized by tertiary contacts with distant portions of the chain. An optimization analysis indicates that the rates are highest if the denatured state has little residual structure, and if there is no accumulation of intermediates.

Implicit in all these models is a description of protein folding in terms of an evolution of structure. Yet, theoretical and computational studies indicate that the
progress between two equilibrium states in the conformational space is a statistical process: it cannot be described by a unique pathway or a single transition state. The trajectories observed in molecular dynamics or Monte Carlo proceed by diverse pathways, depending on the original conformation; and the most probable structure formation/breaking mechanisms on a local scale, or recurring dynamic patterns typifying intermediates, are detected through a statistical examination of multiple trajectories. The folding scheme delineated by Miranker and Dobson (1996) (Figure 2.2.2) is perhaps the most realistic picture in this respect.

2.3.1.3. New View of protein folding kinetics. Energy landscapes

Recent progresses in time-resolved instrumentation and experimental design now permit to detect folding events at the atomic level in the submillisecond regime down to nanosecond time scale [[Jones, 1993 #169; Pascher, 1996 #235; Muñoz, 1997 #230] [Nolting, 1995 #183; Miranker, 1996 #177]. Even the transition states can now be characterized to a certain extent using protein engineering methods based on site-directed mutagenesis [Matouschek, 1990 #146; Fersht, 1994 #194; Otzen, 1994 #227; Itzhaki, 1995 #228], or pressure-jump techniques in conjunction with fluorescence spectroscopy [Vidugiris, 1995 #224].

The accumulated new data reveals the complexity and multiplicity of folding kinetics. For example, for the same proteins whose heat-induced denaturation has been observed by calorimetry to be a two-state process - such as cytochrome c, hen lysozyme, RNase A - passage through intermediates is detectable provided that the denaturation conditions are properly controlled (see § 2.3.3). Acid-denaturation, for example, favors some intermediate structure formation. Furthermore, different levels of structure can be formed during the course of protein, which are probed with different experimental techniques. Some local structures that cannot be detected by CD or fluorescence spectroscopy are distinguished by hydrogen exchange coupled with 2D-NMR, for example. There is a lack of synchrony in the acquisition of protection against hydrogen exchange during folding. All these observations indicate the statistical nature of the folding process, i.e. support the New View of protein folding. The new view matches speed and stability [Baldwin, 1994 #221]. Accordingly, native proteins are both fast-folding and stabilized by a sufficiently
deep global energy minimum. The goals of achieving a global minimum, and doing it at a sufficiently high speed are not probably mutually exclusive, in contrast to the original arguments of Levinthal. The attainment of a global energy minimum, and quickly, is possible provided that the folding structure has the right (designable) sequence of residues, and that they interact via potentials that strongly favor the native conformation. MC simulations of Karplus and coworkers demonstrated that some sequences indeed possess the ability to fold rapidly and, at the same time, form thermodynamically stable structures [Sali, 1994 #237; Shakhnovich, 1994 #238; Sali, 1994 #236].

Harrison and Durbin suggested that protein folding resembles the assembly of a jigsaw puzzle: it can start anywhere and would proceed along numerous pathways [Harrison, 1985 #239]. Statistical mechanical theories and simulations support, on the other hand, the view of an energy landscape: a free energy surface in the multi-dimensional space of conformations. Under native state conditions, the unfolded chains have high energy; they occupy high energy positions in the landscape; whereas the native state is represented by a funnel-like global minimum (Figure 2.3.3a). Folding is viewed as a downhill diffusion-like process in which the individual unfolded chains undergo asynchronous kinetics, each being subject to distinct Brownian motions, until all of them ultimately find their way to the native structure. This view has been likened to the flow of water down mountainsides of complex shapes. In this metaphor, Levinthal's paradox would refer to a golf-course landscape in which the search for state N is random, a two-state folding would correspond to smoother funnels, and multiphasic transitions to rugged landscapes with kinetic traps. Some proteins can still undergo multi-stage unfolding along this funnel. See for example the description for apo-myoglobin folding by Ballew et al [Ballew, 1997 #259] in Figure 2.3.3b, which involve an early formation of an A•(H•G) subunit indicated by HX/2d-NMR studies [Hughson, 1990 #117; Jennings, 1993 #104].
**Figure 2.3.3a.** Energy landscape representative of the free energy distribution in the conformational space of proteins. The native state is represented by the central funnel-like minimum. The ruggedness of the surface increases with increased intermediates or kinetic traps. (Reproduced from [Dill, 1997 #233], Figure 4)

**Figure 2.3.3b.** Multistage folding of apo-Mb deduced by combining the results from laser T-jump induced folding fluorescence spectra (nanoseconds to microseconds range) and pH-induced folding NMR/stopped flow data (above milliseconds). On the left a hypothetical free energy surface conforming with a gradual formation of structure is displayed. (Figure 3 of [Ballew, 1997 #259])
Protein Folding Kinetics
As seen from Transition-State Theory

\[ \ln k \]

(Denaturant, temperature)

\[ (2-States) \]

Chevron plots

\[ D \leftrightarrow N \]

\[ k_f \]

\[ k_u \]
Conformational Entropy Barrier
Degree of Freedom \( (i) \)

Reaction Coordinate \( () \)

Energy Barrier
Macroscopic rate is a collective property, not a property of a single bottleneck.

Series Process

```
+------------------+
|                  |
|  60              |
+------------------+
|                  |
|  50  10          |
+------------------+
|                  |
|  60              |
+------------------+
```

macroscopic rate ∫ bottleneck rate

Parallel Process

```
+------------------+
|                  |
|  40              |
+------------------+
|                  |
|  50  10          |
+------------------+
|                  |
|  60              |
+------------------+
```

macroscopic rate > bottleneck rates

Spectrum

```
+-----------+
|           |
| 60  50  40|
+-----------+
|           |
| 10        |
+-----------+
```

Spectrum

```
+-----------+
|           |
| 60  50  40|
| 10        |
+-----------+
```
Folding is Faster than Microscopic Transition Rates

Folding Rate

Microscopic transition rates

- $\ln k$

Jack Schonbrun
How Structure Emerges Along a Funnel
Value Analysis *

\[ G = -RT \ln K \]

\[ G^\dagger = -RT \ln k_f \]

At mutation site:
TS has Native-like structure

\[ G^\dagger = \frac{G^\dagger}{G} = 1 \]

TS has Denatured-like structure

\[ G = 0 \]

Negative values come from Redirected Flow in Parallel Processes

\[ \text{Rate} = \left( \frac{7}{10} \right) 1 + \left( \frac{3}{10} \right) (0.1) = 0.73 \]

\[ \text{Rate} = \left( \frac{9}{10} \right) 1 + \left( \frac{1}{10} \right) (0.1) = 0.91 \]

Destabilization leads to higher folding rates

• Banu Ozkan
Mechanism of Chaperone Action

Series model dilemma:
- How to recognize specific TS?
- Unfolding can't help a protein fold

Parallel model solution:
- Unfolding a protein can help it fold

Summary--
2-state Kinetics can come from:

Pathways
- single rxn coord,
- bottleneck step,
- macro-rate < slowest micro-rate,
- macrostates correspond to microstates

OR

Funnels
- multiple routes,
- early acceleration,
- macro-rate > fast micro-rates
- macrostates are ensembles
• Kinetics is a collective property of landscapes. Not a property of a single trajectory.

• In 2-state folding, what is the barrier? The whole folding process, not just collapse.

• Transition States are broad. They overlap with Denatured States.

• Nonclassical values are evidence for parallel steps.

• Terminology that applies to series processes, but not necessarily to parallel processes:
  (before, after)
  (backward, forward),
  (productive, unproductive (intermediates))
Pathway Model: Collapse comes first, then detailed structure.


Funnel Model: Collapse and structure formation are simultaneous

This side is about kinetics, reaction diagrams:

This side is about thermodynamics & energy landscapes:
The Classical Transition State

- Macrostates are localized ensembles of microstates.
- States are in series and don't overlap.
- Single reaction coordinate. Forward & backward directions.

\[ G = G_f^+ - G_u^+ \]

Energy

localized ensembles: \((C_1 \; C_2 \; C_3) \ldots \; (C_i \; C_{i+1}) \ldots \; (C_N)\)

microstates

- Macrostates are localized ensembles of microstates.
- States are in series and don't overlap.
- Single reaction coordinate. Forward & backward directions.