# Introduction to Computational Structural Biology

## Part I

### 1. Introduction

The disciplinary character of Computational Structural Biology
The mathematical background required and the topics covered
Bibliography

### 2. Probability Theory

Elementary probability spaces and product spaces
Permutations and combinations
Conditional probability
Problems in combinatorics
Random variables
Cumulative distribution functions and probability density
Expectation value, variance, and sampling

### 3. Statistical Mechanics

The canonical partition function and its relation to thermodynamics
The ideal gas
The statistical character of statistical mechanics
The most probable term approach
Fluctuations and phase transitions
Different ways to solve problems in statistical mechanics
Simulation techniques - the Monte Carlo method and molecular dynamics.

### 4. Polymer Physics

Statistical mechanics of Gaussian (ideal) chains
Real chains
The effect of solvent and temperature on the chain dimensions– the Flory's theta point
Phase transitions in polymer systems
Protein folding in the context of polymer collapse

### 5. Protein Models

Force fields and their derivation
The rugged energy surface – the difficulty to fold a protein

Methods for conformational search – energy and free energy
as criteria of stability.

# Part II

# I. Introduction

## 1. Proteins Are Informational and Functional Biological Polymers

a. Proteins perform many of life's functions
b. Proteins are informational macromolecules

## 2. Protein Folding Problem

a. Proteins have large numbers of conformational degrees of freedom
b. Yet, each protein folds into a *unique* biologically active configuration, its *native* state, under physiological conditions
c. The passage between native and denatured states is often sharp and *reversible*
d. A fundamental paradigm of protein science: the amino acid sequence encodes the three-dimensional structure, which in turn determines the function
e. Definition of the "Protein Folding Problem": deciphering the relationship between amino acid sequence and three-dimensional structure
f. Thermodynamic equilibrium *vs.* kinetic accessibility
g. The concept of conformational energy landscape facilitates our understanding of much of the complex behavior of biomolecules

## 3. Medical Reasons to Pursue

a. Many, if not most of our diseases have their origins in our genes
b. The range of effects of mutations, and their functional and medical implications, are extremely broad
c. An important step in controlling biological activities is to understand signalling and control mechanisms

## 4. Promising Future for Computational Biology

a. There has been recently an exponential growth in sequence and structure data from experiments.
b. Experiments can also probe folding mechanisms and functional motions, and foster the construction of more accurate models for computer-aided studies
c. A major challenge to the structural community: Structural Genomics
d. What is the present task of the computational biologist?
e. Cell simulations – An ultimate goal

# II. Structure

## 1. Conformational Properties of Amino Acids. Implications for Protein Structures

a. Proteins are made of amino acids
b. The different structures of amino acids lead to different properties
c. In proteins, amino acids are connected by peptide bonds
d. Ramachandran maps indicate the accessible ranges of $\phi$ and $\psi$ angles
e. Virtual bond model provides an accurate description of the conformational preferences of the backbone
f. Amino acid sidechains prefer angles near their ideal rotational isomeric states

## 2. Proteins Have Hierarchies of Structure

a. Secondary structures: Helices and sheets are common motifs of proteins
b. Supersecondary structures are combination of secondary structures
c. Tertiary structures result from the packing and higher level organization of secondary and supersecondary structures.
d. Contact maps describe protein topologies
e. Quaternary structures
f. Different structures have different functions, at different hierarchical levels.

## 3. Structural Characteristics of Globular Proteins

a. Proteins are well-packed
b. Globular proteins are compact, but not spheres
c. In globular proteins, nonpolar monomers are most frequently on the interior, with charged and polar groups usually on the outside
d. There is much hydrogen bonding in proteins
e. Sidechain packing: is it an instrumental way of selecting and consolidating a fold?

## Appendix

A1. Macromolecular Conformations. Models and Definitions
a. Sets of coordinates for backbone conformations
b. Torsional angles of backbone bonds are a major determinant of macromolecular conformation
A2. Bond-based coordinates for describing polymeric structures
A3. Transformations between Cartesian and generalized coordinates
A4. Virtual bond representation of inter-residue coordination angles

# III. Thermodynamics and Energetics

## 1. Driving forces in protein folding

a. Hydrophobicity is the dominant force driving protein folding.
b. All potential hydrogen bonding groups are satisfied in folded structures.
c. What is hydrophobicity?
d. Nonpolar molecules show an aversion, entropic in origin, to water.
e. There is a competition between solvent entropy gain and protein entropy loss upon folding.

## 2. Solvation free energies can be estimated by group contribution methods

a. Component modelling is a useful approach for estimating physicochemical properties of multicomponent systems.
b. Thermodynamic data for transferring solutes between different phases and/or solvents give information on the solvation thermodynamics of biomolecules
c. Solvation free energies of proteins can be viewed as the result of the contributions from individual residues
d. Solvation contribution to unfolding free energies can be approximated as the sum of atomic contributions.

## 3. Experiments on folding thermodynamics

a. Calorimetry shows both heat and cold denaturation.
b. Temperature dependence of entropy and enthalpy are important determinants of folding/unfolding free energies.
c. Temperature dependence of hydration energies can be estimated from group contributions, using temperature-dependent heat capacity data
d. Unfolding enthalpies are small compared to hydration enthalpies, revealing the important contribution of intramolecular interactions to stability
e. The configurational entropy decrease accompanying folding can be estimated from statistical analysis of databank structures
f. A delicate balance between large entropic and large enthalpic effects determines folding equilibrium

## 4. Two-state and multiple state transitions

a. All-or-none transitions hold only for small, single domain proteins.
b. Stable intermediate states (molten globules) are revealed in some unfolding experiments.
c. The overall enthalpy and entropy change are independent of the mechanism of unfolding

### 5. Empirical potentials derived from structures

a. Why is there a need for empirical potentials?
b. The extraction of empirical potentials relies on the applicability of the Inverse Boltzmann Law
c. How to use a simple model to extract potentials?
d. Databank structures reveal significant differences in the coordination numbers and optimal interaction distances of hydrophobic and hydophilic residues
e. Reference States are critical for the application of empirical potentials
f. Residue-solvent interaction potentials dominate the effective inter-residue contact energies
g. The simple contact potentials have converged, from a comparison of values derived with a large and with a small numbers of proteins
h. More detailed potential functions can capture the greater specificity of inter-residue interactions at physically closer approach
i. Empirical solvent-mediated inter-residue potentials hold for both intramolecular and intermolecular contacts
j. The dominant role of solvent mediation permits us to express the inter-residue interactions in terms of a reduced set of single-body potentials
k. Using empirical potentials for characterizing stabilities of substitution mutations requires some characterization of the denatured state
l. Although broadly applicable there are still some limitations to applications of empirical potentials
m. Other categories of potential functions have also demonstrated their utility

# V. Kinetics

### 1. Mechanism of folding

a. All-or-none or sequential transition? Single or multiple pathways?
b. An overview of proposed mechanisms
c. New View of protein folding kinetics. Energy landscapes

### 2. Kinetic Intermediates

a. Characterization with Circular Dichroism and Fluorescence Spectroscopy
b. Hydrogen Exchange and Proton NMR for probing kinetic intermediates
c. Time-resolved one-dimensional NMR spectroscopy
d. Processes Occurring in the Submilliseconds Regime
e. Limitations of conventional methods. Need for exploring faster motions
f. Laser Temperature Jump (T-jump) and Ultrafast Mixing
g. Utility of combining data from different techniques

### 3. Classical kinetic modeling of protein folding/unfolding

a. Two-state transition
b. Sequential transition from U to N
c. Steady state approximation for the intermediate
d. Sequential transition with a pre-equilibrium between initial state and intermediate state
e. Off-pathway intermediate formation
f. More complex kinetic schemes

### 4. Transition states. Effects of mutations. Probabilistic description

a. Rate limiting or rate controlling steps. Apparent activated states
b. Effect of point mutations on folding kinetics. Φ-value analysis.
b.Master equation formalism for folding kinetics

## VI. Conformational dynamics. Relationship to function

### 1. Fluctuation-dissipation theorem

a. A protein in its native state is not a static conformation: it samples many fluctuating conformations around the native state coordinates
b. Experiments indicate a correlation between equilibrium fluctuations and functionally important fluctuations
c. The greater the stability, the fewer are the accessible large fluctuations
d. Ligand binding mechanisms conform more closely with energy landscape models than simple mass-action models

### 2. Dynamics of polymeric chains

a. Polymer dynamics results from a superposition of conformational jumps between isomers and fluctuations around isomeric minima
b. Fluctuations and conformational jumps are coupled to localize the motion
c. The simplest model for describing polymer dynamics is the bead-and-spring model of Rouse
d. The Rouse matrix expresses the total potential in a form amenable to analytical treatment
e. Rouse chain obeys Langevin equation of motion
f. Normal mode analysis (NMA): A classical method for investigating collective fluctuations

## 3. Dynamics of folded proteins. Gaussian network model

a. The dynamics of folded structures is dominated by fluctuations near equilibrium coordinates
b. Proteins in the folded state have been modelled as Gaussian networks: nodes are the residues, and connectors are their interactions
c. Two major assumptions in the Gaussian network model (GNM): all inter-residue potentials are harmonic, and the same force constant holds for all interactions
d. The Kirchhoff matrix may be viewed as the counterpart of the Rouse matrix, for proteins: it fully controls the fluctuation dynamics.
e Mean-square fluctuations are inversely proportional to residue coordination numbers to a first approximation
f. Mode decomposition of fluctuation spectrum reveals the dominant mechanism of motion
g. GNM analysis is a special case of conventional NMA for the case isotropic fluctuations without residue specificity
h. Gaussian fluctuations are valid for low resolution descriptions of dense systems

## 4. Contribution of nonlinear effects to equilibrium dynamics

a. Non-linear forces induce anharmonic fluctuations
b. In addition to anharmonic fluctuations, proteins can sample multiple states separated by energy barriers
c. Comparison of results from NMA and MD simulations indicate that anharmonicity essentially resides in the slowest mode(s) of motion
d. Neglect of anharmonic modes leads to an underestimation of the fluctuation amplitudes
e. The slowest modes appearing in MD simulations can be biased by sampling inefficiency
f. Are large scale anharmonic motions frozen out in crystal structures?

## 5. Thermal fluctuations in crystalline forms

a. Theory of X-ray diffraction. What are the Debye-Waller temperature factors?
b. Fluctuations of backbone atoms conform with the Gaussian approximation; those of sidechains do not: they are significantly larger
c. Experimental Debye-Waller factors can be closely reproduced by the GNM
d. The agreement between GNM and experiments implies that number and order of inter-residue contacts are the major determinants of backbone flexibility
e. How important is residue specificity in fluctuation dynamics?
f. Secondary structural elements are characterized by reduced mobilities.
g. There is a significant change in fluctuation dynamics upon complexation of biomolecular structures.

### 6. Fluctuations in solution

a. Broad range of utility of NMR: from characterization of non-crystallizable structures to investigation of protein solution dynamics.
b. Structures in solution from NMR are almost identical to X-ray structures; their fluctuations may differ in some cases
c. The major differences between X-ray and NMR structures occur at regions engaged in the slowest modes of motion
d. H/D exchange coupled with 2D-NMR measures the degree of protection of particular amide bonds under different conditions.
e. Two mechanisms of H/D exchange have been proposed: global unfolding and fluctuations
f. H/D exchange near native state conditions is dominated by fluctuation dynamics
g. There is a correlation between H/D exchange data under mildly denaturing conditions and B factors from X-ray crystallography
h. NMR relaxation experiments probe orientational mobilities on a local scale; and these, in turn, depend on bond torsional flexibilities
i. GNM predictions reveal that the reorientations of the individual amide bonds depend on the accessibility of a mechanism accommodating local conformational changes

### 7. Inferring function from structural mode analysis

a. Observed dynamics results from a superposition of a broad range of modes.
b. Slowest (global) modes define the mechanism of motions relevant to function.
c. Examination of slowest modes accurately locates the hinge-bending sites, binding sites and recognition loops.
d. Conserved residues are highly constrained in the global modes.
e. Recognition loops are distinguished by high flexibility, which facilitates the binding to a diversity of substrates.
f. Ligand binding enhances the overall cooperativity of motions.
g. Catalytically active sites are usually located near hinge-bending sites. Is this a possible prerequisite for efficient conversion between chemical and mechanical energies?
h. Cross-correlations between domain motions are consistent with processing mechanisms of biomolecular complexes.

# VII. Computations in Biology

### 1. An overview of the present state-of-the-art

a. What is the present task of the computational biologist?
b. Why do we need automated methods of sequence and structure analysis?
c. A challenging problem: prediction of folded structure. How can theory help?
d. How well do the different computational approaches work?
e. Computational genomics and proteomics.

## 2. Bioinformatics for sequence analysis

a. Bioinformatics: a discipline gaining importance with the advent of Genomics
b. Methods for efficient database searches. Classification and clustering algorithms
c. Sequence-structure analyses by computational methods
d. It is not currently possible to unambiguously determine structure or function based on sequence alignments only
e. A useful general optimization technique: dynamic programming. Application to sequence alignment
f. Two basic tools for scoring alignments: substitution matrices and gap penalties
g. Global *vs* local alignments using dynamic programming.
h. How well do pairwise sequence comparisons work? What fraction of homologous proteins can be detected?
i. Multiple sequence alignment can detect up to three times as many remote homologs as pairwise methods
j. Sequence similarity matrices derived from structures are a promising way to incorporate structural features.

## 2. Secondary structure prediction

a. Secondary structure propensities of different amino acids can be estimated from statistical analysis of observed frequencies of a-helices and b-sheets
b. Examples of secondary structure prediction methods: PHD
d. Dynamic programming as an efficient tool for secondary structure optimization

## 3. Tertiary Structure Prediction

a. Protein structural classes can be predicted with a high level of accuracy from amino acid composition alone.
b. Threading, and inverse threading and folding protocols are all computational experiments for identifying sequence-structure matches.
c. Importance of considering gaps and insertions in threading experiments
h. Toy models illustrate the utility and limitations of threading.
i. Methods of structure matching
e. An example of classification into structure families by SCOP. Evolutionary assumption
j. How many fold families are there?
k. NMR distance constraints are utilized to generate structures in homology modelling

## 4. Molecular simulations (MC, MD)

a. Development of efficient energy minimization scheme: the first requirement for predicting equilibrium structures
b. Monte Carlo/Metropolis algorithms are efficient coarse grained tools for studying macromolecular dynamics.
c. The most detailed current method of investigating protein dynamics is molecular dynamics (MD) simulations at atomic level.
d. MD simulations have originally demonstrated the existence of a dynamic equilibrium near native state.
e. A major utility of MD simulations/energy minimization methods is for atomic structure refinements.
f. Comparison of vacuum simulations with those performed with explicit solvent molecules indicates the importance of solvation.
g. MD simulations cannot presently examine folding kinetics.
h. Conformational sampling inefficiency is a major drawback limiting the applicability of MD simulations to small systems and subnanosecond time regime.
i. Information on the rate and mechanism of helix formation/dissolution is obtainable from MD simulations.
j. Transition to partially unfolded structures with molten globule characteristics has been observed in MD simulations.
k. Other applications of MD include an examination with free energy perturbation methods of ion diffusion through channels.

## 5. Coarse grained simulations

a. Coarse-grained simulations permit exploration of larger size/longer time dynamic processes that cannot be explored with atomic MD simulations.
b. Two major approaches: On-lattice and off-lattice simulations
c. There is a trade-off between the level of accuracy of the model and the efficiency of computation.
d. Coarse grained simulations necessitate the use of empirical energy parameters.
e. Accurate choice of energy parameters is critically important for the success of coarse grained simulations.
f. Folding with limited success has been observed for small proteins.
g. Lattice models for proteins indicate multiple folding nuclei for a nucleation- collapse mechanism.

## 6. Docking

a. Docking involves recognition of molecular surfaces
b. Optimal burial of hydrophobic surfaces offsets the entropy loss of binding
c. Shape complementarity/geometric fitness is an important condition for docking

d. Automated docking algorithms are essentially based on rigid bodies induced fit binding mechanism
e. Newly developed docking algorithms include the effects of conformational flexibility

## 7. Searching for correlations

a. Combinatorial chemistry: a new field of study with potential pharmaceutical applications
b. Classification of drugs is possible by database screening techniques coupled with correlation analyses.
c. Functional correlations *vis a vis* structural correlations