The rise of computational biology

1. Computer simulations as a new tool of scientific research

The era of computer simulations (CS) started with the development of the first digital computers in '40s and '50s (for example, MANIAC (stands for "MATHEMATICAL AND NUMERICAL INTEGRATOR AND COMPUTER) computer in Los Alamos). First Molecular Dynamics simulations were reported in 1956. Now CS are the standard tool for studying diverse set of problems. The need in CS arises because very few problems in science afford analytical ("pencil and paper") solution.

CS closely interact with experiment and theory: **Experiment ↔ CS ↔Theory**

a) Models used in CS are often based in theoretical approximations. Accordingly, CS may be used to test theories.
b) CS themselves do not offer physical understanding. The results of CS must be "processed" by the theory
c) CS cannot substitute experiment, because CS are "what you put is what you get"-type of numerical experiments. Therefore, CS critically depend on experimental input for verification of their results. Experiments often provide initial conditions for CS.

CS are used to predict properties of materials under extreme conditions (high temperature, pressure etc) or to explore the behavior of complex systems (e.g., protein folding, aggregation of proteins). The latter is the most important because it put CS on the edge of scientific research.

Homework: *Read Chapter 1 of "Understanding Computer Simulations" by Frenkel and Smit)*

2. The rise of Computational Biology

Computational Biology (CB) or Bioinformatics incorporates virtually any application of numerical techniques for biological systems. Explosion of newly available biomolecular experimental data coupled with the remarkable increase in CPU power form the foundation for the rapid advances in CB. Some of the CB problems are (i) comparison and analysis of genome sequences, (ii) making a relation between sequence, structure, and function (functional genomics), (iii) computation (refinement or prediction) of biomolecular structures (structural genomics), (iv) study of biophysics of biomolecules. The last two problems are the primary targets of CS. The use of CS in CB includes (i) the study of the dynamics of structural changes in biomolecules, including the assembly or unfolding of biomolecular native structures, aggregation; (ii) thermodynamics of biomolecules; (iii) the effect of external conditions (such as temperature, denaturant, pH) and sequence mutations on the properties of biomolecules; (iv) prediction of native structures from sequence of amino acids, etc.

3. Sequence → Structure

With the completion of many genome projects, which produce myriads of new sequences of proteins, the focus is shifting towards obtaining structural and biophysical information for these sequences. *CS of biomolecules is the main source of such information.*

<u>Some major steps in genome sequencing:</u>

1995: First genome of the bacterium is determined. *Haemophilus influenzae* bacterium, which causes secondary infections following flu, has 1,700 genes. Yeast genome was sequenced in 1996 (about 6,300 genes).

1998: First genome of the multicellular (round worm has ~10^3 cells) organism is obtained. Its genome contains about 19,100 genes. About one-third of worm proteins are similar to mamalian that provides an opportunity to use worm as a model.

1999: Genome of *Drosophila* fruit fly (about 10^6 cells) contains "only" 13,600 genes (< than worm).
Note: because many genes may be spliced differently during transcription/translation, same regions of DNA may code different proteins.
Both fruit fly and round worm have the same number of core proteins encoded in their genomes.

2000: Genome of mustard plant is sequenced. It contains about 25,500 genes. Many genes have dublicates (perhaps, for radiation protection); there is also some "junk" in the genome as well.

2002: The draft of rice genome is determined promising a huge impact on agrobiotechnology. The number of genes is expected to be from 33,000 to ~60,000.

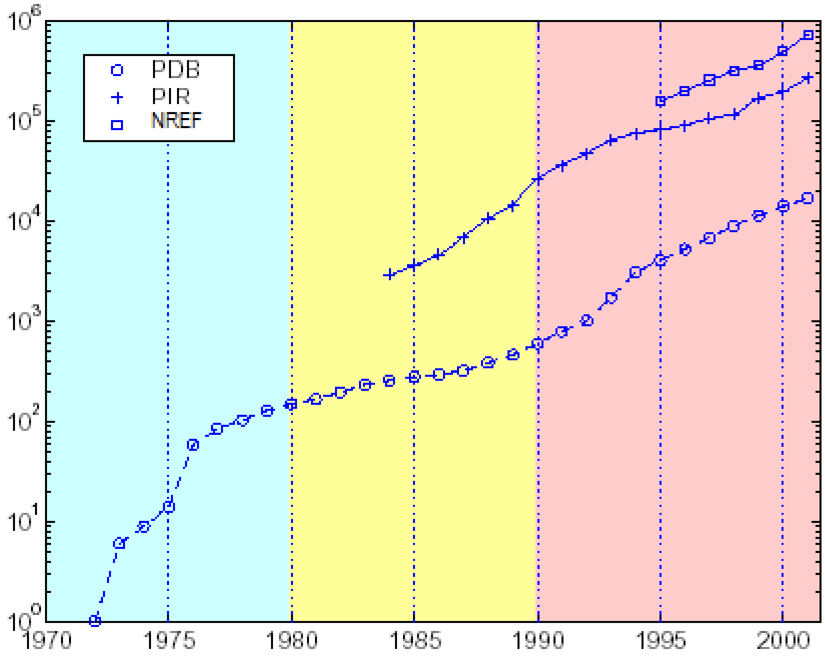
<u>Human Genome:</u>

International Human Genome Project was launched in 1990 to sequence about $3\text{x}10^9$ human DNA pairs. First two chromosomes (#21,#22) were sequenced in 1999 and 2000. In 2000 85% of human genome was mapped accurately with high redundancy, in 2003 with 99.9% accuracy.

How many genes are in human DNA?
Relatively few, the best estimate is roughly about 30,000. However, organism complexity is also determined by the number of proteins and RNAs, not the size of genome alone. Many genes may be "read" in alternative ways ("alternative splicing") up to 3 times. The average human protein consists of about 450 amino acids and is generally more complex than animal counterpart. Furthermore, only 2-3% of DNA appears to be directly involved in gene coding. See www.ncbi.nlm.nih.gov/genome/guide/human for further information.

The most important task now is to translate sequence information into structural information. This implies determination of protein structures and functions coded by genome.

Lag between the number of sequences and the number of resolved structures is demonstrated by the following plot (Schlick "*Molecular Modeling and Simulations*"):



The plot shows the number of protein sequences deposited in Georgetown University PIR-PSD (crosses) and PIR-NREF (squares) databases as well as the number of structures in PDB database (circles) *vs* time. The web sites for PIR and PDB databases are pir.georgetown.edu and www.rcsb.org, respectively. As of now PIR-PSD contains 283,416 non-redundant, fully annotated and classified sequences, while PIR-NREF database, which includes sequences from several different databases, has grown to 1,638,166 sequences. In contrast, PDB contains only 26,013 structures.

Homework: *read Chapter 1 and Chapter2 (section 2.1) of "Molecular Modeling and Simulations" by Schlick.*

4. Computational methods for solving protein structures

A. Comparative methods are based on homology (similarity) modeling.
Basic emperic idea is that similar sequences have similar structures. Consider the following observations:

(i) > 50% similarity implies < 1 angstrom of backbone RMS deviation. Structure deviations are larger in loops, side chain packing;
(ii) 30% thru 50% similarity implies that 90% of the chain can be modeled with 1.5A backbone RMS;
(iii) 30% is a threshold value, below which the structures tend to be dissimilar (predicted fold may be globally wrong).

One factor that helps homology modeling is that active sites in proteins tend to conserve sequence composition, therefore the quality of structure prediction for these important areas is usually high.

Homology modeling involves sequence alignment, i.e., sequence comparison performed using **BLAST** and its variants **(**www.ncbi.nlm.nih.gov/Education/BLASTinfo). Furthermore, we need to know the set of representative structures for each type of protein conformations. The estimate is that about 16,000 non-redundant protein structures will cover 90% of structural families. As a rule, for each solved structure the structures of about 100 new sequences can be modeled.

References:

1. B. Al-Lazikani, J.Jung, Z. Xiang, and B. Honig "Protein structure prediction" Curr. Opion. Struct. Biol. **5**, 51 (2001).
2. M.A. Marti-Renom et al "Comparative protein structure modeling of genes and genomes" Annu. Rev. Biophys. Biomol. Struct. **29**, 291 (2000).

B. *Ab initio* prediction: Can we obtain the structure using sequence information alone? The most direct implementation of *ab initio* prediction is to "plug" the sequence into the program that simulates protein dynamics, hoping that the native structure will be found. The implicit assumption in this approach is that the native structure corresponds to a global energy minimum. Unfortunately, simulations so far cannot routinely predict the structure for several reasons:

(i) limited timescale (< 1 microsec) available for simulations;
(ii) inaccurate description of interactions in biomolecules;
(iii) poor sampling algorithms

Yet this approach is the most desirable, because it is based on first-principle knowledge of atomic interactions and structure. It can be used also to study protein function, folding and misfolding. Furthermore, homology modeling methods cannot account for the environment effects, such as temperature, pH etc. There are many other *ab initio* methods, which typically utilize the interaction parameters derived from structural databases.

Reference:

C. Hardin, T. V. Pogorelov, and Z. Luthey-Schulten "*Ab initio* protein structure prediction" *Curr. Opion. Struct. Biol*.**12**, 176 (2002).

Homework*: Read the article "Molecular dynamics simulations of biomolecules" by Karplus and MaCammon (Nature Structural Biology* ***9****, 646 (2002)).*