Appendix D

Homework Assignments

Please Note: (1) Files mentioned throughout the homeworks can be obtained from the course site, through our group's home page monod.biomath.nyu.edu *or by contacting the author at* schlick@nyu.edu. (2) Insight modeling commands may have changed since the time of this writing and may need updating by instructor and/or students.

Assignment 1: Sequence and Structural Databases, Molecular Modeling Perfective

1. **Molecular Modeling Resources**. Search the web for resources in *molecular modeling*. Look, in particular, for tutorials and instructional material. A good place to start is the NIH site:
   cmm.info.nih.gov/modeling/, which also provides links to many "Related Web Sites". (Make use of bookmark-type browser utilities to keep useful web sites handy for future use).
   Submit information from two of the most valuable sites you discover (printout) along with a description of how you found the material and what you found most useful.

2. **Sequence and Structure Information Databases**. Search the web for protein (amino-acid) sequence and structure databases. Examples of sequence databases are: PIR, Swiss-Prot, GenPept, and NRPR.[1]

   (a) Plot the amount of available *sequence* database as a function of year, going back as far as possible, to the 1970s. Plot the information on both a regular scale and on a logarithm scale.

   (b) Similarly, plot the amount of *structural* information available as a function of year, on both a regular and a logarithm scale.

   (c) Plot the *sequence and structure* information on the *same plot* in both standard and logarithm views. What can you say about the rate of growth of sequence and structural information? Discuss these finding in relation to the Human Genome Project.

3. **Early Molecular Modeling Literature and Current Progress**. Read two articles dealing with early molecular modeling work:

   - B. J. Alder and T. E. Wainwright, "Studies in Molecular Dynamics. I. General Method", *J. Chem. Phys.* **31**, 459–466 (1959).
   - G. Némethy and H. A. Scheraga, "Theoretical Determination of Sterically Allowed Conformations of a Polypeptide Chain by a Computer Method", *Biopolymers* **3**, 155–184 (1965).

   Do not worry about not understanding the technical details for now.
   Then also read more recent articles describing current progress in the field and another discussing issues in validating simulation results:

[1] PIR was established in 1984 by the National Biomedical Research Foundation (NBRF) and is a good starting point for protein database searching. PIR is somewhat more comprehensive than SwissProt but smaller and better annotated than GenPept (which also includes many hypothetical sequences of unknown function. Since 1999, the NBRF has added a new section to PIR called PATCHX which contains a non-redundant set of all other protein sequences not included in PIR (from other databases), with subsequences removed. Thus, PIR supplemented by PATCHX provides a comprehensive collection of protein sequence data in the public domain. Any search through PIR will automatically include PATCHX. The SwissProt database is useful for searches limited to well annotated sequences, and GenPept is useful for searching all possible sequences, including those that have unknown functions.

The best known structural databases are the Protein Data Bank (PDB) and the Nucleic Acid Database (NDB).

The PDB, managed from 1971 through June 1999 by the Brookhaven National Laboratory, is now operated by the Research Collaboratory for Structural Bioinformatics (RCSB) (www.rcsb.org/), a consortium among Rutgers University, the University of California at San Diego, and the National Institute of Standards and Technology. The RCSB has introduced new features, such as a web-based tool for data deposition, fast data processing systems, and new search engines (text-based and data-based), both with extensive reporting capabilities.

The NDB, pioneered in 1992 by the Rutgers RCSB leader Helen Berman, similarly assembles and distributes structural information about nucleic acids (ndbserver.rutgers.edu/). NDB contains an atlas, an archive, and a sophisticated search engine to access the data.

- J. A. Board, Jr., L. V. Kalé, K. Schulten, R. D. Skeel, and T. Schlick, "Modeling Biomolecules: Larger Scales, Longer Durations", *IEEE Comp. Sci. Eng.* **1**, 19–30 (1994).
- W. F. van Gunsteren and A. E. Mark, "Validation of Molecular Dynamics Simulation", *J. Chem. Phys.* **108**, 6109–6116 (1998).

First describe (in about two pages) the difficulties that Alder and Wainwright enumerate in 1959 regarding molecular dynamics simulations. Then discuss the issues that are still serious limiting factors today, as well as others, such as those noted by van Gunsteren & Mark. Have any of the original limitations been resolved or are likely to be resolved in the near future?