Appendix D

Homework Assignments

Please Note: (1) Files mentioned throughout the homeworks can be obtained from the course site. (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 Perspective

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).
   - A. Rahman and F. H. Stillinger, "Molecular Dynamics Study of Liquid Water", *J. Chem. Phys.* **55**, 3336–3359 (1971).

[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) (home.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.

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

- 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).
- F. S. Collins, E. D. Green, A. E. Guttmacher, and M. S. Guyer, "A Vision for the Future of Genomics Research", *Nature* **422**, 835–847 (2003).
- J. Norberg and L. Nilsson, "Advances in Biomolecular Simulations: Methodology and Applications", *Quart. Rev. Biophys.* **36**, 257–306 (2003).
- J. E. Cohen, "Mathematics is Biology's Next Microscope, Only Better; Biology is Mathematics' Next Physics, Only Better", *PLoS Biology* **2 (e439)**, 2017–2023 (2004).
- T. Schlick, "The Critical Collaboration Between Art and Science: Applying *An Experiment on a Bird in an Air Pump* to the Ramifications of Genomics on Society", *Leonardo* **38** (4), 323–329 (2005).
- W. F. van Gunsteren et al., "Biomolecular Modeling: Goals, Problems, Perspectives", *Angew. Chem. Int. Ed.* **45**, 4064–4092 (2006).
- M. A. Gerstein et al., "What is a Gene, post ENCODE? History and Updated Definition", *Genome Research* **17**, 669–681 (2007).

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. Have any of the original limitations been resolved or are likely to be resolved in the near future?

Assignment 2: Introduction to the Insight II Modeling Package and the PDB File Structure and Retrieval

At this writing, Insight II is available as part of the Discovery Studio from Accelrys (accelrys.com/products/insight/). Other molecular modeling packages can be used instead as available.

1. **Introduction to Insight II**. If you are not familiar with Insight II you might start by reading the short manual and the tutorial from the web site.
   The manual also contains a short list of basic UNIX commands and a description of simple text editors. Some of the displays on our computers are slightly different from those described in the tutorial but the differences are not critical. For example, in our version the help window automatically follows the tasks invoked from the pulldown menus, and the dial boxes are located on the left, rather than the right, side of the screen. For more information, refer to the Insight II User Guide.

   Note: after opening Insight II, ignore the message about detection of the unlicensed mode. Our site does not have license for the Sketch module. Repeated warning messages might occur during the "build in" Insight II Pilot *tutorial.*

2. **Running Insight II**. Before starting Insight II, you must define the set of environmental variables by running two commands:
   `source /local/msi/cshrc`
   Alternately, you can insert these lines into your `.cshrc` file and they will be executed by the system automatically every time you log on. In that case, you will only need to specify the command
   `source .cshrc`
   after editing your `.cshrc` file the first time. When this insertion is complete, you can run Insight II by specifying the command
   `insightII`
   at the UNIX prompt. Remember, UNIX commands are case sensitive.

   Note: NYU staff may have already inserted these commands in your `.cshrc` *file.*

3. PDB **Structures**. Check the PDB web site for information about the type of stored 3D structures (i.e., proteins, DNA, RNA, DNA/protein complexes, etc.) and the amount in each group. Report your findings.

4. **Retrieval of** PDB **Files**. Using the web PDB browser, find coordinate files for the crystal forms II and III of Bovine Pancreatic Trypsin Inhibitor (BPTI) among the many BPTI entries. From the PDB Home page, go to **Searching and Browsing PDB** and then choose **PDB's web Browser**. You can search by specifying the abbreviation "BPTI" in the Compound window. When the search is completed, records containing BPTI (including its mutated forms) will be displayed at the bottom of the page. Note their ID codes (a number followed by three letters). The two middle let-

ters in this code constitute the name of the subdirectory where the file of interest resides. For example, the subdirectory name for ID code `1abb` is `ab` and the file name `pdb1abb.ent` . Ftp to `ftp.rcsb.org` and login as `anonymous` (the password instructions will be on the screen). Change directory to `pub/pdb/data/structures/divided/pdb/ab` and get the desired file with the command
`get pdb1abb.ent.Z`

5. **Format of PDB Files**. Read the text in the top of both PDB files and describe the differences in the number of recorded residues, structure resolution, number of solvent molecules, experimental conditions, etc.
   Attach to the assignment sheet a printout of a few lines, starting with the word ATOM, from a PDB file; mark with arrows and describe the content of each specific format field. (The PDB browser contains information about the format; see also the original paper on PDB files: *J. Mol. Biol.* **112**, 535–542, 1977).

6. **Displaying a Protein in Insight II**. Retrieve from PDB the file of mutated form of BPTI (ID = `7pti`), and start Insight II. From the top menu bar select Molecule [2] and then choose Get. Press the PDB button in Get File Type and the User button specify the directory. Select the file of the `7pti` structure. (Do not press Heteroatom button). Execute. The structure of BPTI should now be on the screen. Use Object / DepthCue and then Transform / Clip for viewing the protein.
   Change the display (Molecule / Display) to Backbone only to speed up the response time. Label the mutated residues (Molecule / Label).
   Repeat the operations described above to display the structure of BPTI's crystal form II (keep both structures on the screen).
   Now, overlay both structures by selecting Overlay from Transform. In few paragraphs describe the structural differences between both forms of BPTI.

7. **Ramachandran Plots**. To create a file listing with all the dihedral angles ϕ and ψ for a protein, you can use Protein / List from the **Biopolymer** module. Select Dihedrals and the protein (any of the structures). Press List_to_file button, specify the file name, and execute. From the recorded data create a scatter plot (phase diagram), so that each point corresponds to

[2] The following notation will be used throughout the homework assignments:

- Pulldown corresponds to a menu bar pulldown.
- Command corresponds to a command from the pulldown menu.
- Option corresponds to an option in the dialog box of a given command.

one (ϕ, ψ) value.

Summary of Items to Hand in:

(a) Data with the amount of PDB 3D structures for each category of biomolecules.
(b) Description of the differences in the informational part of the PDB files for form II and form III of BPTI.
(c) Explanation of the PDB format for storing atomic coordinates.
(d) Description of the structural differences between BPTI (form II) and the mutated form (7pti).
(e) The table with dihedral angles ϕ and ψ listed for BPTI.
(f) Scatter plot with points (ψ, ϕ) for each residue of BPTI.

Background Reading from Coursepack (Appendix B)

- M. Levitt and A. Warshel, "Computer Simulation of Protein Folding", *Nature* **253**, 694–698 (1975).
- M. Karplus and J. A. McCammon, "The Dynamics of Proteins", *Sci. Amer.* **254**, 42–51 (1986).
- Y. Duan and P. A. Kollman, "Pathways to a Protein Folding Intermediate Observed in a 1-Microsecond Simulation in Aqueous Solution", *Science* **282**, 740–744 (1998).
- H. J. C. Berendsen, "A Glimpse of the Holy Grail", *Science* **282**, 642–643 (1998).
- X. Daura, B. Juan, D. Seebach, W. F. Van Gunsteren, and A. Mark, "Reversible Peptide Folding in Solution by Molecular Dynamics Simulation", *J. Mol. Biol.* **280**, 925–932 (1998).
- R. Bonneau and D. Baker, "Ab Initio Protein Structure Prediction: Progress and Prospects", *Annu. Rev. Biophys. Struc.* **30**, 173–189 (2001).

Assignment 3: Construction and Analysis of the Pentapeptide Met-enkephalin with the Insight II Program

1. **Building a Pentapeptide**. To build the molecule met-enkephalin, whose amino acid sequence is **Tyr–Gly–Gly–Phe–Met**, invoke the **Biopolymer** module. From Residue select Append.
   Specify the molecule's name and choose Extended to specify the structural motif of the backbone. (At this stage, do not worry whether the structure is correct). Select the first residue, **Tyr**. Then add each of the remaining four residues in turn.
   You can *center* the molecule on your screen by clicking on it with the center mouse button and dragging it to the desired position. You can also *rotate* the molecule by pressing the right mouse button and dragging the molecule. By pressing both (center and right) mouse buttons and dragging the molecule you can change its position along the *z*-axis, perpendicular to the screen. You can rotate the molecule around the *z*-axis by dragging the mouse while both left and right mouse buttons are pressed. To change the representation style of the molecule, select Molecule / Render and choose any of the options. Note how the speed of executing commands (e.g., translation, rotation) is affected by the representation display.
   Now you must amend the ends of the structure you built. Switch to Protein and choose Cap. Change both the N-terminus and C-terminus moieties to the zwitterionic form (NH_3^+ and COO^-) to get a proper oligopeptide.

2. **Measurements of Met-enkephalin's Structural Parameters**. Generate a table listing all the dihedral angles in your met-enkephalin molecule by using Protein / List from the **Biopolymer** module. Select the appropriate command (Dihedrals) and direct the data to a file (List_to_file button on). This file can be viewed and edited later.
   To measure individual bond lengths or distances, bond angles, dihedral angles, etc., use Measure. Select the atoms for the measurement by clicking on them with the left mouse button.

3. **Rotameric Structures**. By a *rotamer*, Insight II refers to a different conformational arrangement of a side chain of a given amino acid. The rotameric structures identified in Insight II are correlated with the ϕ and ψ angles for a given residue (i.e., are sterically compatible).
   Select Manual_Rotamer from Residue of the **Biopolymer** module. Press the Evaluate_Energy button. Energy for a given rotameric structure will be printed in the information window at the bottom of the screen (you can scroll by pressing on arrows to its right). Keep the Nonbond Cutoff value — an option which is displayed on the screen once Evaluate_Energy is chosen — at the default value of 8.0 Å.
   Select a residue by clicking on it and sweep through all the rotamers

of that residue (while holding other rotamer conformations fixed). Next will trigger the execution of the command. In a table, report the rotamer structures for each residue by specifying the dihedral angles χ_1, χ_2, χ_3, ...(Protein / List), along with associated energies.
Now assemble the pentapeptide structure with the lowest rotamer energy and save it. Thus far we have ignored a global optimization of the structure and have only built it up from low-energy conformations of its components. *We will return to optimization later in the course, after studying minimization techniques.*

4. **Main-chain Structure**. Choose Protein / Secondary from the **Biopolymer** module. Change the main-chain configuration by selecting different motifs (Alpha_R_Helix, Alpha_L_Helix, 3–10_Helix, etc.). For each motif:

   (a) prepare a table with the dihedral angles ϕ and ψ for each of the residues (Protein / List),
   (b) list all the hydrogen bonds present in the structure.
   Measure / HBonds and Molecule / Label will be helpful here.

5. **Torsional Rotation**. Bring back your Met-enkephalin's backbone structure in the extended form. You can use the structure saved in part 3 of this assignment (File / Restore_Folder). Change the ψ dihedral angle on the second residue, **Gly**, to $60°$ and the ϕ dihedral angle on the forth residue, **Phe**, to $-60°$.
   Torsional motion around a chosen dihedral angle can be performed with Transform / Torsion command or by pressing the Torsion button on the left of the Insight II screen.
   First, click with the left mouse button on the bond which constitutes the axis of rotation, then press the Torsion button. A little cone, defining the direction of the torsion, will pop up on the screen at one of the bond ends. Now you can change the torsion angle by horizontally dragging the mouse with the center button pressed in. To exit the torsion mode press the Torsion button again (the cone will disappear).
   Calculate the distance between the N-terminus (N atom) and C-terminus (C atom) atoms. Use the Measure / Distance command. Keep the Monitor button on, and select Monitor Mode/Add. Atom 1 and Atom 2 can be selected by clicking on them with the left mouse button.
   Print a picture of your molecule (keep the distance between the N-terminus and C-terminus atoms on the screen). To save ink, the background color should be changed to white every time you print a color or black/white picture. To do so, go to Session / Environment, press the Background button and change the color to white. Now go to File and choose Export_Plot.

Select postscript, Gray_Scale and optionally Ball_and_Stick. Save the file as postscript by using Save_Device_File (the file will have the ".ps" extension). This file can be printed on any postscript printer. Hand in your printout as part of the homework.

Summary of Items to Hand In:

(a) The table with the dihedral angles for met-enkephalin.
(b) The table with the dihedral angles and energies for the rotameric structures of met-enkephalin.
(c) The table with the $\{\phi, \psi\}$ dihedral angles for each different backbone motif of met-enkephalin.
(d) The listing of the hydrogen bonds for each of the backbone motif for met-enkephalin.
(e) Printout of the met-enkephalin structure with the end-to-end link marked.

Background Reading from Coursepack

- 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).
- P. Y. Chou and G. D. Fasman, "Prediction of Protein Conformation", *Biochemistry* **13**, 222–245 (1974).
- M. Levitt and C. Chothia, "Structural Patterns in Globular Proteins", *Nature* **261**, 552–558 (1976).

Assignment 4: Creating Ramachandran Plots from Known Protein Structures and the NDB

1. **Ramachandran Plots**. Our goal is to generate Ramachandran plots for a particular amino acid residue or a group of residues. We have assembled a database of 50 proteins based on the article: M.A. Williams, J.M. Goodfellow, and J.M. Thornton, "Buried Water and Internal Cavities", *Protein Science* **3**, 1224 (1994). These files can be found in the PDB directory of Insight II prepared for our course.
   Each of you must generate two Ramachandran plots. Check the chart below for your particular assignment (on the basis of your last name).

First letter of your last name	Subgroup 1	Subgroup 2
A–N	**Ala**, **Val**, **Leu**, **Ile**	**Gly**
O–R	**Asp**, **Asn**, **Glu**, **Gln**	**Pro**
S–V	**Lys**, **Arg**, **His**	**Ser**, **Thr**
W–Z	**Trp**, **Tyr**, **Phe**	**Cys**, **Met**

   Each plot should have the data points for the (ϕ, ψ) dihedral angles, corresponding to the assigned group of residues from all the proteins in the database. To find the values of the ϕ and ψ dihedral angles in a protein you can use Protein / List command from the **Biopolymer** module. Record these angles to a file. You can use the Fortran code posted on the website (aa_select.f) for searching the Protein / List output files (called PDA files) for the dihedral angles of selected residues. Alternatively, you can write a suitable program in a different language.
   If you use the code from the website, you will need to edit it to replace all occurrences of the names `ALA, VAL`, etc. by the abbreviations of the residues you are searching for. These abbreviations must be capitalized. Compile the code and execute it with each PDA file as input. The output file should contain only two numbers per line, the ϕ and ψ angles for the specified residues. Check the numbers for correctness by comparing a few lines from this output with the numbers in the PDA file.

   Note: For plotting, you must use Insight II *so that all plots are uniform in size. We will overlay them in class! Follow the instructions below*.

   Prepare a file with your data points collected from all the proteins for each of the assigned groups of residues. These files should have the extension `tbl (filename.tbl)` . In addition, you must format these files for Insight II by inserting the 12 lines as indicated:

```
#
TITLE: Phi (deg)
MEASUREMENT TYPE: Ang
UNITS OF MEASUREMENT: Deg
FUNCTION: dihedral
#
TITLE: Psi (deg)
MEASUREMENT TYPE: Ang
UNITS OF MEASUREMENT: Deg
FUNCTION: dihedral
#
#
−34.5      144.9                    This is the first line of your numeric data.
⋮
```

Note that at the top of the file there is space for your own comments and you can use as many lines as you need. Insight II will start reading the data from the line without the character # on the first column. That first line should be as indicated above.

If you have done everything correctly, you are ready for plotting. Press the Graph button on the left of the screen and select Get. Specify the data file name, dihedral1 as X_Function and dihedral2 as Y_Function. Keep the New_Graph on and execute. The zigzag appearance of the plot now requires fixing. Move the graph box near the bottom-left corner of the screen. First, connect to the object (your plot) with the command Transform / Connect. Second, move the plot by clicking on it and dragging to the desired position. The scatter plot will be produced in the next step. Select Point (only!) from the Graph / Modify_Display dialog box. Choose Star as the Point Symbol, scale it ten times and execute. Use Graph / Threshold to change the minimum value to -180.0 and the maximum value to 180.0 for both, X Graph Axis and Y Graph Axis. Scale each axis 4 times with Graph / Scale_Axis command. Change the color of any white elements in your plot with Graph / Color command. Change the background color to white. Print your plot (see instructions below) and repeat the procedure for the second data file.

Prepare a transparency for each of the two plots and bring to class. (Hand in the printed version of the plots with your homework.) Also e-mail the $\{\phi, \psi\}$ files you generated, sending each file separately and specifying your name and the group of residues in the `Subject` line.

2. **The** NDB. The next part of the homework will acquaint you to the Nucleic Acid Database, NDB, on which we will have a guest lecturer.

First, explore the database to discover what is available, and look through the newsletter archives for current update information. Then, describe the structures available in the database and the numbers in each class (e.g., B-DNA, ribozymes). Explore the different features of NDB. There are many exciting structures and utilities.

3. **Sugar Conformations for Canonical A, B, and Z-DNA**. To appreciate the different features of canonical A, B, and Z-DNA forms, choose through the NDB entries one unmodified form in each DNA class with the largest number of residues possible. You can view the structure within NDB, or by porting the PDB file to Insight II and using the Molecule / Get from the **Viewer** module (though the latter may be more difficult).
   Learn how to use the Report facility in the NDB Query Interface to generate a list of all the sugar pseudorotation angles (P) for each deoxyribose in each of the three structures you have chosen for your study. (Remember that there are two sugars per base pair). Print a table for each structure.
   Now, on one figure for all the three structures, plot P versus the residue number. Exclude the two terminal base pairs of each structure for plotting purposes. You may need to connect the points corresponding to each structure for clarity.
   Label the pattern clearly to indicate the A, B, and Z-DNA data. Hand in this plot, with a description of the patterns you had identified for the sugar conformation for the three canonical DNA forms, indicating the specific structures you have chosen.

Summary of Items to Hand In:

(a) Ramachandran plots for the two subgroups of amino acid residues assigned to you.
(b) Data on the structure class types and the amount of structures in each class available in NDB.
(c) The figure with P versus residue number for the A, B, and Z-DNA forms, with complete discussion.

Background Reading from Coursepack

- B. Cipra, "Computer Science Discovers DNA", in *What's Happening in the Mathematical Sciences*, pp. 26–37 (P. Zorn, Ed.), American Mathematical Society, Colonial Printing, Cranston, RI (1996).
- M. Gerstein and M. Levitt, "Simulating Water and the Molecules of Life", *Sci. Amer.* **279**, 101–105 (1998).

- J. E. Cohen, "Mathematics is Biology's Next Microscope, Only Better; Biology is Mathematics' Next Physics, Only Better", *PLoS Biology* **2 (e439)**, 2017–2023 (2004).
- A. Hastings et al., "Quantitative Bioscience for the 21st Century", *Bioscience* **55**, 511–517 (2005).

Printing Instructions

To save ink, change the background color from black to white at each printing of a color or black/white image. Go to Session / Environment, press the Background button and change the color to white.

Proceed to File, choose Export_Plot, postscript, Gray_Scale and optionally Ball_and_Stick. Use Save_Device_File to save a postscript file (the file will have the ".ps" extension). This file can be printed on any postscript printer you can access. Hand in your printout as part of the homework.

Fortran **program for selecting Ala, Ile, Val, and Leu data**
(see website)

Assignment 5: Analysis of Protein/DNA Complexes with Insight and NDB: Canonical vs. Protein/Bound DNA, and DNA/Protein Interactions

In this assignment, we will study three nucleic acids structures, two of which have been crystallized with regulatory proteins: a nucleic acid dodecamer, a DNA oligomer bound to a prokaryotic protein in the helix-turn-helix (HTH) motif, and a DNA oligomer (known as the TATA-box sequence) bound to a eukaryotic transcription factor.

NDB ID code	Structure
• **BDL078**	DNA dodecamer
• **PDR010**	DNA (20 bp) bound to bacteriophage λ cI repressor
• **PDT034**	DNA (16 bp) bound to human TATA-Box Binding protein

1. **Structure Downloading**. Download each structure from the NDB, already explored in Assignment 4 (ndbserver.rutgers.edu:80/). Use the Archives, Atlas or Search entry points to the NDB; all are useful. In particular, the Atlas allows you to quickly view the structures. The Search entry point will be utilized later in this assignment.
   Load each structure into Insight II. Separate the two complexes into DNA and Protein objects using the Modify / Unmerge command from either the **Biopolymer** or **Builder** modules. Both the DNA and Protein objects will be used later in this assignment. Use caution with the wildcard character * in selecting multiple atoms/residues/nucleotides/ proteins in Insight II. A quick test which you can used to test which object you've created with the Modify / Unmerge is to blank or blink the object with the Object / Blank or Object / Blink commands.
   In order to create the B-DNA models for specified nucleic acid sequences (see below) and manipulate the downloaded structures (as in the Unmerge command), you will need to understand certain general parameters related to the structure as used in both the PDB file and Insight. These parameters refer to the DNA sequence, strand and residue labels, and so on. Explore the text in the downloaded files as well as in the structural displays. Information about relevant formatted lines in the coordinate files (such as SEQRES and ATOM) is available from the PDB web site (see Assignment 2).

2. **Canonical vs. Protein/Bound DNA Analysis**. Create an idealized B-DNA structure using the Nucleotide / Append command in the **Biopolymer** module corresponding to each of the three nucleic acid sequences in the above structures. The Nucleotide / Append command both creates new nucleic acid molecules and appends to existing molecules; when it is creat-

ing a new molecule, the Append Point is None.[3] You will need to use the Nucleic_Acid / Cap pulldown menu to replace the phosphate group with a hydroxyl group at the 5′ ends of each strand.
Nucleotide / Append will create your B-DNA model, defining each strand as a separate object. Your downloaded DNAs, however, are each composed of a single object in Insight II. To properly superimpose the two structures (the task in the next part), you will need to Modify / Merge the two strands as one object.
Superimpose each B-DNA model upon its respective crystal structure from the NDB using the Transform / Superimpose pulldown menu. If your DNAs are each composed of a single object in Insight II, you will need to Modify / Merge the two strands as one object to properly superimpose the two structures. Use the Heavy option to avoid superpositioning of hydrogen atoms. After selecting the B-DNA model and the crystal structure in the selection boxes, you will need to click the End Definition box to Execute the Superimpose command. The root-mean-square deviation (RMSD) value will be printed at the bottom of your screen.

(a) Record the RMSD values relative to idealized B-DNA for each superimposed model/structure by repeating this procedure for each crystal structure.

(b) Use the Search entry point to the NDB to extract the following parameters for each base pair of the three structures: *P* (pseudorotation sugar pucker); the dihedral angles χ, α, β, γ, δ, ϵ, ζ; and the helical parameters twist, tilt, roll, and propeller twist: Ω, τ, ρ, and ω, respectively.
For each conformational variable (excluding those from the end residues), calculate the average (μ) and standard deviation (σ) from the data per structure. Prepare a table in the following form:
Now discuss your results in terms of the differences noted between the protein-bound DNA and canonical B-DNA (which the BDL078 structure represents):

(c) Which structure is most deformed from B-DNA? Which parameters display the largest changes from B-DNA (consider both μ and σ)? Based on these parameters, what is similar in the way the two complexes deform their recognition sites away from B-DNA (look for

[3]The **PDR010** structure has an overhanging base at each end, that is a base without a Watson-Crick partner on the other strand. (This procedure promotes crystal formation.) The recommended procedure for creating the overhanging base is to use Nucleotide / Append to create a 21-base-pair duplex of the correct sequence and then use Nucleotide / Delete command to delete one base from each strand prior to using Nucleic_Acid / Cap.

	BDL078 μ σ	**PDR010** μ σ	**PDT034** μ σ
P			
χ			
α			
β			
γ			
δ			
ϵ			
ζ			
Ω			
τ			
ρ			
ω			
RMSD			

similarities in columns 3 and 4 above which are different from column 2)?

(d) Are any of the changes observed localized to particular regions in the DNA? (Consider properties with μ values similar to B-DNA but large associated σ values). Plot one of these parameters as a function of position (base pair) along the DNA.

(e) Generate a *side-by-side* picture of the three DNA structures. A recommended utility for this is the File / Export_Plot facility.

3. **Analysis of Interface Between Proteins and DNA**. Next, we will examine some of the interactions formed at the interface between the regulatory proteins and their DNA binding sites. Load the PDB files of each DNA/protein complex in turn and unmerge the DNA part (but leave the DNA and protein together in space).

   The main tool used here is the Subset / Interface pulldown in the central **Viewer** module. This menu allows subsets to be defined in one molecule that satisfy a certain spatial relationship with respect to the other molecule. For example, we would like to use this menu to define subsets of atoms in the protein that are near functional groups of the DNA. A contour level of 3.5 Å is useful in this menu for defining interactions between non-hydrogen atoms, since it roughly corresponds to distances for strong interactions.

   Open the Subset / Interface pulldown menu. You can define the Subset Name as you please. You can define the Center of Subset to be a specific functional group in DNA. For example, DNA:T:C5M refers to atom C5M of thymine's methyl group in the DNA. Define the Search_Domain to be the protein. The Radius of Subset should be set to the value 3.5.

   For your reference, these are names of some DNA atoms:

- Phosphate groups: Atoms P, O1P, O2P
- Thymine methyl groups: Atom C5M
- Adenine amino groups: Atom N6
- Pyrimidine carbonyls: Atom O2
- Purine amines: Atom N3

(a) Save listings of these subsets into output files. Subset / List is recommended for this task.

> [Do not be alarmed if you get the error message "Invalid Comparison Object"; it simply means that the comparison could not be performed since no member of the set fulfilled the criteria. If an attempt is made to analyze all atoms B that are 3.5 Å from protein A, but all atoms of protein A are more than 3.51 Å from B, this error will occur.]

Combine the analyses of the two complexes (if you can) and construct *histograms* of the residue types in each subset. That is, from the listings of all residues within 3.5 Å of the atom groups above, count the number of times each residue appears (e.g., Methionine appears 60 times, Glutamate 3) and generate histogram plots as illustrated below; use the one-letter mnemonic for the amino acids. (You may also want to count the frequencies of residues grouped by type, like polar, hydrophobic, charged, etc.).

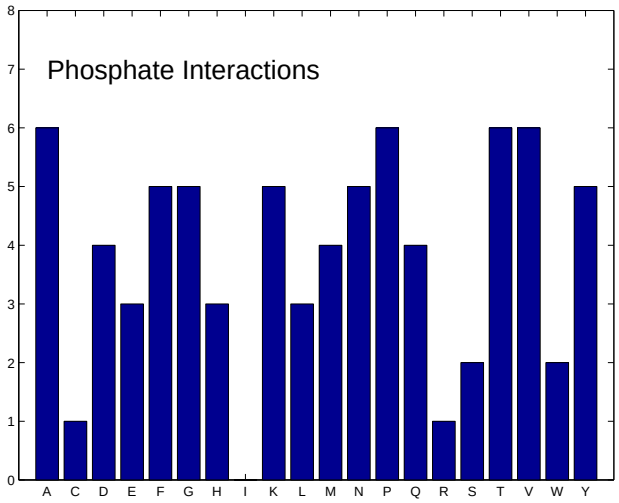

Figure D.1. Sample histogram for protein/DNA interaction analysis

(b) Do you observe common patterns in the two complexes? Are certain amino acids likely to be found interacting with a particular functional group? What types of interactions are being formed between these nu-

cleic acid functional groups and the regulatory protein (e.g. attribute to each of the nucleic acid groups above a type of interaction such as hydrophobic, hydrogen bonding, electrostatic, intercalation/insertion motif, etc.)?

(c) Is there anything unusual about the subsets formed between the proteins and the O2 carbonyl/N3 amine atoms?

(d) Is there any relation between the interactions observed in these subsets and the deviations from canonical B-DNA structure observed above (i.e. how do the interactions you observe explain any of the parameter variances you diagnosed)?

Note: A trick to identify atoms/residues is via Molecule / Color *for assigning a color to an atom/residue. Other labeling tools such as* Molecule / Render *and* Molecule / Label *can similarly be used.*

(e) **Bonus Question:**[4] **BDL078 Homologues.** We have used the BDL078 structure as an example of B-DNA in the analysis above. However, sequence-dependent variations in local structure are also important. Therefore, a more sensitive analysis of free versus protein/bound DNA employs the same nucleotide sequence, both with and without bound proteins. There are few cases, however, in which high-resolution DNA structures are available in both the free and protein-bound states. Such analyses are illuminating and show how intrinsic DNA preferences are amplified in the DNA/protein complexes. See recent reports regarding the complex between DNA and the bovine papillomavirus E2 protein in D.M. Crothers (*Proc. Natl. Acad. Sci.* **95**:15163–15165, 1998) and H. Rozenberg *et al.* (*Proc. Natl. Acad. Sci.* **95**:15194–15199, 1998).

Such analyses have not been done with our BDL078 sequence, but there are protein-DNA complexes in the NDB which are closely related to BDL078. This close relationship means that: (*i*) the related sequence has many similar or closely related residues to BDL078 (e.g., GG**G**A**A**AA**T**TT is closely related to GG**C**A**T**AA**C**TT), and (*ii)* the protein would bind to BDL078 and this related sequence.

Determine which protein/DNA complexes these are, and briefly describe what these complexed proteins do.

Background Reading from Coursepack

- K. B. Lipkowitz, "Abuses of Molecular Mechanics. Pitfalls to Avoid", *J. Chem. Educ.* **72**, 1070–1075 (1995).

[4]The correct solution will allow you to drop lowest homework grade in any assignment.

- S. Lifson, "Potential Energy Functions for Structural Molecular Biology", in *Methods in Structural Molecular Biology*, pp. 359–385, D. B. Davies, W. Saenger, and S. S. Danyluk, Eds., Plenum Press, London (1981).

Assignment 6: MIDTERM: Homology Contest! Exploring Sequence/Structure/Function Relationships (& Related Tools/Databases like SCOP, IMAGE, BLAST, NDB, PDB)

With the rapidly growing information on genomic sequences, *comparative modeling* — structure prediction based on sequence similarity — is becoming increasingly valuable. Indeed, structural and functional genomics, the three-dimensional (3D) structure and functional analysis of genomic products, are rising disciplines in bioinformatics. It has been reported, for example, that a sequence homology of larger than 40% usually implies more than 90% 3D-structure overlap (see below for precise definitions of similarity). Thus, with the growing amount of genomic information, we may eventually be able to predict reliably 3D structures of proteins. Since structural similarity is often preserved more strongly than sequence through evolution, reliable homology-based predictions might provide crucial functional properties of new gene products in the near future.

Through this assignment, you will gain some experience in quantifying and analyzing sequence and 3D structure similarity for proteins. You will also explore sequence and structure databases in search of interesting examples, and learn how to use important computational and database resources. You will have to be resourceful in looking for suitable programs for alignment and structure analysis besides those below; no simple recipes will be given here.

This assignment can be done by teams of two students; choose a partner with complementary skills. You will have to present your results to the class.

The 5 Tasks

Find and demonstrate the following four relationships for proteins:

1. [**EASY**] Two proteins with very *high sequence similarity* (but less than 95%) and very *high structural similarity*. Excluded from consideration are trivial examples, such as involving multiple PDB entries for the same protein.

2. [**EASY**] Two proteins with very *high sequence similarity* (but less than 95%) and *very high structural similarity* but markedly *different biological/functional* properties.

3. [**MODERATE**] Two proteins with *low sequence similarity* but *high structural similarity*. Also comment on the *functional* properties of the pair.

4. [**HARD**] Two proteins with very *high sequence similarity* but very *low structural similarity*. Also comment on the *functional* properties in your example.

> For problems 3 and 4 above, the class contest will be won by the students that find the most extreme examples (i.e., the maximal sequence similarity / minimal

structural similarity, minimal sequence similarity / maximal structural similarity).

5. [**EASY WARMUP**] Search and identify all the determined structures in the PDB/NDB that contain the nucleic acid sequence TATAAAAG. Discuss these structures and their significance.

For each task, generate color molecular views, report the analyses in detail, and include a description of how you found the example. Also discuss your similarity/dissimilarity criteria (see below), and prepare a class presentation on your results.

Ground Rules

1. Homology, or sequence similarity, will be defined by the percentage of sequence identity.
2. 3D-structure similarity will be defined in two ways:
   (a) the percentage of C^{α} atoms of the proteins that "overlap", i.e., are within 3.5 Å of each other in a rigid-body alignment of the protein;
   (b) the root-mean-square-deviation (RMSD) between C^{α} atoms of the proteins in a rigid-body alignment of the protein. (Recall your experience with RMSD measurements in the previous assignment).

   You should first experiment with overlapping several protein structures to determine what RMSD values and/or percentages of C^{α} overlap indicate random similarity. *Discuss this in your submission.*

Tools of the Trade

1. **Sequence and Structure Databases**. You have already navigated through the structural PDB and NDB databases and various sequence databases. Continue to work with these and the RCSB facilities.
2. **SCOP**. This site for the *Structural Classification of Proteins* (scop.mrc-lmb.cam.ac.uk/scop/ categorizes proteins according to the levels (top-to-bottom) of: class, fold, superfamily, family, domain, and reference PDB structure.
3. **Insight II**. Continue to use Insight II for structure display and analysis.
4. **NCBI Tools like BLAST and Its Cousins**. BLAST is a library of heuristic similarity search programs (Basic Local Alignment Search Tools) that explore relationships involving protein and nucleic-acid sequences and 3D structures. This library contains blastp, blastn, blastx, tblastn, tblastx,

and others, developed at the National Center for Biotechnology Information at the National Library of Medicine of the National Institutes of Health. Get started at their web site www.ncbi.nlm.nih.gov/BLAST/. This page leads to the BLAST suites as well as contains usage information. See, for example, Overview, Manual, BLAST FAQs, References.

BLAST, one of the most popular tools among molecular biology researchers, has evolved rapidly since its inauguration in 1990. BLAST searches a database in two stages, finding small sequence lengths that match the target exactly and then attempting to extend the length of the match from this subset of sequences in the database. Not only are the alignment algorithms improving continuously (e.g., allowing alignments of DNA or protein sequences with insertions or deletions in Gapped BLAST; forming families of aligned sequences and quick profiles of them in Position-Specific Iterated (PSI)-BLAST; or incorporating biological-function hypotheses into sequence queries to restrict the analysis to subset of protein sequences as in Pattern-Hit Initiated (PHI)-BLAST), but performance has been greatly accelerated. Algorithmic features include dynamic programming tools, hidden Markov models, and various optimization strategies.

To align two protein or nucleotide sequences, go to the link of BLAST 2 sequences (www.ncbi.nlm.nih.gov/gorf/bl2.html) and set up the computation according to the instructions. Take care to choose the options of the computation with care, and explore different options. The server will send the results to the web browser being used.

Some available programs are:

- blastp: compares an amino acid query sequence against a protein sequence database.
- blastn: compares a nucleotide query sequence against a nucleotide sequence database.
- blastx: compares the six-frame conceptual translation products of a nucleotide query sequence (both strands) against a protein sequence database.
- tblastn: compares a protein query sequence against a nucleotide sequence database dynamically translated in all six reading frames (both strands).
- tblastx: compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database.

See www.ncbi.nlm.nih.gov/BLAST/newblast.html#introduction for further information.

Other similarity programs are available (such as MEME *and* MAST *from SDSC); use anything appropriate for the task.*

5. **An Image Library**. The *Image Library of Biological Macromolecules* organized by the Institute for Molecular Biotechnology in Jena, Germany (www.imb-jena.de/IMAGE.html) offers a colorful library of biomolecular images corresponding to structures available in databases like the NDB and PDB. Besides detailed colorful illustrations of the structure in a variety

of styles, relevant structural information and publication links are available. Basic tutorials on structural biology are under preparation at this site.

HINTS for the Assignment

1. Scan the literature for related papers on comparative or homology modeling but do not repeat known examples. You **CAN** be original.
2. Large changes in 3D structure despite high sequence similarity can result from the following situations:
   - mutations in critical regions of the proteins such as active sites
   - mutations in ligand binding sites (as in immunoglobulins)
   - mutations in regions that connect two secondary-structural elements (as in helix-loop-helix motifs)
   - structure determination of the same system at different environmental conditions (e.g., different solvent, different crystal packing forms for mutant proteins)
   - proteins containing the same subunits but a different number of subunits, with a structure/fold/topology that depends critically on that number.

   Search PDB and SCOP for examples in this spirit.
3. Look for groups of proteins in the same family, or for proteins sharing the same fold in the SCOP site. The structural classification information should generate ideas.
4. General structure alignment via Insight is not very sophisticated and may be entirely unsuitable for sequences of disparate lengths and for structures with two similar subdomains adopting a different relative orientation. Search for suitable programs for these cases (e.g., from the RCSB, home.rcsb.org and from SDSC) and also write/use your own programs to perform certain analyses, such as structure similarity measurements upon alignment (e.g., criterion 2a under **Ground Rules**).

Background Reading

- D. Baker and A. Sali, "Protein Structure Prediction and Structural Genomics" *Science* **294**, 93–96 (2001). [**From Coursepack**].
- J. C. Whisstock and A. M. Lesk, "Prediction of Protein Function from Protein Sequence and Structure", *Quart. Rev. Biophys.* **36**, 173–189 (2001). [**From Coursepack**].
- J.-M. Chandonia and S. E. Brenner, "The Impact of Structural Genomics: Expectations and Outcomes", *Science* **311**, 347–351 (2006). [**From Coursepack**].

- B. Honig and A. Nicholls, "Classical Electrostatics in Biology and Chemistry", *Science* **268**, 1144–1149 (1995) [**From Coursepack**].
- D. Case, "NMR Refinement", in P. von Ragué Schleyer (Editor-in Chief), N. L. Allinger, T. Clark, J. Gasteiger, P. A. Kollman, and H. F. Schaefer, III, editors, *Encyclopedia of Computational Chemistry*, volume 3, pages 1866–1876. John Wiley & Sons, West Sussex, England, 1998.

Assignment 7: Molecular Mechanics Force Fields: Approximations, Variations, and the Assessment of Results with respect to Experiment and other Simulations

1. **Reading**. This assignment deals with the series of four articles below, which raise both general and specific problems in biomolecular simulations. At issue is the validation of conformational predictions by various molecular mechanics force fields. You may also wish to refer to the Lipkowitz article from Assignment 5 (on the pitfalls of molecular mechanics) and the van Gunsteren and Mark article from Assignment 1 (on validating molecular dynamics simulations). Begin by reading these papers (included in the Coursepack, see Appendix B) and thinking about the modeling issues as you read them.
   - I. K. Roterman, M. H. Lambert, K. D. Gibson, and H. A. Scheraga, "Comparison of the CHARMM, AMBER and ECEPP Potentials for Peptides. I. Conformational Predictions for the Tandemly Repeated Peptide $(\text{Asn-Ala-Asn-Pro})_9$", *J. Biomol. Struct. Dyn.* **7**, 391–419 (1989a).
   - I. K. Roterman, M. H. Lambert, K. D. Gibson, and H. A. Scheraga, "Comparison of the CHARMM, AMBER and ECEPP Potentials for Peptides. II. ϕ–ψ Maps for N′-Methyl Amide: Comparisons, Contrasts and Simple Experimental Tests", *J. Biomol. Struct. Dyn.* **7**, 421–453 (1989b).
   - P. A. Kollman and K. A. Dill, "Decisions in Force Field Development: An Alternative to Those Described by Roterman *et al.*", *J. Biomol. Struct. Dyn.* **8**, 1103–1107 (1991).
   - K. B. Gibson and H. A. Scheraga", "Decisions in Force Field Development: Reply to Kollman and Dill", *J. Biomol. Struct. Dyn.* **8**, 1109–1111 (1991).

2. **Preparation for Class Discussion.** You will be divided into three groups (assignments will be given in class): (1) the moderators, (2) the ECEPP group, and (3) the AMBER and CHARMM group. Each group will have to prepare material, as described below, for class presentation and discussion. *All materials should be prepared on overhead projector slides.* You should meet with your group members in advance to plan your presentation and debate strategies.

   The *moderators* will be in charge of presenting in detail the *facts*: what studies were performed, what questions were asked, and what analyses were made. You should be prepared to answer any background questions (e.g., definitions of polymer quantities analyzed).

   The *ECEPP* group will endorse the point of view taken by Roterman, Gibson, Scheraga, and co-workers. Besides understanding your po-

sition well, you will need to bring to the debate *concrete examples from the literature* to support your position. Be creative and try to find interesting examples.

The *AMBER* folks and *CHARMMers* will endorse the approach taken in these two molecular packages and, in particular, the point of view taken by Kollman and Dill in their reply to Roterman *et al.*. As above, besides understanding well your molecular mechanics packages and position taken in the reply, you will need to bring to the debate *concrete examples from the literature* to support your position. Be creative in your supporting materials and strategies.

3. **Useful Recommendations**. Summarize in brief the useful recommendations and comments that emerged from all the above articles, as well as additional ones, for practitioners of molecular modeling. That is, propose *concrete procedures* that biomolecular simulators can use to gain as much confidence as possible in their conclusions and predictions.

   Remember, uncertainties and approximations in numerical modeling and simulations will always exist! The field of modeling biomolecules on modern computers involves as much art as science. But despite their obvious limitations, modeling methodologies are improving continuously. The goal of every practitioner should be to realize the highest possible accuracy as is compatible with the model and methods utilized. Like any calculation, 'error bars' in the broad sense should be attributed to the results and conclusions claimed.

4. **Points to Keep in Mind**. Throughout this assignment, think about the following important issues in molecular modeling:
   - Accuracy versus approximation
   - Theory versus experiment
   - Dependence of simulation results on the protocols used
     - starting configuration
     - model assumptions
     - force field
     - algorithms (minimization, adiabatic mapping, etc.)
   - Assessment of Results:
     - How can you distinguish between bona fide physical *trends* and numerical *artifacts*?
     - How can you decide whether the model is wrong (energy, assumptions, etc.) or the method is inappropriate?
     - What are appropriate comparisons with experimental results?

Summary of Items to Hand in:

(a) Brief description of the issues raised in the four articles regarding molecular mechanics predictions.
(b) Your work in preparation of the class debate.
(c) Proposals of procedures to be used to attain the maximum possible confidence from biomolecular simulations.

Have Fun!

Background Reading from Coursepack

- J. Skolnick and A. Kolinski, "Simulations of the Folding of a Globular Protein", *Science* **250**, 1121–1125 (1990).
- F. M. Richards, "The Protein Folding Problem", *Sci. Amer.* **264**, 54–63 (1991).
- H. A. Scheraga, "Predicting Three-Dimensional Structures of Oligopeptides", in Reviews in Computational Chemistry, K. B. Lipkowitz and D. B. Boyd, Editors, Vol. 3, pp. 73–142, VCH Publishers, New York (1992).
- A. Neumaier, "Molecular Modeling of Proteins and Mathematical Prediction of Protein Structure", *SIAM Review* **39**, 407–460 (1997).

Background Reading for Scheraga's Lecture

- J. Pillardy, Y. A. Arnautova, C. Czaplewski, K. D. Gibson, and H. A. Scheraga, "Conformation-Family Monte Carlo: A New Method for Crystal Structure Prediction", *Proc. Natl. Acad. Sci., USA* **98**, 12351–12356 (2001).
- J. Pillardy, C. Czaplewski, A. Liwo, W. J. Wedemeyer, J. Lee, D. R. Ripoll, P. Arlukowicz, S. Oldziej, Y. A. Arnautova and H. A. Scheraga, "Development of Physics-Based Energy Functions that Predict Medium-Resolution Structures for Protein of the α, β, and α/β Structural Classes", *J. Phys. Chem. B* **105**, 7299–7311 (2001).
- J. Lee, D. R. Ripoll, C. Czaplewski, J. Pillardy, W. J. Wedemeyer and H. A. Scheraga, "Optimization of Parameters in Macromolecular Potential Energy Functions by Conformational Space Annealing", *J. Phys. Chem. B* **105**, 7291–7298 (2001).
- A. Liwo, C. Czaplewski, J. Pillardy and H. A. Scheraga, "Cumulant-Based Expressions for the Multibody Terms for the Correlation Between Local and Electrostatic Interactions in the United-Residue Force Field", *J. Chem. Phys.* **115**, 2323–2347 (2001).

- J. Pillardy, C. Czaplewski, A. Liwo, J. Lee, D. R. Ripoll, R. Kazmierkiewicz, S. Oldziej, W. J. Wedemeyer, K. D. Gibson, Y. A. Arnautova, J. Saunders, Y.-J. Ye and H. A. Scheraga, "Recent Improvements in Prediction of Protein Structure by Global Optimization of a Potential Energy Function", *Proc. Natl. Acad. Sci., USA* **98**, 2329–2333 (2001).
- J. Pillardy, C. Czaplewski, W. J. Wedemeyer and H. A. Scheraga, "Conformation-Family Monte Carlo (CFMC): An Efficient Computational Method for Identifying the Low-Energy States of a Macromolecule", *Helv. Chim. Acta* **83**, 2214–2230 (2000).
- J. Lee, J. Pillardy, C. Czaplewski, Y. Arnautova, D. R. Ripoll, A. Liwo, K. D. Gibson, R. J. Wawak, and H. A. Scheraga, "Efficient Parallel Algorithms in Global Optimization of Potential Energy Functions", *Comput. Physics Commun.* **128**, 399–411 (2000).
- J. Lee, A. Liwo, D. R. Ripoll, J. Pillardy, J. A. Saunders, K. D. Gibson and H. A. Scheraga, "Hierarchical Energy-Based Approach to Protein-Structure Prediction: Blind-Test Evaluation with CASP3 Targets", *Intl. J. Quantum Chem.* **71**, 90–117 (2000).
- J. Pillardy, R. J. Wawak, Y. A. Arnautova, C. Czaplewski, and H. A. Scheraga, "Crystal Structure Prediction by Global Optimization as a Tool for Evaluating Potentials: Role of the Dipole Moment Correction Term in Successful Predictions", *J. Am. Chem. Soc.* **122**, 907–921 (2000).
- H. A. Scheraga, J. Lee, J. Pillardy, Y.-J. Ye, A. Liwo, and D. R. Ripoll, "Surmounting the Multiple-Minima Problem in Protein Folding", *J. Global Optimization* **15**, 235–260 (1999).
- J. Lee, A. Liwo and H. A. Scheraga, "Energy-Based *de novo* Protein Folding by Conformational Space Annealing and an Off-lattice United-Residue Force Field: Application to the 10-55 Fragment of Staphylococcal Protein A and to apo calbindin D9K", *Proc. Natl. Acad. Sci., USA* **96**, 2025–2030 (1999).
- J. Lee, H. A. Scheraga and S. Rackovsky, "Conformational Analysis of The 20-Residue Membrane-Bound Portion of Melittin by Conformational Space Annealing", *Biopolymers* **46**, 103–115 (1998).
- R. J. Wawak, J. Pillardy, A. Liwo, K.D. Gibson and H. A. Scheraga, "Diffusion Equation and Distance Scaling Methods of Global Optimization: Applications to Crystal Structure Prediction", *J. Phys. Chem.* **102**, 2904–2918 (1998).
- A. Liwo, R. Kazmierkiewicz, C. Czaplewski, M. Groth, S. Oldziej, R. J. Wawak, S. Rackovsky, M. R. Pincus, and H. A. Scheraga, "United-Residue Force Field for Off-Lattice Protein-Structure Simulations; III. Origin of Backbone Hydrogen-Bonding Cooperativity in United-Residue Potentials", *J. Comput. Chem.* **19**, 259–276 (1998).

Assignment 8: A Bit of Programming: Nonbonded Versus Bonded Energy Computations

This is a small programming assignment. At least one assignment in this modeling course should give you such first-hand experience! If you are a novice in programming, NYU staff, the TA, and the course assistant can help you, so set up an appointment with them early. You will have *two weeks* for this assignment.

1. **Programming Nonbonded Energy Computations**.
   We will begin with the nonbonded energy computations since they are most straightforward (but most expensive!)

   Write a simple program to compute the nonbonded energy of a system of 1000 atoms. The nonbonded energy, Lennard Jones and Coulomb terms, should have the form:

$$E_{NONB} = \sum_{i<j} \left[\frac{-A_{ij}}{R_{ij}^3} + \frac{B_{ij}}{R_{ij}^6} + \frac{q_i q_j}{\sqrt{R_{ij}}} \right]. \qquad \text{(D.1)}$$

   Here R_{ij} is an interatomic distance *squared*, and A_{ij}, B_{ij}, q_i, and q_j are the familiar energy parameters.
   For an atom $\mathbf{x}_k$ of Cartesian components $\{x_{k1}, x_{k2}, x_{k3}\}$,

$$R_{ij} = (x_{j1} - x_{i1})^2 + (x_{j2} - x_{i2})^2 + (x_{j3} - x_{i3})^2 , \qquad \text{(D.2)}$$

   and the interatomic distance r_{ij} is

$$r_{ij} = \sqrt{R_{ij}}.$$

   Set up your program to read in arbitrary atomic coordinate data — a file will be sent to you electronically in case you want to use it[5] — and *repeat the nonbonded energy calculation* 10,000 times for all $\{i, j\}$ pairs with $i < j$ for $j = 1, \ldots 1000$. The R_{ij} calculations can be placed in some inline function.
   Perform the calculations for the nonbonded energy evaluations for both single and double precision, and record the total CPU time in each case.
   Also report how much CPU time and CPU percentage the square-root ($\sqrt{\ }$) operation consumes. There are special timing functions that describe the distribution of CPU time among the various program parts.
   Describe the machine you are using, the precision, and attach the subroutine and program output, along with the above results.

[5] The file can be obtained through the link to the course web site or directly from the author.

2. **Programming the Bonded Energy Computations**.
   Next we will write three additional subprograms to compute the bond energy, bond-angle energy, and dihedral-angle energy of a molecular system. For now, we will assume there are 1000 bonds, 1000 bond angles, and 1000 dihedral angles. We will again perform 10,000 energy evaluations of each energy term, with each sweep here involving 1000 internal variables. This large number is necessary to get reliable timing values for the bonded interactions.

   You can choose in each case any representative potential form. For example, you may use:

$$E_{BOND} = \sum_{i,j \in S_B} S_{ij} \big(r_{ij} - \bar{r}_{ij}\big)^2, \tag{D.3}$$

$$E_{BANG} = \sum_{i,j,k \in S_{BA}} K_{ijk} (\cos\theta_{ijk} - \cos\bar{\theta}_{ijk})^2, \tag{D.4}$$

$$E_{TOR} = \sum_{ijk\ell \in S_{DA}} \left(\frac{V3_{ijk\ell}}{2} \big[1 + \cos(3\tau_{ijk\ell})\big] \right), \tag{D.5}$$

   where the sets S_B, S_{BA}, and S_{DA} contain all bonds, bond angles, and dihedral angles, respectively. Here θ and τ denote a bond angle and dihedral angle, respectively, of a given triplet or quadruplet of atoms. The values with overhead bar symbols indicate reference values.

   Since we are interested only in timing for now, you can use any pairs, triplets, or quadruplets in your sample energy routines — even the same sequence — repeatedly, as long as the total number of interactions used to obtain each energy term is 1000.

   Some program segments which you may find helpful are posted on the website. The derivative components are present in the angle routines, *but you do not need them* for this assignment. You can find details of the cosba and cosda subroutines in an article.[6]

   For your convenience, an addendum to this assignment also summarizes the basic geometric relations involved in defining internal variables.

   Report the CPU time required for each routine in a table, including absolute time as well as percentage of the total time of *bonded* energy components. Again, attach your programs and output to the report of the results.

[6] "A Recipe for Evaluating and Differentiating $cos\phi$ Expressions", *J. Comp. Chem.* **10**, 951–956, 1989.

3. **Setting up A Polymer Model**. For obtaining realistic CPU estimates, we will now consider a simple n-alkane chain with the chemical formula CH_3–$(CH_2)_m$–CH_3, where m is an integer. For large m, this is polyethylene. For $m = 2$, for example, we have butane, chemical formula C_4H_{10}. To have about 1000 atoms, we will use $m = 330$ for our model calculations.

   Determine the number of bonds, bond angles, dihedral angles, and unique interatomic distances (atom pairs) that polyethylene has as a function of m. Consider all the distinct possibilities for the bonds and angles. Report these expressions.

   Then report how many bonds, bond angles, dihedral angles, and unique atom pairs the polymer has for the case $m = 330$.

4. **Bonded Versus Nonbonded Energy Computations**.
   Now we will combine the timing above to estimate the CPU time spent in bonded versus nonbonded energy computations for 10,000 iterations (of energy evaluations) for our polymer of 998 atoms.
   Scale the timing you obtained above (10,000 iterations for 1000 atoms for the nonbonded terms, and 10,000 iterations for 1000 bonds, bond angles, and dihedral angles) so that they correspond to the numbers relevant for our polymer with $m = 330$, as determined in item 3 above.

   Collect the data in one table which reports the CPU time and percentage required for each of the four subroutines.

   What can you conclude? What can you suggest to speed up the nonbonded computations, especially if derivatives are also required?

5. **Extra Credit!**
   For extra credit (the grade on this will replace your lowest homework grade), write the four subroutines above specifically for polyethylene. This means that you should use realistic coordinates, as well as correct data structures so that you consider all relevant bonds and angles for this polymer. Similarly, for energy parameters, associate values according to atom, bond, and angle types (e.g., C–C and C–H bonds, C–C–C, H–C–H, and H–C–C bond angles, and rotations about C–C bonds). You can use any resources on Insight to help you.

   Hand in all programs and results as requested above.

Addendum to Assignment 8:
Definitions of Internal Variables in Molecules

A bond angle θ_{ijk} formed by a bonded triplet of atoms i–j–k is expressed as an inner product:

$$\cos\theta_{ijk} = \frac{(\mathbf{x}_k - \mathbf{x}_j) \bullet (\mathbf{x}_i - \mathbf{x}_j)}{r_{jk}\ r_{ji}}\ , \tag{D.6}$$

or

$$\cos\theta_{ijk} = (\mathbf{r}_{jk} \bullet \mathbf{r}_{ji})/r_{jk}\ r_{ji}\ ,$$

where the distance *vector* from atom j to i is given by

$$\mathbf{r}_{ji}\ =\ \mathbf{x}_i - \mathbf{x}_j\ =\ [x_{i1} - x_{j1},\ \ x_{i2} - x_{j2},\ \ x_{i3} - x_{j3}]^{\mathrm{T}}. \tag{D.7}$$

A dihedral angle $\tau_{ijk\ell}$, defining the rotation of bond i–j about bond j–k with respect to k–l, is expressed as

$$\cos\tau_{ijk\ell} = \mathbf{n}_{ab} \bullet \mathbf{n}_{bc}. \tag{D.8}$$

The vectors $\mathbf{n}_{ab}$ and $\mathbf{n}_{bc}$ denote unit normals to planes spanned by vectors $\{\mathbf{a}\ ,\mathbf{b}\}$ and $\{\mathbf{b}\ ,\mathbf{c}\}$, respectively, where $\mathbf{a} = \mathbf{r}_{ij}$, $\mathbf{b} = \mathbf{r}_{jk}$, and $\mathbf{c} = \mathbf{r}_{k\ell}$. Denoting θ_{ab} and θ_{bc} as angles θ_{ijk} and $\theta_{jk\ell}$, respectively, we write:

$$\cos\tau_{ijk\ell} = \frac{\mathbf{a}\times\mathbf{b}}{\|\mathbf{a}\|\,\|\mathbf{b}\|\sin\theta_{ab}} \bullet \frac{\mathbf{b}\times\mathbf{c}}{\|\mathbf{b}\|\,\|\mathbf{c}\|\sin\theta_{bc}}\ . \tag{D.9}$$

The sign of $\tau_{ijk\ell}$ is determined by the sign of the triple scalar product $\mathbf{a} \bullet (\mathbf{b} \times \mathbf{c})$.

To simplify potential energy equations (and differentiation when needed) it is convenient to work with inner product expressions and use Lagrange's identity $(\mathbf{a} \times \mathbf{b}) \bullet (\mathbf{c} \times \mathbf{d})$ $= (\mathbf{a} \bullet \mathbf{c})\ (\mathbf{b} \bullet \mathbf{d})\ -(\mathbf{b} \bullet \mathbf{c})(\mathbf{a} \bullet \mathbf{d})$. This produces the alternative expression:

$$\cos\tau_{ijk\ell} = \frac{(\mathbf{a}\times\mathbf{b}) \bullet (\mathbf{b}\times\mathbf{c})}{[\ (\mathbf{a}\times\mathbf{b}) \bullet (\mathbf{a}\times\mathbf{b})\ \ (\mathbf{b}\times\mathbf{c}) \bullet (\mathbf{b}\times\mathbf{c})\ ]^{\,1/2}}$$

$$= \frac{(\mathbf{a} \bullet \mathbf{b})(\mathbf{b} \bullet \mathbf{c}) - (\mathbf{a} \bullet \mathbf{c})(\mathbf{b} \bullet \mathbf{b})}{\left\{\ [\ (\mathbf{a} \bullet \mathbf{a})(\mathbf{b} \bullet \mathbf{b}) - (\mathbf{a} \bullet \mathbf{b})^2\ ]\ \ [\ (\mathbf{b} \bullet \mathbf{b})(\mathbf{c} \bullet \mathbf{c}) - (\mathbf{b} \bullet \mathbf{c})^2\ ]\ \right\}^{\,1/2}}\ . \tag{D.10}$$

According to this convention, $\tau = 0^{\mathrm{o}}$ defines a cis coplanar orientation for atoms i–j–k–l, $\tau = 180^{\mathrm{o}}$ defines a $trans$ coplanar orientation, and a positive sign corresponds to a clockwise rotation of the far bond with respect to the near bond (when viewed along the j–k bond).

(See code segments on website)

Coordinate file (available electronically) for the 1000-atom molecule $CH_2OH–(CH_2)_{330}–CH_2OH$:

Atom	X	Y	Z	ID
1	4.988000	2.012136	-7.818089	O
2	5.915912	2.025234	-7.572307	H
3	4.596542	0.674197	-8.131964	C
4	5.198226	0.305558	-8.962735	H
5	4.751310	0.037135	-7.261160	H
6	3.108016	0.653187	-8.526237	C
7	2.517394	1.015322	-7.710808	H
8	2.956134	1.278341	-9.381230	H
9	2.686321	-0.787809	-8.863891	C
10	2.838205	-1.412964	-8.008898	H
..	.....	.....	.....	.

Etc.

Assignment 9: TERM PROJECT

The Successes (Failures?) of Molecular Modeling

The year is 2006. You have graduated and moved on with your life. Due to your outstanding academic record at NYU, you have landed a high-profile job as a staff research scientist for PBS (Public Broadcasting Service) in the nation's capital.

You are now assigned to prepare for an internationally televised scientific program entitled *Biocomputing in the Third Millennium*. In this program, a team of scientific experts will respond to live questions transmitted by comphones from the general public. Since these scientists are busy traveling, consulting, reviewing papers, writing grants, researching, and teaching, your group is in charge of preparing all background information for the panelists.

Specifically, you are told to prepare for the following questions:

> *Can the panel describe some concrete examples where computational tools have significantly enhanced our understanding of molecular systems — from small organic systems to macromolecules — by offering new insights, interpretations, and predictions, of practical and scientific importance, that were impossible to obtain by experimental techniques?*
>
> *What modeling/simulation tools were used in each case, and what can be credited to each success (computing power, algorithms, intuition, right time, sheer luck, persistence, etc.)?*

You are promised by your boss a hefty bonus for each complete and satisfactory item provided. However, the minimal requirement (for obtaining a B-level mark on your monthly evaluation form, given that you produce truly outstanding examples) is detailing FOUR "SENSATIONAL" EXAMPLES.

Each example must be clearly described and entered under the following subheadings: Problem, Methodology, Success, Significance, References. The second item, Methodology, requires the most comprehensive coverage, followed by Significance. You are asked to attach to your meticulous writeup any visual aids (charts, figures, sketches) that will enhance the presentation, both to a general (nonspecialist) audience and to a highly informed scientist. Creativity is highly desired. Try also to analyze the findings in a larger context.

Back at your ergonomic desk, with your feet up and glancing at regal Washington monuments against a glorious background of blossoming cherry trees with occasional views of ambitious runners and politicians, you recall a molecular modeling course you took in the good ol' days at NYU. Memories come back of many homework assignments inflicted upon you weekly by your professor — dealing with web resources, sequence and structural databases, Insight,

sequence/structure contests, force fields, tedious programming, difficult minimization, and Monte Carlo simulations. You find fragments of lecture notes and transparency copies inside an old purple gym bag and begin to follow up on, and explore, some of those key words, resources, authors, and topics. You also begin to wonder if there are any interesting and instructive examples of *failures in molecular modeling* and decide to pursue those for an extra bonus. (Maybe the boss will let *you* design the next scientific program?)

Your deadline in early May is rapidly approaching and you begin to work early and diligently. The promise that the best examples provided by the crew will be published, if appropriate, in an article provides further motivation for the assignment. You also decide to contact your professor if she is still at NYU when you get stuck or have questions.

You find the project more interesting now, and vow to become *famous* (and maybe even *rich*)!

Background Reading from Coursepack

- M. S. Friedrichs and P. G. Wolynes, "Toward Protein Tertiary Structure Recognition by Means of Associative Memory Hamiltonians", *Science* **246**, 371–373 (1989).
- T. Schlick, "Optimization Methods in Computational Chemistry", in Reviews in Computational Chemistry, K. B. Lipkowitz and D. B. Boyd, Editors, Vol. 3, pp. 1–71, VCH Publishers, New York (1992).
- See also an updated version titled "Geometry Optimization" in the Encyclopedia of Computational Chemistry, P. von Ragué Schleyer (Editor-in-Chief) and N. L. Allinger, T. Clark, J. Gasteiger, P. A. Kollman, and H. F. Schaefer, III, Editors, Vol. 3, pp. 1136–1157, John Wiley & Sons, West Sussex, England (1998).
- R. A. Abagyan and M. M. Totrov, "Biased Probability Monte Carlo Conformational Searches and Electrostatic Calculations for Peptides and Proteins", *J. Mol. Biol.* **235**, 983–1002 (1994).

Assignment 10: Experiments in Molecular Geometry Optimization: Biphenyl Minimization

See Insight II *and* Discover *manuals for reference.*

1. **Brief introduction to the Discover module of** Insight II.

   The Discover[7] software performs energy minimization and molecular dynamics simulations. This program constitutes a powerful modeling tool since it offers many features such as constrained and restrained minimization, calculation of vibrational frequencies, and analysis tools. Many variations of simulation conditions (e.g., constant temperature, constant pressure) are available.

   We will access the Discover software from the Insight II environment. The **Discover** module of Insight II is a convenient interface to the Discover program. This module builds Discover input files from information provided through graphical interfaces, and it allows users to run Discover jobs interactively. Though more advanced users may prefer to use the independent version of Discover, the Insight II environment is more appropriate for a novice.

   Before using Discover, make sure that Insight II contains all of the necessary information to define the topology, coordinates, and force field parameters. These include, for example, atom types and partial charges (see lecture notes for structure definitions).

   If you succeed in displaying the molecule correctly on the screen, the topological and coordinate information is most likely in order. However, selecting the appropriate force field and assigning atom types and parameters is a separate task.

   (a) To select the force field, use Forcefield / Select.
   (b) To assign atom types, use the Fix option for Potential Action in Forcefield / Potentials. Alternatively, first assign atom types with Atom / Potential in the **Biopolymer** module, and then use the Accept option for Potential Action in Forcefield / Potentials.
   (c) To assign charges, use the Fix option for both Partial Chg Action and Formal Chg Action under Forcefield / Potentials.

   Note that after each change in the force field you must assign atom types and charges anew.

[7] Note that the name *Discover* has two separate meanings. The first, Discover, stands for the software package with minimization and molecular dynamics routines. The second, typed in bold (**Discover**), refers to the module available in Insight II.

To check if the assigned atom types and partial charges are correct, you can select Potential or Partial_charge in Molecule / Label to label each atom. Once you specify the information about the structure and parameters, you are ready to move to the **Discover** module. (We will not use **Discover_3** in this course).

The Constraint pulldown menu contains various atom-constraining and restraining procedures that you can select. In Parameters, you select the simulation type for Discover (Minimize, Dynamics, etc.), as well as the choice for cutoff parameters for nonbonded interactions, periodic boundary conditions (Variables), and dielectric constant (Set). Take time to familiarize yourself with the first three pulldown menus Constraint, Parameters, and Run, with Insight_help active, to learn about the various commands they contain.

To start a simulation, go to Run / Run, select desired options, choose the object for calculations, and execute.
Each Discover run is assigned a number in the order of the execution start time. The files created during the execution are identified by the calculation object (molecular system name) and the job integer (appended to the name). The file extension specifies the file type. Examples are listed below.

Discover Input Files:

- Commands (`.inp`)
- Cartesian Coordinates (`.car`)
- Molecular Data (`.mdf`)
- Force field Parameters (`.frc`)
- Restraints (`.rstrnt`)

Discover Output Files:

- Standard Output (`.out`)
- Cartesian Coordinates (final structure) (`.cor`)
- Cartesian Coordinate Archive (multiple frames) (`.arc`)
- Automatic Potential Parameter Assignment (`.prm`)
- Discover Dynamics Restart Information (`.rst`)

You can specify the files to save with Run / Files. By default, all are saved.

2. **Setting up biphenyl minimization**.

We will begin to learn about potential energy minimization for a simple yet interesting system, biphenyl (see Fig. D.2).

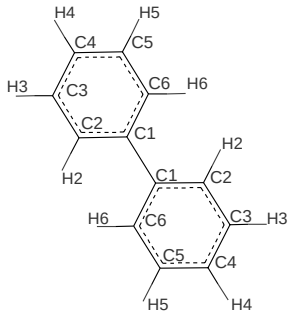

Figure D.2. Biphenyl

You will receive electronically two files containing the coordinates of biphenyl[8] The first, `biphenyl.car`, includes the structure with coplanar phenyl rings. This configuration was created with the **Builder** module by connecting two benzene rings.
The second file, `biphenyl_distorted.car`, contains the structure with each of the phenyl rings distorted from planarity.
Before displaying these structures, check that the AMBER force field is chosen. This will save work in assigning AMBER force field parameters. It will also permit you to proceed to Discover directly. (Note: For other force fields, you would have to assign parameters through Forcefield / Potentials). To open a coordinate file and display a structure, use Molecule / Get, specify Archive as the File Type, and select the desired file.

3. **Generation of energy profiles by restrained minimization**.

   A potential energy profile along some molecular coordinate, X (such as the rotamer dihedral angle χ_1), describes the dependence of the energy, minimized with respect to the remaining coordinates, on X. The simplest way to generate such a profile is to use minimization with *restraints*. Restraining a coordinate X to a specified value X^0 can be accomplished by adding harmonic penalty term,

   $$E_{\mathrm{rstr}} = K\,(X - X^0)^2,$$

   to the potential energy. After minimization, X should not deviate significantly from X^0 when the force constant K is large.[9] For a complete profile, minimum energy values must be calculated for a series of values $\{X_1^0, X_2^0,$

[8] Files can be obtained through the link to the course web site or directly from the author.

[9] *Constrained*, as opposed to restrained, minimization entails a more complex procedure to guarantee that $X = X^0$.

$X_3^0, \ldots\}$ in the range of X.

For biphenyl, we will analyze the dependence of energy on the torsion angle between the planes of phenyl rings. Four dihedral angles are defined about the C1–C1 bond connecting the two rings. They are specified by the following atom quadruplets {1B:C2, 1B:C1, 1:C1, 1:C6}, {1B:C6, 1B:C1, 1:C1, 1:C2}, {1B:C2, 1B:C1, 1:C1, 1:C2}, and {1B:C6, 1B:C1, 1:C1, 1:C6}. Restraining only one of them will result in a nonplanar geometry of phenyl rings (since the remaining dihedral angles will tend to assume values associated with a lower energy). To ensure that the phenyl-ring planes are not distorted, it is necessary to restrain a *pair* of dihedral angles to the same value. You can choose the first two or the last two atom quadruplets from the list above.

The plot for the full range of the angle, $[-180.0°, 180.0°]$, can be created by first computing energy minima for a sequence of values in the range $[0.0°, 90.0°]$ (e.g., $0.0°$, $10.0°$, $20.0°$, ..., $90.0°$), and then using symmetry operations.

Start with the coplanar structure (`biphenyl.car`). Make sure that potential parameters are properly assigned.

Then select Constraint / TorsionForce. You can now proceed in different ways to calculate the energy values for the profile. For instance, you can make 10 separate minimization runs, each time specifying both restraints (Intervals set to 1). Alternatively, you can execute one run specifying the range of values for both restraints (Intervals set to 9, Starting_Angle set to 0.0, and Angle_Size set to 90.0). In the latter case, you must extract the appropriate energy values from the output file. For two restraints, defined at ten points each, 100 energy values (corresponding to all restraints) will be listed as output. Extract only those values for which the restraint targets on both angles are identical.

Use Force Constant set to the range of 2000–5000.

Switch to Parameters / Minimize and select Conjugate gradient algorithm with Gradient tolerance set to 0.001.

Note that Parameters / Set and Parameters / Variables are left at their default values. Now proceed to Run.

Another possibility is to use Run / Files to limit the number of output files. Before executing the Run / Run command check the restraints and selected minimization options using the List option.

After minimization, the dihedral angle might deviate somewhat from the value specified in the restraint. Save the final structure (both dihedral angles are around $90°$) to `biphenyl.psv` using File / Save_Folder.

You can view these structures (frames) with Trajectory / Get and Trajectory / Conformation from the **Analysis** module and determine the torsion angle value.

Plotting the profile should be done only after completing the next section of the assignment.

4. **Unrestrained minimization for biphenyl**.

   In addition to the restrained minimization calculations, perform unrestrained minimization to find the "global" energy minimum, $E_{\min}$, for biphenyl.

   Now express the profile energy E from the previous section relative to the $E_{\min}$ (i.e., $E - E_{\min}$), and plot against the dihedral angle for the full range $[-180.0°, 180.0°]$.

   Note that $E_{\min}$ may be larger than some E values. Why is that?

5. **Comparison of different force fields**.

   Repeat the energy profile calculations with the cff91 force field. Plot the results obtained with the AMBER and cff91 force fields on one plot and discuss your findings.

6. **Dependence on initial conditions**.

   Perform unrestrained minimization of biphenyl starting with the structure specified in the `biphenyl.psv` file from Part 3. Use the cff91 or AMBER force field and any minimization algorithm you wish, but use the Derivative tolerance of 0.001.

   Describe the minimization algorithm briefly and discuss your results.

7. **Assessment of the performance of various minimization algorithms in Insight II**.

   For each minimization algorithm offered in Discover record the CPU time required for convergence of the energy gradient to the target values of 10.0, 0.1, 0.001, and 0.00001 kcal/Å. Use the AMBER force field.

   For each algorithm, begin with the structure contained in the file `biphenyl_distorted.car`. Select the desired Algorithm from Parameters / Minimize; set Iterations to 5000; specify the first target value of the derivative; execute; and proceed to execute the Run/Run command. After this job is completed, change the derivative tolerance to the

next target value and repeat minimization. Extract the computational times and values of minima from each output file. Do not increase the number of Iterations above 5000. If the specified convergence is not reached with this threshold, note that in your report.

Repeat this procedure for each of the remaining algorithms. (Remember to start with the structure from `biphenyl_distorted.car` file.) Construct a table comparing the performance of minimization algorithms in the different regions of derivative tolerance (report the timing and energy minimum values).

On the basis of these results, and the information you have learned in class, suggest a simulation schedule to achieve an optimal minimization of a large molecule. Note that for our small system the gradient norm associated with the initial configuration of biphenyl is not extremely large.

Background Reading from Coursepack

- M. Karplus and G. A. Petsko, "Molecular Dynamics Simulations in Biology", *Nature* **347**, 631–639 (1990).

Assignment 11: A Global Optimization Contest!

Our goal is to compute the lowest energy structure for the pentapeptide met-enkephalin, whose sequence is **Tyr–Gly–Gly–Phe–Met**. Many local minima exist for this molecule, so it is a challenge to reach the global minimum. *The student who finds the structure of the lowest energy will receive a prize from the instructor.*

The rules of this contest are:

1. use a molecule with *charged* COO^- and NH_3^+ ends
2. use the AMBER force field
3. use the distance dependent dielectric constant (**Discover** module, Parameters / Set command, Dist_Dependent button on)
4. use $1/2$ as the scale factor for 1–4 nonbonded interactions (i.e., Parameters / Scale_Terms command, p1_4 button on, and specify 0.5)

You can use *any* technique mentioned in this course (energy minimization, molecular dynamics, Monte Carlo sampling), as well as any other resources (e.g., web and literature), to find the global minimum of the pentapeptide.

Be Creative.

Hand in a detailed report describing how you reached the minimum for met-enkephalin and any particular difficulties, or interesting observations, you encountered along the way. Attach the Cartesian coordinate file and the energy value reached.

Also submit a three-dimensional picture of the configuration of lowest energy along with a table specifying all associated bond lengths and bond angle values, and the $\{\phi, \psi\}$ and χ dihedral-angle values per residue.

To qualify for consideration of the prize, send electronically the coordinate file with the minimized structure to the instructor and TA.

Good Luck!

Background Reading from Coursepack

- K. A. Dill and H. S. Chan, "From Levinthal to Pathways to Funnels", *Nature Struc. Biol.* **4**, 10–19 (1997).
- T. Lazaridis and M. Karplus, " 'New View' of Protein Folding Reconciled with the Old Through Multiple Unfolding Simulations", *Science* **278**, 1928–1931 (1997).

Assignment 12: Monte Carlo Simulations

1. **Random Number Generators**. Investigate the types of random number generators available on: (a) your local computing environment and (b) a mathematical package that you frequently use. How good are they? Is either one adequate for long molecular dynamics runs? Suggest how to improve them and test your ideas.

   To understand some of the defects in linear congruential random number generators, consider the sequence defined by the formula $y_{i+1} = (a\, y_i + c) \bmod M$, with $a = 65539$, $M = 2^{31}$, and $c = 0$. (This defines the infamous random number generator known as RANDU developed by IBM in the 1960s, which subsequent research showed to be seriously flawed). A relatively small number of numbers in the sequence (e.g., 2500) can already reveal a structure in three dimensions when triplets of consecutive random numbers are plotted on the unit cube. Specifically, plot consecutive pairs and triplets of numbers in two and three-dimensional plots, respectively, for an increasing number of generated random numbers in the sequence, e.g., 2500, 50,000, and 1 million. (Hint: Figure D.3 shows results from 2500 numbers in the sequence).

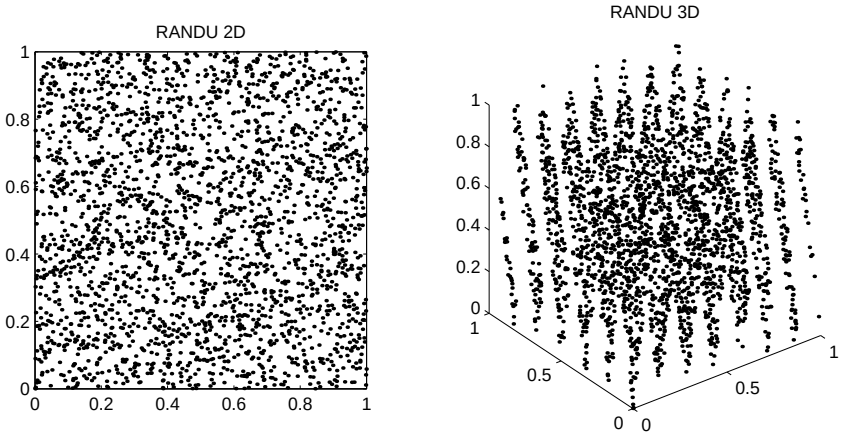

Figure D.3. Plots generated from pairs and triplets of consecutive points in the linear congruential generator known as RANDU defined by $a = 65539$, $M = 2^{31}$, and $c = 0$ when 2500 total points in the sequence are generated.

2. **MC Means**. Propose and implement a Monte Carlo procedure to calculate π based on integration. How many MC data points are needed to yield an answer correct up to 5 decimal places? What is the computational time

involved? Show a table of your results displaying the number of MC steps, the associated π estimate, and the calculated error.

3. **Gaussian Variates**. You are stranded in an airport with your faithful laptop with one hour to spare until the deadline for emailing your homework assignment to your instructor. The assignment (next item) relies on a *Gaussian random number generator*, but you have forgotten the appropriate formulas involved in the commonly used Box/Muller/Marsaglia transformation approach. Fortunately, however, you remember the powerful Central Limit Theorem in basic probability and decide to form a random Gaussian variate by sampling N uniform random variates $\{x_i\}$ on the unit interval as

$$\bar{y} = \sum_{i=1}^{N} x_i \, .$$

You quickly program the expression:

$$y = \sqrt{\frac{1}{\sigma^2(\bar{y})}} \sum_{i=1}^{N} [x_i - \mu(\bar{y})]$$

where above σ^2 is the standard deviation of $\bar{y} = N\sigma^2(x)$ and the mean $\mu(\bar{y}) = N\mu(x)$. [Recall that the uniform distribution has a mean of 1/2 and variance of 1/12].

How large should N be, you wonder. You must finish the assignment in a hurry. To have confidence in your choice, you set up some tests to determine when N is sufficiently large, and send your resulting routine, along with your testing reports, and results for several choices of N.

4. **Brownian Motion**. Now you can use the Gaussian variate generator above for propagating *Brownian motion* for a single particle governed by the biharmonic potential $U(x) = kx^4/4$. Recall that Brownian motion can be mimicked by simulating the following iterative process for the particle's position:

$$x^{n+1} = x^n + \frac{\Delta t}{m\gamma} F^n + R^n$$

where

$$\langle R^i R^j \rangle = \frac{2k_B T \Delta t}{m\gamma} \delta_{ij}, \qquad \langle R^i \rangle = 0 \, .$$

Here m is the particle's mass; γ is the collision frequency, also equal to ξ/m where ξ is the frictional constant; and F is the systematic force. You are required to test the obtained mean square atomic fluctuations against the

known result due to Einstein:

$$\langle x^2 \rangle = 2 \left(\frac{k_B T}{m\gamma} \right) t = 2D\,t\,,$$

where D is the diffusion constant.
The following parameters may be useful to simulate a single particle of mass $m = 4 \times 10^{-18}$ kg and radius $a = 100$ nm in water: by Stokes' law, this particle's friction coefficient is $\xi = 6\pi\eta a = 1.9 \times 10^{-9}$ kg/s, and $D = k_B T/\xi = 2.2 \times 10^{-12}\text{m}^2/\text{s}$. You may, however, need to scale the units appropriately to make the computations reasonable.

Plot the mean square fluctuations of the particle as a function of time, compare to the expected results, and show that for $t \gg 1/\gamma = 2 \times 10^{-9}$ s the particle's motion is well described by random-walk or diffusion process.

Background Reading from Coursepack

- T. Schlick, E. Barth, and M. Mandziuk, "Biomolecular Dynamics at Long Timesteps: Bridging the Timescale Gap Between Simulation and Experimentation", *Ann. Rev. Biophys. Biomol. Struc.* **26**, 179–220 (1997).
- E. Barth and T. Schlick, "Overcoming Stability Limitations in Biomolecular Dynamics: I. Combining Force Splitting via Extrapolation with Langevin Dynamics in LN", *J. Chem. Phys.* **109**, 1617–1632 (1998).
- L. S. D. Caves, J. D. Evanseck, and M. Karplus, "Locally Accessible Conformations of Proteins: Multiple Molecular Dynamics Simulations of Crambin", *Prot. Sci.* **7**, 649–666 (1998).
- M. Karplus and J. A. McCammon, "Molecular Dynamics simulations of Biomolecules", *Nat. Struc. Biol.* **9**, 307–340 (2003).
- M. Karplus and J. Kuriyan, "Molecular Dynamics and Protein Function", *Proc. Natl. Acad. Sci. USA* **102**, 6679–6685 (2005).
- S. A. Adcock and J. A. McCammon, "Molecular dynamics: survey of methods for simulating the activity of proteins", *Chem. Rev.* **106**: 1589–1615 (2006).
- E. H. Lee, J. Hsin, M. Sotomayor, G. Comellas, and K. Schulten, "Discovery Through the Computational Microscope", *Structure* **17**: 1295–1306 (2009).

Assignment 13: Advanced Exercises in Monte Carlo and Minimization Techniques

1. Study the function:

$$E(x,y) = ax^2 + by^2 + c\,(1-\cos\gamma x) + d\,(1-\cos\delta y)\,. \qquad \text{(D.11)}$$

Note that is has many local minima and a global minimum at $(x,y) = (0,0)$. Minimize $E(x,y)$ with $a = 1, b = 2, c = 0.3$, $\gamma = 3\pi$, $d = 0.4$, and $\delta = 4\pi$ by the standard simulated annealing method. Use the starting point $(1,1)$ and step perturbations $\Delta x = 0.15$, and set β in the range of 3.5 to 4.5. Limit the number of steps to ~ 150. Now implement the *variant* of the simulated annealing method where acceptance probabilities for steps with $\Delta E < 0$ are proportional to $\exp(-\beta E^g \Delta E)$, with the exponent $g = -1$. Analyze and compare the efficiency of the searches in both cases. It will be useful to plot all pairs of points (x,y) that are generated by the method and distinguish 'accepted' from 'rejected' points.

2. Devise a different variant of the basic simulated annealing minimization method that would incorporate *gradient* information to make the searches more efficient.

3. Consider the following global optimization deterministic approach based on the *diffusion equation* as first suggested by Scheraga and colleagues (L. Piela, J. Kostrowicki, and H. A. Scheraga, "The Multiple-Minima Problem in Conformational Analysis of Molecules. Deformation of the Potential Energy Hypersurface by the Diffusion Equation Method", *J. Chem. Phys.* **93**, 3339–3346 (1989)).
The basic idea is to deform the energy surface smoothly. That is, we seek to make "shallow" wells in the potential energy landscape disappear iteratively until we reach a global minimum of the deformed function. Then we "backtrack" by successive minimization from the global minimum of the transformed surface in the hope of reaching the global minimum of the real potential energy surface. This idea can be implemented by using the heat equation where T represents the temperature distribution in space x, and t represents time:

$$\frac{\partial^2 T}{\partial x^2} = \frac{\partial T}{\partial t} \qquad \text{(D.12)}$$

$$T(x,0) = E(x)\,. \qquad \text{(D.13)}$$

Here, the boundary condition at time $t = 0$ equates the initial temperature distribution with the potential energy function $E(x)$. Under certain conditions (e.g., E is bounded), a solution exists. Physically, the application of this equation exploits the fact that the heat flow (temperature distribution) should eventually settle down.

To formulate this idea, let us for simplicity consider first a one-dimensional problem where the energy function E depends on a scalar x. Let $E^{(n)}(x)$ represent the nth derivative of E with respect to x and define the transformation operator $\mathcal{S}$ on the energy function E for $\beta > 0$ as follows:

$$\mathcal{S}[E(x)] = E(x) + \beta E^{(2)}(x)\,. \tag{D.14}$$

That is, we have:

$$\begin{aligned}
\mathcal{S}^0 E &= E \\
\mathcal{S}^1 E &= E + \beta E^{(2)} \\
\mathcal{S}^2 E &= E + 2\beta E^{(2)} + \beta^2 E^{(4)} \\
\mathcal{S}^3 E &= E + 3\beta E^{(2)} + 3\beta^2 E^{(4)} + \beta^3 E^{(8)} \\
\vdots &\qquad\qquad \vdots \\
\mathcal{S}^N E &= (1 + \beta d^2/dx^2)^N E\,.
\end{aligned}$$

Now writing $\beta = t/N$ where t is the time variable, and letting $N \to \infty$, we write:

$$\exp\left(td^2/dx^2\right) E \equiv \exp(A(t))\, E = \left[1 + A + \frac{A^2}{2!} + \frac{A^3}{3!} + \cdots\right]. \tag{D.15}$$

Thus we can define $T(t)$ as

$$T(t) = \exp((A(t)) = \exp(td^2/dx^2)\,. \tag{D.16}$$

In higher dimensions, let x represent the collective vector of n independent variables; we replace the differential operator above d^2/dx^2 by the *Laplacian operator*, that is

$$\Delta = \sum_{i=1}^{n} \partial^2/\partial x_i\,.$$

Using this definition, we can also write

$$T(t) = T_1(t)\, T_2(t)\, \cdots\, T_n(t)$$

where

$$T_i = \exp(t\partial^2/\partial x_i)\,.$$

This definition produces the heat equation (D.12, D.13) since

$$\begin{aligned}
\frac{\partial T(t)[E(x)]}{\partial t} &= \left[\frac{dA}{dt} + \frac{2A}{2}\frac{dA}{dt} + \frac{3A^2}{3!}\frac{dA}{dt} + \cdots\right][E] \\
&= \left[1 + A + \frac{A^2}{2} + \cdots\right]\frac{d^2}{dx^2}[E]
\end{aligned}$$

$$= \frac{\partial^2 T(t)}{\partial x^2}[E(x)]\,.$$

In practice, the diffusion equation method for global optimization is implemented by solving the heat equation by Fourier techniques (easy, for example, if we have dihedral potentials only) or by solving for T up to a sufficiently large time t. This solution, or approximate solution (representing $E(x,t)$ for some large t), is expected to yield a deformed surface with one (global) minimum. With a local minimization algorithm, we compute the global minimum x^* of the deformed surface, and then begin an iterative deformation/minimization procedure from x^* and $E(x,t)$ so that at each step we deform backwards the potential energy surface and obtain its associated global minimum ($E(x,t) \longrightarrow E(x,t-\Delta t)$ and x^* to x^1, $E(x,t-\Delta t) \longrightarrow E(x,t-2\Delta t)$ and x^1 to x^2, $\cdots E(x,0) \longrightarrow x^k$). Of course, depending on how the backtracking is performed, different final solutions can be obtained.

(a) To experiment with this interesting diffusion-equation approach for global minimization, derive a general form for the deformation operator $T(t) = \exp(td^2/dx^2)$ on the following special functions $E(x)$: (*i*) polynomial functions of degree n, and (*ii*) trigonometric functions $sin\omega x$ and $cos\omega x$, where ω is a real-valued number (frequency). What is the significance of your result for (*ii*)?

(b) Apply the deformation operator $T(t) = \exp(td^2/dx^2)$ to the quadratic function

$$E(x) = x^4 + ax^3 + bx^2\,, \tag{D.17}$$

with $a = 3$ and $b = 1$. Evaluate and plot your resulting $T(t)E(x)$ function at $t = 0, \Delta t, 2\Delta t, ...$, for small time increments Δt until the global minimum is obtained.

(c) Apply the deformation operator $T(t)$ for the two-variable function in eq. (D.11). Examine behavior of the deformation as $t \to \infty$ as a function of the constants a and b. Under what conditions will a unique minimum be obtained as $t \to \infty$?

4. Use Newton minimization to find the minimum of the two-variable function in equation (D.11) and the one-variable function in equation (D.17). It is sufficient for the line search to use simple bisection: $\lambda = 1, 0.5$, etc., or some other simple backtracking strategy. For the quartic function, experiment with various starting points.

Read remaining paper from Coursepack

Assignment 14: Advanced Exercises in Molecular Dynamics

1. Calculate the 'natural' time unit for molecular dynamics simulations of biomolecules from the relation: energy = mass $*$ (length/time)^2, to obtain the time unit τ corresponding to the following units:

length	(l):	1 Å = 10^{-10}m
mass	(m):	1 amu = 1 g/mol
energy	(v):	1 kcal/mol = 4.184 kJ/mol.

   Estimate the "quantum mechanical cutoff frequency", $\omega_c = k\mathrm{T}/\hbar$ at room temperature ($\sim 300^\circ$K).

2. Derive the amplitude decay rate of $\gamma/2$ for an underdamped harmonic oscillator due to *friction* by solving the equations of motion:

   $$m\frac{d^2x}{dt} = -kx - m\gamma\frac{dx}{dt} \tag{D.18}$$

   and examining time behavior of the solution.

3. Derive the amplitude decay rate of $\omega^2(\Delta t)/2$ *intrinsic* to the *implicit-Euler* scheme by solving the discretized form of eq. (D.18).

4. Compare your answer in problem 2 above with behavior of the *explicit-Euler* solution of eq. (D.18).

5. Compare molecular and Langevin dynamics simulations of two water molecules by the Verlet discretization of the equation of motion and its Langevin analog. Use the "SPC" *intermolecular* potential, given by:

   $$E = \sum_{\substack{(i,j)\equiv(\mathrm{O},\mathrm{O})\ \text{pairs}\\ i<j}} \left[\frac{-A}{r_{ij}^6} + \frac{B}{r_{ij}^{12}}\right] + \sum_{\substack{(k,\ell)\equiv\text{intermolecular}\\ (\mathrm{O},\mathrm{O}),(\mathrm{O},\mathrm{H}),(\mathrm{H},\mathrm{H})\ \text{pairs}\\ k<\ell}} \left[\frac{Q_kQ_\ell}{r_{k\ell}}\right]$$

   where

   $$\begin{aligned} A &= 626\ (\text{kcal Å}^6)/\text{mol} \\ B &= 629\times 10^3(\text{kcal Å}^{12})/\text{mol} \\ Q_\mathrm{H} &= 0.41\ e \\ Q_\mathrm{O} &= -0.82\ e\ . \end{aligned}$$

   A numerical factor of 332 is needed in the electrostatic potential to obtain energies in kcal/mol with the coefficients above. For simplicity, assume that *intramolecular* geometries are rigid: r_{OH} = 1 Å, $\cos\theta_{\mathrm{HOH}} = -1/3$. (You can use harmonic soft constraints). Begin by first minimizing the energy

of the water dimer and examining the hydrogen bond geometry (hydrogen-bond distance and angle θ between the hydrogen-bond vector and bisector of the acceptor molecule). Then study numerical behavior of the two models/schemes as a function of Δt, and examine the hydrogen bond geometry. Experiment with Δt = 1, 2, 5, and 10 fs and use γ values in the range of 1 to 50 ps^{-1}. If you are more ambitious, continue to study energy-minimized structures of water clusters of larger sizes and their dynamics. Analyze the hydrogen bonding networks of these clusters.

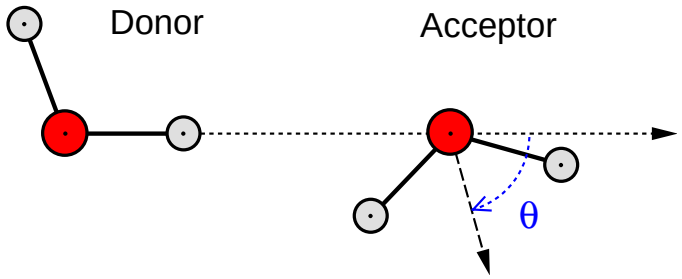

Figure D.4. Hydrogen bond geometry: the angle θ is defined between the hydrogen-bond vector and the bisector of the acceptor molecule.

Some Useful Constants and Conversion Factors

Avogadro's Number	$N_A = 6.0221 \times 10^{23}$ mol^{-1}
Planck's Constant	$h = 6.6261 \times 10^{-34}$ Jsec
	$\hbar = h/2\pi = 1.055 \times 10^{-34}$ Jsec
Boltzmann's Constant	$k_B = 1.38066 \times 10^{-23}$ JK^{-1}
Gas Constant	$R = k_B N_A = 8.3145$ JK^{-1} mol^{-1}
Atomic Mass Unit, amu	$(1/N_A) = 1$ g/mol = 1.6605 $\times 10^{-27}$ kg
	$\pi = 3.14159$
	1 kcal = 4.184 kJ

Assignment 15: BONUS PROJECT

The Scaling of Protein-Conformer Number with Size and Solvability of the Protein Folding Problem

The phone rings one morning as you sip through your Sunluck vanilla latté grande and check your emails in your office at the University of Seabeetle. The editor of the journal *Proteomics Today* is on the line.

Given your expertise in biomolecular modeling, she asks your help in writing a brief *Folding In Silico* Perspectives article for the next issue discussing the following series of interesting papers debating the nature of scaling of the number of protein conformers as function of chain length (exponential or nonexponential) and the solvability of the protein folding problem by computer simulation:

- W. F. van Gunsteren, R. Bürgi, C. Peter, and X. Daura, "The Key to Solving the Protein-Folding Problem Lies in an Accurate Description of the Denatured State", *Angew. Chem. Int. Ed.* **40**, 352–355 (2001).
- A. R. Dinner and M. Karplus, "Comment on the Communication 'The Key to Solving the Protein-Folding Problem Lies in an Accurate Description of the Denatured State' by van Gunsteren et al.", *Angew. Chem. Int. Ed.* **40**, 4615–4616 (2001).
- W. F. van Gunsteren, R. Bürgi, C. Peter, and X. Daura, "Reply", *Angew. Chem. Int. Ed.* **40**, 4616–4618 (2001).

Since many of *Proteomics Today* readers and authors perform computer simulations of proteins, both macroscopic and all-atom based, the editor wants you also to mention the strengths and weaknesses of these different approaches.

Though already overloaded with preparing final examinations for your classes, writing grant proposals, supervising your students and postdocs, and reviewing several articles for journals (long-overdue), you agree to take this challenging and potentially rewarding assignment. You immerse yourself in the papers, send your graduate students to collect background articles and information, order through your cellphone Sunluck's caramel-macchiato enormoso, and decide to make your article not only objective and interesting but also fun to read.

Some Suggestions:

- Discuss the Levinthal paradox.
- What is the relation among the number of conformers, timescales, and folding pathways?
- Analyze lattice or other protein simulations reported in the literature to estimate the number of possible conformers and the timescale of protein folding.