Lecture 1: Computational Biology

I. Logistics

Course Instructor: Dr. Patrick Fleming
pat.fleming@jhu.edu
G85 UTL
Office Hours: Tuesday and Thursday, 9 – 10:30 am

Classes will meet Tuesdays and Thursdays 10:30 – 11:45 am in the UTL G98 (Computer Lab). Both lectures and labs will be in this room.

Lecture Notes will be available for downloading from `pages.jhu.edu/pfleming/compbio/`. Access to the lecture notes is password protected. Login information will be given in class.

Text. There is no single text that covers the material in the course. A list of reference materials for the course is available at `pages.jhu.edu/pfleming/compbio/`. These are listed for those students who wish to read some of the primary sources for topics presented in class. All necessary reading material will be provided on-line or in class.

Computer lab. A flash drive would be useful for you to bring to class. Access to the computer room outside of class is permitted and encouraged. Some computer experiments will take longer than lab time and completion of the experiments will be considered part of the home work assignments. The computer lab is for students enrolled in Biophysics classes who have been given card access. **Please do not allow access to others.**

Some students prefer to bring their laptops (Mac) to the computer labs and use those rather than the desktop computers. You will need to load a number of open source software applications if you want to do this and I will help you with the installations.

Lab Reports. There will be highlighted questions in the computer lab guides. The answers to these questions, submitted by email to the TA, are considered your lab reports.

Homework Assignments. Specifics will be given at the time of the assignments.

Exams. There will be one midterm and one final exam. The midterm exam will be written in class and will cover material given in lectures and learned in the labs. **Background material provided in the lab guides will also be covered**. The final exam will be a report in the style of a scientific publication describing a research project you will carry out near the end of the class. Details will be given in class.

Grading. Grades on the lab reports, homework assignments, the mid-term exam and the final exam will contribute to the final grade for the course. The mid-term and final exams

each will count 30 percent, the homework average grade will count 30 percent and the average lab report grade will count 10 percent toward the final grade.

***Learning Objectives*.**

1. Develop and apply knowledge of UNIX and Python computer languages to the manipulation and analysis of protein models.
2. Develop knowledge of and practice using the VIM text editor.
3. Interpret protein structure PDB format files.
4. Recognize and apply Monte Carlo and molecular dynamics methods to generate ensembles of molecules at different energetic states.
5. Apply computer methods to analyze and evaluate conformational diversity in protein populations.
6. Understand, analyze and discuss the role of conformational diversity in the allosteric control of protein function.
7. Understand and discuss the differences and relationships between molecular population average and stochastic approaches to studying protein function.
8. Practice and asses the role of conformational diversity in protein binding events.

Ethics. The administration has asked us to include the following statement relating to academic ethics in our course descriptions and we all take this very seriously. **You are encouraged to collaborate with your fellow students during the computer labs. However, all material submitted to the instructor must be your own work.**

Cheating is wrong. Cheating hurts our community by undermining academic integrity, creating mistrust, and fostering unfair competition. The university will punish cheaters with failure on an assignment, failure in a course, permanent transcript notation, suspension, and/or expulsion. Offenses may be reported to medical, law, or other professional or graduate schools when a cheater applies.

Violations can include cheating on exams, plagiarism, reuse of assignments without permission, improper use of the Internet and electronic devices, unauthorized collaboration, alteration of graded assignments, forgery and falsification, lying, facilitating academic dishonesty, and unfair competition. Ignorance of these rules is not an excuse.

You can read a full description of the ethics code at http://e-catalog.jhu.edu/undergrad-students/student-life-policies/.

II. Overview

In a cell you may have a hundred copies of a particular protein. But these copies are not identical in conformation. Rather the hundred copies exist as an ensemble of structures that vary considerably in overall conformation and/or configuration. The structure of a biomolecule one typically views with molecular graphics images represents an *average* of the molecular population and may project a false impression of the underlying structural diversity. An ensemble, or population, of structures results from the many degrees of conformational freedom which each molecule samples. An understanding of this structural diversity and conformational flexibility is important to the understanding of the cellular function of biological macromolecules.

One can envision an ensemble of protein conformations existing instantaneously among all the copies of that protein or one can envision a single protein visiting each of the different conformations over time. A recent paper on protein function states the case better than I can. "Cellular function requires biomolecules to undergo dynamic transitions that include folding, conformational rearrangements, and large-scale assembly. The result is a highly interdependent network of processes that is maintained by a balance of thermodynamic and kinetic factors. In molecular machines, each constituent biopolymer (i.e., a chain of residues) first folds to a low-energy configuration/ensemble. These ordered polymers can then assemble into sophisticated architectures, which undergo conformational transitions during function. In contrast to the dynamics of macroscopic machines, molecular-level processes are stochastic, where the molecular interactions that ensure structural integrity are weak (i.e., on the scale of energetic fluctuations from solvent). In this dynamic environment, biomolecules constantly fluctuate, and the extent of disorder is heterogeneous between residues." (Paul Whitford, PNAS, 118:7114, 2013). We will come back to many of the concepts alluded to by Whitford in this statement.

In this course we will examine the dynamic nature of proteins and how these dynamics impact on the experimental determination of protein structure and on protein function. We will also explore how dynamic systems are generated computationally and how they are analyzed especially with regard to how these analyses relate to experimental data on biological systems.

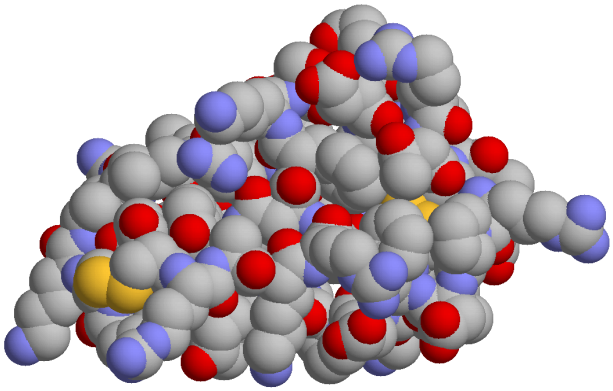
Exercises in using computer programs will be given. Minimal coding will be required but we will develop an ability to read and understand Python code, run simple and complex Python programs, use common research tools for molecular simulations, manipulate, analyze and view atomic and molecular models and characterize conformational ensembles.

III. Main themes

There are two major take-home messages from the course:

1. The structural model represented by a molecular graphics image of an X-ray crystal structure is a simplification of the true structure of a protein. Below is shown a typical molecular graphics image of a small protein (58 residues). The function of this particular protein is to bind to a protease and inhibit the protease. As we will see from a

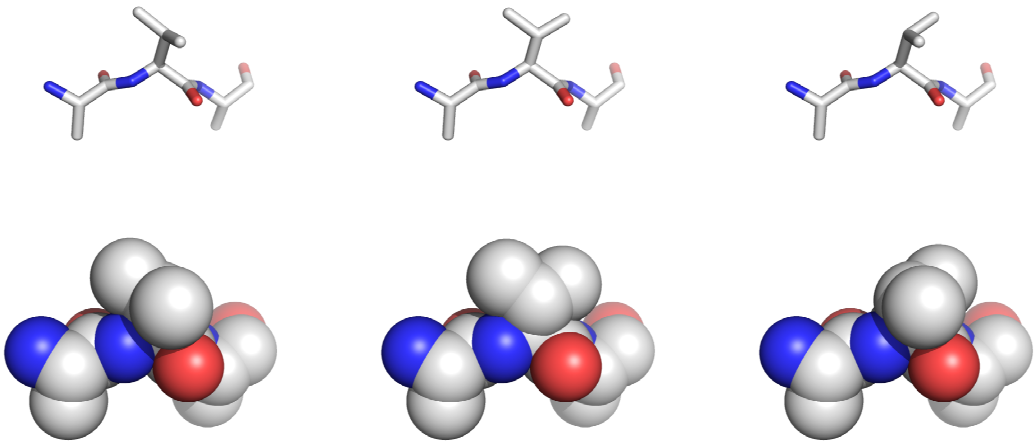
molecular dynamics movie in class the binding event must occur in the context of **significant fluctuations in the structure** that are not represented by the picture.



How does nature ensure that the correct, binding-competent conformation will be found in this highly mobile structure? The answer to this question involves a probabilistic, ensemble description of the molecule.

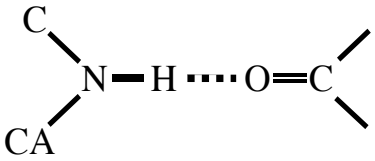
We will investigate two major classes of protein dynamics that affect protein function. The first class of conformational dynamics occurs on a time scale of picoseconds (ps) or 10^{-12} sec. The molecular dynamics movie shown in class represents 10 picoseconds (ps) of protein fluctuations, each fluctuation (or frame of the movie) is 0.1 ps or 10^{-13} sec. This is the time range in which exposed amino acid residue side chains can change conformation.

For example, valine is an amino acid with a branched alkane side chain. Below are shown three commonly observed conformations of the $-CH_2-(CH3)_2$ side chain groups of valine (hydrogens not shown). Different side chain conformations are called **rotamers**.



In the movie you will see mostly fluctuations of side chains *within* a single rotameric conformation (<+/- 60° rotation around the average dihedral angle) and only occasionally a *flip* between rotamers. In the static model of an X-ray crystal structure, fluctuations *within* a rotamer are considered as the same conformation.

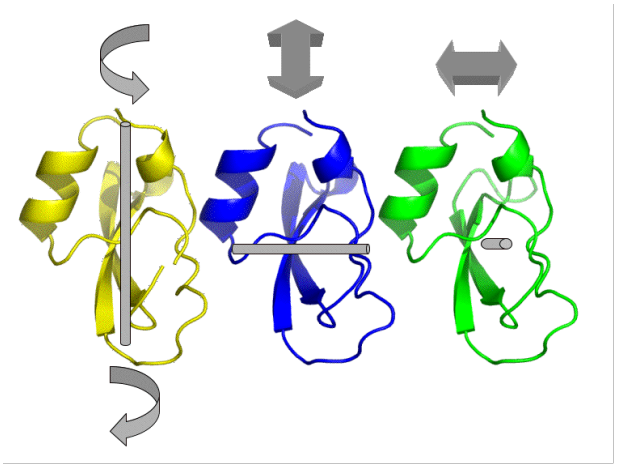
As we will see in the movie hydrogen bonds also break and re-form on the time scale of picoseconds. An example of a hydrogen bond between the backbone amide groups found in proteins is shown below as the dotted line,



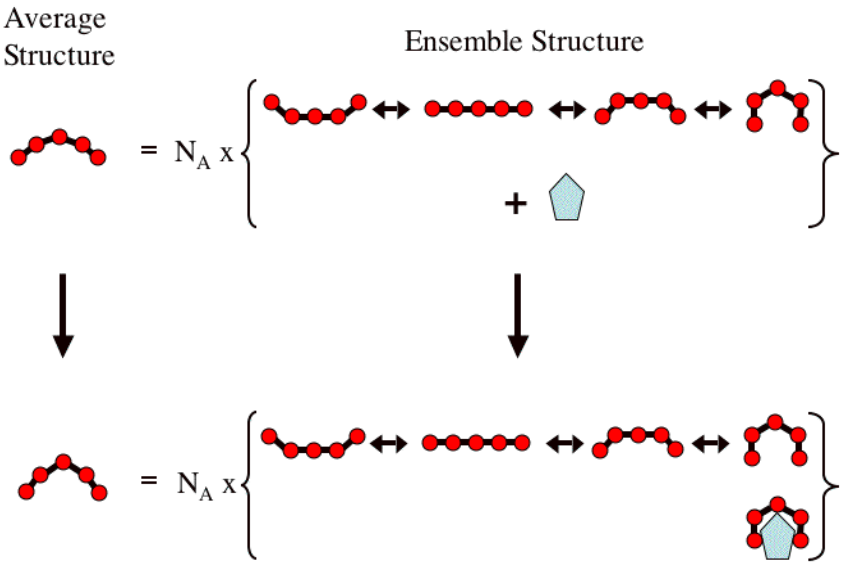
Hydrogen bonds are important for the stability and function of proteins, yet they are often transiently broken. This paradox may be understood from probabilistic approaches learned in this course.

After viewing the movies one should ask the question of what it means to define a specific residue side chain position or the presence of a hydrogen bond in the average structure represented by a crystallographic model. If the atoms are close to being in position for a hydrogen bond but are not quite there should one count it? As you will come to learn this is really not a valid question. The question should not be about all-or-none. Rather, the question should be about frequencies or probabilities.

The second class of protein dynamics occurs on a much slower time scale of nanoseconds (ns) to microseconds (μs) or 10^{-9} to 10^{-6} seconds. These motions involve fluctuations of the protein backbone as well as side chain motions. Many proteins undergo opening and closing or twisting movements on this slower time scale. The period of time captured in the movie shown in class is not long enough to see significant backbone fluctuations. However, many proteins undergo opening and closing or twisting movements during their functions. These motions are due to **collective** movement of groups of atoms in the protein. We can observe these larger scale, slower motions with a simulation of **normal modes** of the protein. Normal modes are equivalent to deconvolving the complex fluctuations into a distribution of simpler motions. Below is an illustration of the types of relative normal modes protein backbones may undergo. The rods in the figure below represent the axes of rotation and the arrows indicate the directions of rotation.



2. The second take-home message of the course is that the "induced fit" explanation for specific binding of a ligand by a protein is not the only mechanism of protein function. A different explanation will be examined in this course. In this second mechanism, binding events between a protein and a ligand (small molecule or other macromolecule) are examples of **conformational selection** from the ensemble. Consider the following figure. The red bead polymer represents a protein. It exists in a distribution of structures called an ensemble (here the N_A indicates that in a true ensemble one would have on the order of Avagadro's number of molecules, not just four). When we "determine" the structure in solution (or even in a crystal lattice) we actually determine an "average" structure. Notice that the conformation of the average structure on the left may not actually exist in the ensemble!



When a ligand, represented by the pentagonal structure in the figure, is added to the ensemble of protein structures it is likely that only one or a few ensemble structures are able to bind the ligand. This is what we mean by **conformational selection** from the ensemble. If the relative concentration of the ligand is high and many of the protein molecules in the ensemble bind the ligand, the distribution of ensemble structures will be biased toward the conformation that binds the ligand (Le Chatelier's principle or law of mass action). The result would be an apparent conformational change in the average structure.

What may appear to be a structural change between two static structures (shown on the left in the figure) is better understood as a shift in a **distribution of structures**. We will look at the experimental results that inform us about actual protein molecular fluctuations and how these results are consistent with an ensemble view of molecular function.

That's it! If you remember nothing else from this course but the above two take-home messages, I will consider it a successful course. Naturally, we will work through a few of the details during the semester.

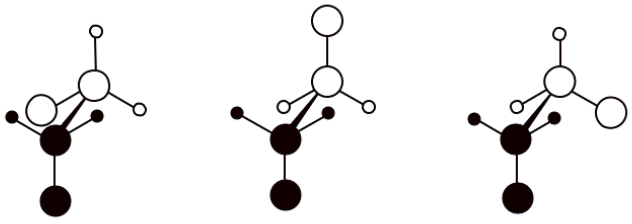
IV. Scales of Conformational Fluctuation

The conformational fluctuations of polypeptides are due to thermal motion or kinetic energy of the atoms (Brownian motion). These conformational fluctuations occur on different **time scales** depending on the type of motion involved.

Time scale		Amplitude	Type
femto - picoseconds	10^{-15} s - 10^{-12} s	0.001 Å - 0.1 Å	Bond stretching Angle bending
pico - nanoseconds	**10^{-12} s - 10^{-9} s**	**0.1 Å - 10 Å**	**Exposed side chain rotation, loop motion**
nano - microseconds	**10^{-9} s - 10^{-6} s**	**1 Å - 100 Å**	**Helix - coil transition, hinge bending**
micro - seconds	10^{-6} s - 10^{-0} s	10 Å - 100 Å	Macroscopic fluctuations, protein folding

These categories represent ranges on a continuum. Significant overlap occurs between the time scales of different types of fluctuations, but the categories are useful for discussion. The time regimes discussed in this course are in bold font. **You should memorize this table.**

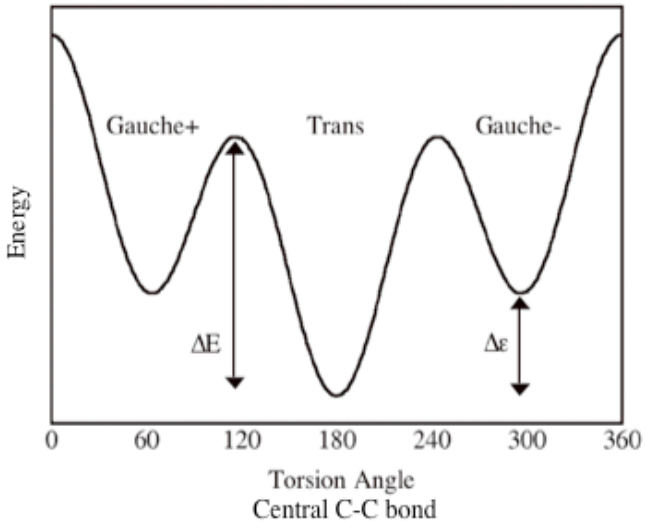
The time scale of conformational interconversion is dependent on the relative energies involved in the conformational path of fluctuations. To understand this concept, it is helpful to classify conformational flexibility into two categories; **static flexibility** and **dynamic flexibility**. Consider a carbon-carbon chain of saturated tetrahedral bonding (sp^3 orbitals for the chemists). There are three energy minima in the torsion angle around the middle C-C bond as shown in the plot below (hydrogens shown on central two carbons only). These conformations are called *gauche+*, *trans* and *gauche-*.



gauche + trans gauche -

There are two key energy differences that determine whether a chain will have what is called static or dynamic flexibility. These energy differences are **ΔE**, the energy **barrier** separating the minima and **Δε**, the energy **difference** between the minima as shown

schematically in the simplified figure below.



When the **ΔE** is smaller than **thermal energy, RT** (~ 0.6 kcal/mol at 300K, where **R** is the gas constant (1.987 cal K^{-1} mol^{-1}) and **T** is the temperature in Kelvin) the energy barrier between the *trans* and *gauche* conformations (i.e. the transition) is frequently reached and the *trans-gauche* transition can take place on a time scale of picoseconds to nanoseconds. This type of motion is considered dynamic flexibility. Dynamic flexibility is associated with amino acid residue side chain flexibility if the side chain is exposed on the surface of a protein and is free to rotate. We observed this type of motion in the molecular dynamics movie. Also, exposed polypeptide loops are considered to have dynamic flexibility. Dynamic flexibility causes a smearing of the atomic electron density in X-ray crystallography maps but the structural model is built with *one average conformation*.

On the other hand, if **ΔE** is significantly greater than **RT** then the frequency of transition is limited and the flexibility is considered static, i.e. specific conformations will persist for a relatively longer time. These are the types of motions observed in the normal mode description above. Static flexibility can cause a "bi-lobed" smearing of the electron density in X-ray maps and the structure is built with *two* (or more) *alternate conformations*. The value of **ΔE** would only determine the **rate** at which the chains reached this state but not the distribution of the states; **Δε** determines the **distribution** of states at equilibrium.

When **Δε** is smaller than **RT** the difference in energy between the *trans* and *gauche* conformations is inconsequential and both conformations are populated essentially equally. When **Δε** is significantly larger than thermal energy the *trans* conformation (in the plot above) is more highly populated than the *gauche* conformations (because it has a lower, more favorable energy). In this case a population of long chains with such rotating bonds would have a high proportion of trans conformation at any instant in time.
To summarize: **ΔE** controls the rate of change, **Δε** determines the equilibrium ratio or distribution of conformations.

Unfolded polypeptides in water appear to have **Δε** values for the rotatable backbone bonds larger than **RT** and therefore *certain conformations are preferred* as we will see later in the course. The **ΔE** for backbone bond rotation in unfolded proteins is near **RT** so *many conformations are sampled* even though there is a preference for certain conformations.

If both **ΔE** and **Δε** are larger than **RT** the chain would adopt most frequently the lowest energy conformation and stay there for a significant time - but importantly - other conformations would occasionally be observed although much less frequently. This is the case for folded polypeptides, i.e. most native proteins.

The important point here is that the energy/RT ratios describe **probabilities** of the distributions and fluctuations, not all-or-none states. Even a folded protein will infrequently sample an unfolded conformation!

V. Theory

Ludwig Boltzmann revolutionized physics in the late 1800's with the introduction of statistical methods to the understanding of fundamental laws of physics. Irwin Schroedinger said that "no perception in physics has ever seemed more important to me than that of Boltzmann – despite Planck and Einstein" (What is Life?: The physical Aspect of the Living Cell with Mind and Matter & Autobiographical Sketches, Cambridge University Press, 1992).

Before Boltzmann much of physical theory was believed to be strictly deterministic as in Newtonian mechanics. Newton's three laws of motion attempt to say something about every molecule in a system. In contrast the laws of thermodynamics say something about the macroscopic properties of a system, e.g. heat, energy and entropy, without reference to any individual molecules as the source of these properties. The laws of thermodynamics were considered basic axioms that were true *a priori* and unquestionably accepted.

Boltzmann introduced the notion that probability was fundamental to physics and that thermodynamic laws could be derived from the statistics of large ensembles of atomic particles. He developed what is now called **statistical mechanics** that is a bridge between Newtonian mechanics and the laws of thermodynamics.

Boltzmann's formative starting point was that all physics was based on the idea of collections of atoms giving rise to macroscopic observables. Perhaps his greatest contribution was an expression for entropy **S** based on probability – this expression is carved on his tombstone (**S = k logW**). The idea of entropy was described by Rudolph Clausius in 1865 as the amount of energy released during combustion that was always lost and could not be converted into work. Boltzmann gave a statistical definition of entropy and also provided the energetic value associated with entropy. His constant, **k** ($1.38\ E^{-16}$ erg/K) is the molecular gas constant divided by Avagadro's number ($\mathbf{R/N_A}$).

Let's use an example to introduce some terminology concerning Boltzmann's work; this is terminology that we will use throughout the course. Assume we place four coins on the table in different random combinations of heads (H) and tails (T). The collection of four coins is a **system** of coins and the different combinations of H and T in each system represent the **microscopic ways** that exist for each **macroscopic average** number of H and T in the system. There is no preference for whether the coin lands heads up or tails up (they have the same **potential energy**). The combinations that make up the different macroscopic averages are listed below:

Macrostate	Microstates	# Microstates
0 Tails	HHHH	1 way
1 Tail	THHH, HTHH, HHTH, HHHT	4 ways
2 Tails	TTHH, HTTH, HHTT, THHT, HTHT, THTH	6 ways
3 Tails	HTTT, THTT, TTHT, TTTH	4 ways
4 Tails	TTTT	1 way
		---
		16 total microstates

A specific combination (configuration) of coins (e.g., HTHH) may be considered a **microstate**. In contrast a **macrostate** (one tail, two tails, etc) is defined by the **average** number of tails or heads and may have many different microstates (THHH, HTHH, etc). If we divide the number of ways (microstates) to achieve each macrostate by the total number of microstates in all possible macrostates we obtain the probability to find the system in the specific macrostate,

P(0 Tails) = 1/16 = 0.0625
P(1 Tail) = 4/16 = 0.25
P(2 Tails) = 6/16 = 0.375
P(3 Tails) = 4/16 = 0.25
P(4 Tails) = 1/16 = 0.0625

The macroscopic result of 2 tails and 2 heads has the most microscopic ways and this is the most probable macrostate (i.e. for a large number of trials, the average would be 50% heads and 50% tails). The probabilities in the above list describe a **Boltzmann distribution**.

We may say that the macrostate with the most different ways (microstates) has the most **entropy**. As we shall see, this same description (and terminology) may also be applied to a large system of particles where each particle has different configurational (potential) energies.

For each macrostate in the simple example of coins every microstate has the same probability (formally the same potential energy). *In real cases microstates have different energies and the actual probabilities are products of both entropy and energy*.

Now we can state the same ideas in more general terms. Boltzmann assumed that the entropy **S** of a system of atomic particles in a particular state is proportional to the phase space volume **Ω** of that macrostate,

$$S \propto \log\Omega \qquad (1)$$

Phase space is just the collection of different positions and momenta (ways to exist) that a macrostate of particles may have. Boltzmann cleverly broke the continuous phase space up into many small finite cells and called each cell a microstate. The number of *recognizably different* microstates **W** (ways to exist) occupied by a system in a particular macrostate is equivalent to the phase space of that macrostate and is proportional to the entropy,

$$S \propto \log W \qquad (2)$$

Again, we are defining **W** as the number of different microstates which define a macrostate of the system. A macrostate is defined by macroscopic variables such as volume, temperature and pressure. For example, the temperature of a system is an average value over the microscopic particles in the system. We can usually observe and measure macrostate variables - we usually cannot observe and measure microstates (except from molecular simulations).

Boltzmann went on to demonstrate that the macrostate with the most microstates, and therefore the largest entropy, corresponds to the state of **thermodynamic equilibrium**. This is a restatement of the second law of thermodynamics that all systems tend toward maximum entropy.

Furthermore, Boltzmann established the link between statistical properties of atoms (number of ways or microstates) and measurable macroscopic properties such as energies. His famous equation includes a factor, the **Boltzmann constant**, (**k** = 0.33×10^{-23} cal/K), relating entropy to energy

$$S = k \log W$$

where **log** is $\log_e$, now commonly designated **ln**.

For systems where energies of the microstates are different and all macrostates have the same entropies (the opposite of the heads and tails example above), Boltzmann established that the **most probable** macrostate has the **lowest energy**. This relationship is the foundation for molecular simulations and for much of statistical mechanics. It may be simply stated as: **The probability of a molecular state decreases exponentially with the increased energy of the state.**

$$P(\text{state}) \propto e^{-E} \qquad \text{(notice the negative exponent)}$$

In other words, molecular conformations with lower energy will be more frequently populated (assuming they all have the same entropy). We will return to these ideas in more detail many times during the class.

VI. Methods

We will learn two computational methods of molecular simulation commonly used to create and study molecular ensembles: **Monte Carlo** and **Molecular Dynamics** simulations. These methods will be presented at an introductory level but in enough detail to actually implement them in computer code for simple systems. Simulations of simple particles will be performed and the configurational distributions analyzed. We will also carry out molecular simulations on a complex system including a protein in water and analyze the conformational distribution of the protein. However, this is not a course on the full range of important topics in molecular simulations. The emphasis is on the generation of ensembles and structural analysis; advanced methods for free energy calculation (frequent topics in a molecular simulations course) are not covered.

Molecular simulations are attempts to model chemical reality. Simulations must be validated by continued comparison of computer predicted results to experimental results. It is all too easy to fall into the trap of "garbage in; garbage out". We will learn (very briefly) about an experimental technique, **hydrogen exchange**, used to study protein dynamics. The focus for this topic will be to understand the relationship between simulation results and experimental results and how one method serves to validate the other.

Proteins undergo conformational change during their function. Frequently, these conformational changes are influenced by regulatory molecules that bind to the protein and alter the protein conformation (or dynamics), not just at the binding site, but throughout the protein matrix. How is the binding of a regulator at one site communicated to the rest of the protein? We will address this question in both lecture and lab under the topic of **collective motion**.

The ideas learned in this course apply to areas beyond molecular simulations. One of the

main themes of the course is that molecular systems may be studied as average entities that appear to have **deterministic** mechanisms (e.g. move a lever and obtain an action), or the systems can be studied as populations undergoing **stochastic** changes that result in macroscopic actions. We will explore how these two approaches apply to the modeling of molecular reactions as well as molecular conformational changes.

The last laboratory will include exercises in **docking** a ligand to a protein molecule when the ligand is undergoing conformational flexibility. This type of experiment is important in drug discovery.

