Math. Model. Nat. Phenom.
Vol. 6, No. 6, 2011, pp. 278-294
DOI: 10.1051/mmnp/20116615

Using normal mode analysis in teaching mathematical modeling to biology students

D. A. Kondrashov [1]

University of Chicago, BSCD, 924 E 57th St, Chicago, IL 60637, USA

Abstract. Linear oscillators are used for modeling a diverse array of natural systems, for instance acoustics, materials science, and chemical spectroscopy. In this paper I describe simple models of structural interactions in biological molecules, known as elastic network models, as a useful topic for undergraduate biology instruction in mathematical modeling. These models use coupled linear oscillators to model the fluctuations of molecular structures around the equilibrium state. I present many learning activities associated with building and understanding these models, ranging from analytical to computational. I provide a number of web resources where students can obtain structural data, perform calculations, and suggest research directions for independent projects.

Key words: mathematical modeling, elastic network models, normal mode analysis, curriculum development
AMS subject classification: 92B01

1. Role of modeling in modern biology curriculum

Mathematical modeling has played a role in biology for centuries; in the 17th century William Harvey used a simple calculation of the volume of blood ejected by a pumping heart to demonstrate that blood circulated in the body. However, the traditional practice of biology is strongly empirical, with a robust skepticism of theoretical approaches. To many experimental researchers, a quantitative model either agrees with data, in which case it apparently adds no new knowledge, or disagrees with data, in which case it is wrong. This paradigm appears to be shifting in recent years, due to the dramatic explosion of quantitative biological data driven by technological advances in molecular

[1]E-mail: dkon@uchicago.edu

biology, biochemistry, and medicine. The recognition of the necessity of quantitative modeling in research militates for changes in biological curriculum, particularly the improved training of future biologists in mathematical and computational skills. Policymakers have elucidated a vision for a transformed biological curriculum in the BIO 2010 Report, which includes recommendations for greater quantitative and interdisciplinary training for biology students.

One of the main challenges for teaching mathematical and computational methods to students who are not primarily interested in mathematics is convincing them of the relevance of these topics to their interests. The recommendations of the BIO 2010 Report, among many other proposals for biological curriculum reform, specify two approaches to address this issue. The first is the use of real experimental data in teaching quantitative methods. While using data is more difficult for the instructor than using clean mathematical examples, the experience is immeasurably more useful for students who will have to deal with experimental data in order to be scientists. The second approach is to use open-ended projects, akin to research, instead of step-by-step, sanitized assignments. Projects engage students' creativity, and they learn about dealing with the joys and frustrations of doing science.

In this paper I present normal mode analysis of systems of coupled linear oscillators applied to the study of flexibility of biological macromolecules. I have found these models to be useful for illustrating and learning a number of quantitative skills. Among the areas of mathematics used in normal mode analysis are: differential equations, matrix diagonalization, variance and covariance between random variables, and numerical linear algebra. The suggested projects allow students to explore biological data bases, such as the Protein Data Bank, download their own data, and conduct computational experiments with unknown results, that is, conduct their own research. Normal mode analysis could make a good unit of study in a course of mathematical modeling aimed at biology students.

2. Normal mode analysis of biological macromolecules

The linear spring potential is used to model interactions in numerous natural situations. Compression waves in air or other materials are well represented by linear oscillations, which are used to predict the frequencies of sounds that have the longest lifetime and the largest amplitude, given the shape of the acoustical cavity. One can predict which oscillatory frequencies are favored by solid objects, such as bridges and buildings, and thus try to prevent the destructive effects of resonance. In chemistry, the bonds between atoms are represented by linear springs to predict the natural frequencies of bending and stretching, which allows for experimental identification of molecules using vibrational spectroscopy, such as IR (infrared) and Raman. Normal modes are a part of the educations of chemists [5], but have not been commonly used in biological curricula. In this section, I introduce protein structures and describe the role of normal modes in biophysical research.

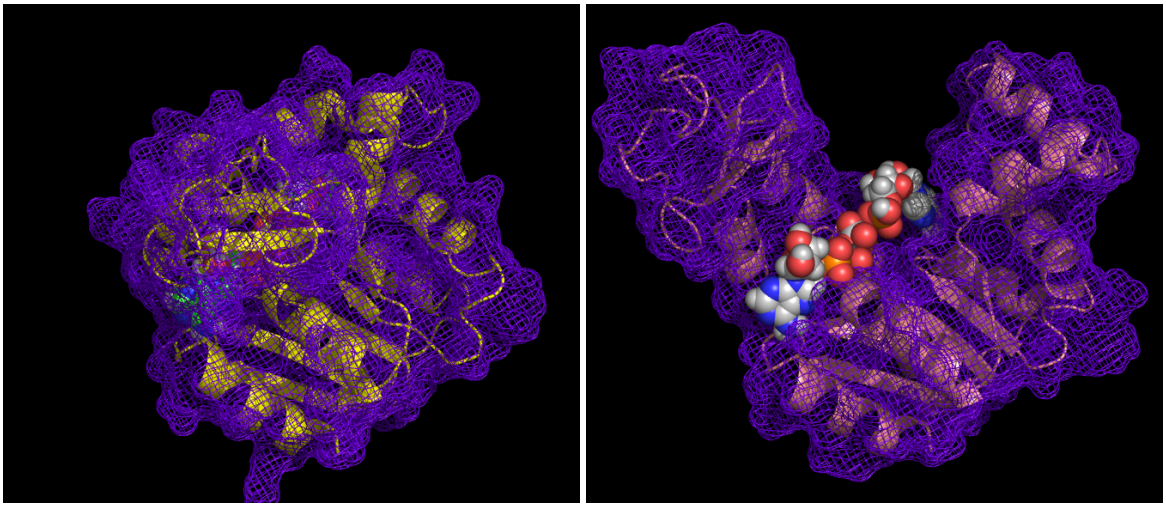

Figure 1: Representation of the structure of two states of the protein adenylate kinase: a) closed; b) open, with substrate shown in the active site.

2.1. Biomolecular structures are flexible

Proteins are polymers of amino acids, which are linked by covalent bonds into an unbranched chain. The length of the sequence and the identity of amino acids (there are 20 different types in all life forms) is encoded by the DNA sequence of a given gene, which is then transcribed into an RNA sequence, from which the protein is composed by the ribosome. Once the chain of amino acids is linked, the protein arranges itself (sometimes with assistance from chaperones) into a precise three-dimensional arrangement of its constituent amino acids, which in turn consist of atoms. The coordinates of all the atoms comprising a protein is called the protein's structure. It is a remarkable fact that, with some exceptions, the sequence of amino acids determines the unique three-dimensional structure of the chain. In fact, these structures are generally quite robust to changes in the amino acid sequence, and groups of related proteins from different organisms, called protein families, tend to share the same structural geometry, known as protein folds.

Protein structures are determined experimentally, most commonly using X-ray crystallography, and deposited in the publicly accessible Protein Data Bank (`http://www.pdb.org/`) in the form of PDB files, which contain three dimensional coordinates of a protein's atoms. Determining the structure of a protein is laborious, although thanks to technological advances, structural determination is less difficult than in the past. The knowledge of a protein's structure provides a great deal of important information to biochemists, for instance the location of the catalytic active site of an enzyme. However, the structure does not tell the whole story of how a protein functions.

The structures of proteins and other biomolecules are not static. The mobility of atoms inside a protein is somewhere between that of a liquid and a solid, with atoms on the exterior more mobile, and those in the center (core) of a protein resembling a solid. The atomic coordinates reported in the PDB actually represent the mean positions of atoms, but the proteins in a reasonable physiological environment undergo substantial fluctuations about the mean positions. Furthermore, a change in

the environment, e.g. a shift in pH, or binding of a substrate to a ligand, may induce a change in the mean atomic positions, called a conformational change.

Conformational changes often play important functional roles, for instance by admitting the substrate into the active site of an enzyme, where the catalyzed reaction takes place. One such example is shown in Figure 1, where the enzyme adenylate kinase is shown in the closed state, in which its active site is not accessible, and the open state, which allows both entry for the substrate and exit for the products. Unfortunately, conformational changes in proteins are difficult to observe directly due to the rates of the transitions, which are usually in the microsecond to nanosecond range. Sometimes crystallographers can observe the two endpoints of a conformational change by trapping the protein in different conformations. This is accomplished by changing the conditions of crystallization or adding substrates; the two structures of adenylate kinase in Figure 1 were obtained in that way.

The sequence of amino acids, which is encoded by DNA, defines the properties of a protein. Based on this sequence, the protein folds into a specific structure. However, structural flexibility is necessary for a protein to perform its function. A particular protein not only forms a particular shape, but this structure also undergoes particular motions, which are also essential characteristics of that sequence. While structures may be determined experimentally, dynamics are much harder to observe, and therefore are modeled using computational methods.

2.2. Elastic network models for biomolecules

Biological macromolecules are complex systems of thousands of atoms, each interacting with each other and with the surrounding waters, ions, and other macromolecules. All these interactions may be faithfully simulated computationally by adding every single bond, electrostatic interaction, and other physical forces, and then proceeding in tiny time steps to produce a movie of the molecular motion. These molecular dynamics simulations are generally successful, but are very expensive to run, and complicated to manage and understand. Instead, we will use very simple models of interactions within a biological molecule to try to simulate conformational changes around the native state.

A protein molecule can be thought of as a system with a potential energy function, composed of the interactions between all the atoms in the molecule, and with the surrounding solvent. The state variables of the system are the positions of all the atoms, which means that the system has thousands or tens of thousands of variables. In equilibrium, one expects the system to fluctuate around the minimum of that potential function, which is presumed to be the experimentally determined structure, also known as the "native state" of the protein. Thermal noise adds random kinetic energy to the system, causing the conformations to fluctuate around the native state. By Taylor expansion, the fluctuations near the equilibrium occur in a roughly quadratic well.

Harmonic approximations to the potential well have been used in the study of flexibility of molecular structures, in order to simplify the complexity of the system. This assumption makes physical sense at the bottom of the potential well, near the native conformation, where the potential must have close to quadratic shape, and therefore the restoring forces are nearly linear with displacement. This scenario implies that the molecules behave as coupled harmonic oscillators, with

each atom connected to other atoms by harmonic potentials. Various models exist for defining the interaction between different particles within a protein structure. The connections may be based on physical chemical forces, such as chemical bonds, van der Waals forces, and electrostatic interactions, or may be based on a simple model where parts of the protein in proximity are assumed to interact as if bound by a linear spring.

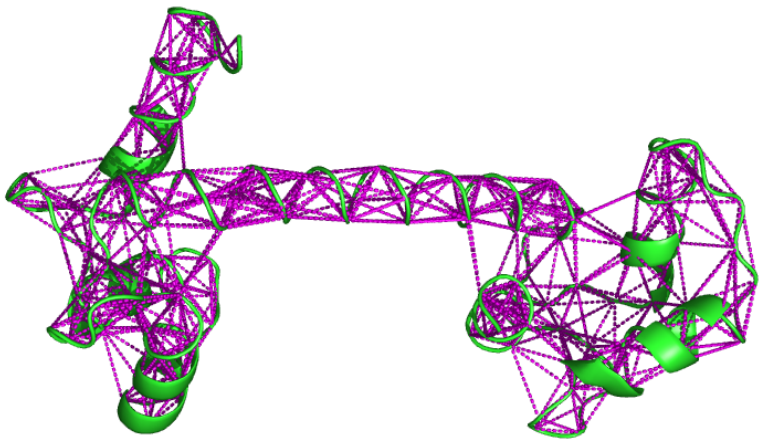

Figure 2: Harmonic potential model of the protein calmodulin. Green indicates the backbone of the molecule, maroon lines indicated harmonic interactions between residues [6].

A simple class of models, known as elastic network models (ENM), was first proposed for studying protein conformational dynamics by Tirion [12]. The potential is constructed as follows: any pair of atoms closer than a cutoff distance are coupled by a harmonic spring, while those farther away have no interaction at all. The approach was borrowed from materials science models of elastic materials, such as rubber, in which the interactions between neighboring atoms are determined by their proximity. This assumption is reasonable for simple polymers, composed of identical short monomers, but it seems to grossly oversimplify the description of proteins. In these models, the only determinant of interaction is the geometry of the folded protein, and thus much of the chemical information is discarded. Despite this limitation, many researchers have used these models, with a considerable amount of success, to predict and understand the conformational changes of complex proteins [3]. Figure 2 shows the harmonic potentials used to model the structural dynamics of the protein calmodulin, using a variation on the ENM called the Gaussian network model (GNM) [2], which further reduces the protein to a set of nodes corresponding to the amino acid residues rather than atoms, as in the original ENM.

Once the interactions are set up, one can calculate the collective modes of motion for the system of coupled oscillators. These normal modes and the corresponding frequencies are determined by computationally finding the eigenvectors and eigenvalues of the Hessian matrix of the potential function [4]. For practical purposes, the most interesting modes are those with lowest (but nonzero) frequencies, because they correspond to the slowest and most global collective motions, as opposed to high-frequency vibrations, which are restricted both in amplitude and in scope. Intuitively, the

lowest frequency modes correspond to the shallowest directions in the potential energy well. Given a reasonable amount of thermal noise, the protein structure is most likely to be deformed along the shallow directions, instead of climbing up the steep directions.

The utility of normal mode analysis of biological molecules lies in obtaining the preferred modes of flexibility from a static structure, which allows biochemists to better understand the mechanism of the molecular function. For instance, in studying the mechanism of opening or closing of an enzyme binding site, normal modes can generate a hypothesis about the intermediate conformations, and help predict which residues play a key role. Figure 3 shows the directions of the lowest frequency mode of calmodulin, which undergoes a large conformational change in response to binding of calcium ions. The arrows show the extent of involvement of each amino acid residue, as well as the direction of preferred fluctuations. Simple elastic network models may be used to decompose the fluctuations of atomic posititions in terms of collective normal modes of motion, which simplifies the systems and generates predictions relevant for understanding the function of biomolecules [3].

3. Learning activities with normal modes

Modeling biological molecules as systems of coupled harmonic oscillators involves many kinds of quantitative skills. The theory requires three areas of mathematics: differential equations, linear algebra, and basic probability to understand the concept of variance and covariance. In order for students to implement the theory and analyze the flexibility of a macromolecule, they need to be able to perform numerical diagonalization and inversion of matrices, make informed decisions about setting up the model, and use databases of molecular structures to read input files. All of these skills are accessible to biology undergraduates, and normal mode analysis of biological macromolecules provides a context in which they can learn some or all of them. Below, I suggest some independent projects in which the students can gain exeperience with the mathematical and numerical topics.

3.1. Differential equations for linear oscillators

Harmonic oscillator modeling begins with the equation of motion for a Hookean spring with force constant k, an object with mass m, with x the displacement from the equilibrium position:

$$m\frac{d^2x}{dt^2} = -kx. \tag{3.1}$$

Students can discover that the solution is composed of periodic functions, specifically $x(t) = A\sin(\omega t) + B\cos(\omega t)$, where the frequency $\omega = \sqrt{k/m}$ and the weighting constants A and B are determined by two initial conditions, typically $x(0)$ and $dx/dt(0)$. Because the sine and cosine with the same frequency are equivalent with a $\pi/2$ phase shift, the solution can also be represented by a single sine or cosine function with a phase shift ϕ determined by the initial conditions: $x(t) =$

$A\sin(\omega t + \phi)$. In either formulation, the deviation from equilibrium of the harmonic oscillator is a periodic oscillation with a constant amplitude, as there is no friction.

To make things more interesting, consider a model describing the dynamics of two objects connected by a spring with constant k_c, and connected to a fixed external object by a spring with constant k. As above, the forces are linear, with the coupling between two masses depending on the difference in the deviations $(x_1 - x_2)$, while the coupling to the external objects only depend on the individual deviations x_1 and x_2. This results in the equations of motion that are a linear system of second-order ODEs, assuming both masses are equal:

$$\begin{aligned} m\frac{d^2x_1}{dt^2} &= -kx_1 + k_c(x_2 - x_1) \\ m\frac{d^2x_2}{dt^2} &= -kx_2 + k_c(x_1 - x_2) \end{aligned} \tag{3.2}$$

These equations can be expressed in a more concise form by using matrices and vectors. If the displacements of the two masses are written as components of a single vector $\vec{x}$, then the two equations can be written as a single matrix equation:

$$m\frac{d^2\vec{x}}{dt^2} = -H\vec{x} \tag{3.3}$$

$$\text{with } H = \begin{pmatrix} k_c + k & -k_c \\ -k_c & k_c + k \end{pmatrix}.$$

The students can see that this equation is analogous to equation (3.1) above, with the difference that the matrix H now plays the role of the scalar spring constant k. The solution is still going to be oscillatory, with frequencies determined by the spring constants, but we need to use linear algebra to express them.

3.2. Linear algebra

The equations of motion for coupled oscillators are a linear, coupled ODE system, which may be expressed in matrix form. The displacements of N coupled oscillators can be written down as a vector of dependent variables $\vec{x}$. The matrix which consists of the spring constants for forces coupling each pair of oscillators can be described as the Hessian matrix of the potential function $V(\vec{x})$, that is the matrix of second partial derivatives with respect to all N variables:

$$H_{ij} = \frac{\partial^2 V(\vec{x})}{\partial x_i \partial x_j}.$$

Once given the Hessian matrix, one can write the equations of motion for any set of coupled oscillators in the form of equation (3.3). Then the power of linearity allows us to decompose the solution in terms of N different terms, each one corresponding to an eigenvector of the matrix H. Decomposing a displacement vector in the basis of eigenvectors reduces the N-dimensional equation to N one-dimensional ones, as follows. Suppose that a given displacement vector is

colinear to an eigenvector of H, $\vec{y} = c\vec{u}$. Then by definition of an eigenvector, $H\vec{u} = \lambda\vec{u}$, and the matrix equation turns to a scalar:

$$m\frac{d^2\vec{y}}{dt^2} = -H\vec{y} = -\lambda\vec{y}.$$

We saw above that the solution of the one-dimensional equation is a sine, with frequency $\omega = \sqrt{\lambda/m}$. This is true for any of the N eigenvectors of H, which all represent oscillations with different frequencies, determined by their corresponding eigenvalues. As long as we can write an initial displacement vector as a linear combination of the eigenvectors, $\vec{x}_0 = \sum c_i\vec{u}_i$, the solution can be written as a linear combination of oscillations:

$$\vec{x}(t) = \sum_i^N c_i\vec{u}_i \sin(\omega_i t + \phi_i). \tag{3.4}$$

The eigenvectors of the Hessian of a system of coupled linear oscillators are called *normal modes*. Each one describes a *collective vibrational motion* with a particular frequency. The normal modes contain coefficients that correspond to displacements of each oscillator. These coefficients describe the relative displacements for each mass, since eigenvalues are invariant to multiplication by a constant; the magnitude of each eigenvector is determined by the initial conditions. The frequencies of the collective oscillations are determined by the eigenvalues of the normal modes, as follows: $\omega = \sqrt{\lambda/m}$.

Example: Let us return to the model with two coupled masses described above. The eigenvalues of its Hessian matrix can be found by solving the characteristic equation, and are $\lambda_1 = k$ and $\lambda_2 = k + 2k_c$. The corresponding eigenvectors are:

$$\vec{u}_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \lambda_1 = k;\ \vec{u}_2 = \begin{pmatrix} 1 \\ -1 \end{pmatrix} \lambda_2 = k + 2k_c$$

These two eigenvectors describe two different collective motions. The first eigenvector stands for a motion with identical, parallel displacements of the two masses. The second eigenvector represents a motion with anti-parallel displacements with equal magnitudes of the two masses. Notice that the first normal mode has a lower vibrational frequency than the second.

Example: Consider three nodes connected as a linear chain, with node 1 connected to node 2 with force constant k_1, and node 2 connected to node 3 with force constant k_2. Notice that this system is not tethered to an external object. The Hessian matrix for this model is:

$$H = \begin{pmatrix} k_1 & -k_1 & 0 \\ -k_1 & k_1 + k_2 & -k_2 \\ 0 & -k_2 & k_2 \end{pmatrix}.$$

For this system of three nodes, let both springs have the same force constant of 1 ($k_1 = k_2 = 1$). Then the system has the following eigenvectors and eigenvalues:

$$\vec{v}_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix},\ \lambda_1 = 0;\ \vec{v}_2 = \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix},\ \lambda_2 = 1;\ \vec{v}_3 = \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix},\ \lambda_3 = 3$$

This demonstrates that a linear chain of three oscillators that are not attached to any other object has three normal modes of different frequencies. The first mode has zero frequency, which is known as a *rigid-body* mode, in which the entire system moves together, without changes in relative distances. This happens because the system is not coupled to any external object, and can thus move as a rigid body without stretching any of the springs. The second mode has frequency 1, and in it the two end nodes move in opposite directions of each other. The third mode has frequency $\sqrt{3}$, and in it the end points move in the same direction, while the middle node moves in the opposite direction with twice the amplitude. This analysis predicts that all motions of the three nodes can be described in terms of the three normal modes.

3.3. Covariance and normal modes

The potential energy function of any system of harmonic oscillators has a particular form: it consists of quadratic terms, which together are known as a quadratic form. Any such potential function can be written in terms of its Hessian matrix as follows:

$$V(\vec{x}) = \vec{x}^T H \vec{x}.$$

Multiplication of H on both sides by the vectors of displacements results in a scalar function with second-degree terms of the displacements.

Above, we have solved the equations of motion of coupled oscillators, to obtain precise, deterministic positions for each mass. However, the biological objects that we would like to model, macromolecules, are subject to a barrage of impacts from thermal motion, both of themselves and their environment. Therefore, we pose the question in stochastic formulation: can we describe the fluctuations of N coupled harmonic oscillators, subject to what is called a heat bath in thermodynamics?

Statistical physics provides a general answer to this question. The heat bath is defined as a source of random, independent and identically distributed impulses with a normal probability mass function. The width of this normal distribution of impulses is determined by the temperature T, which in classical thermodynamics is proportional to the mean kinetic energy of particles in equilibrium with a heat bath, with proportionality constant k_b known as the Botlzmann constant: $Tk_b = \frac{1}{2}\langle mv^2 \rangle$, where v is the velocity of a particle, m is mass, and brackets denote ensemble mean.

For a system subject to a potential energy function $V(\vec{x})$, and coupled to a heat bath with temperature T, the equilibrium probability density function is given as follows:

$$\rho(\vec{x}) = Z^{-1}(T) \exp\left(-\frac{V(\vec{x})}{k_b T}\right).$$

This is known as the Boltzmann distribution, in which k_B is the Boltzmann constant and $Z(T)$ is a normalization constant that is known as the partition function, to make the integral of the density function is 1 over the whole space of $\vec{x}$.

If we substitute the potential energy function with the quadratic form into the Boltzmann equation above, we obtain a form of the probability density function which is the canonical Gaussian distribution in N dimensions:

$$\rho(\vec{x}) = Z^{-1}(T) \exp\left(-\frac{\vec{x}^T H \vec{x}}{k_b T}\right). \tag{3.5}$$

We know a lot about the properties of Gaussian functions, and in particular their statistics. The single-variable Gaussian, $\rho(x) = Z^{-1} \exp(-hx^2)$ has mean 0 and variance $2/h$. When dealing with the multidimensional Gaussian in equation (3.5), there is an analogous situation, but with matrices instead of scalars. To find the variance-covariance matrix Σ of the displacement $\vec{x}$, one needs to invert the Hessian matrix H:

$$\Sigma = \frac{2}{k_b T} H^{-1}$$

This result will be used below to compute the degree of covariance of different masses in a system of coupled of oscillators subject to random thermal noise.

3.4. Modeling and parametrization

In the first three subsections we were able to express the solutions of differential equations describing coupled harmonic oscillators in terms of eigenvectors and eigenvalues of the Hessian matrix, and to connect this solution with the variance-covariance matrix for oscillators receiving random, uncorrelated kicks from a heat bath. In order to actually perform these calculations, we need to do two things: a) construct the Hessian matrix, and b) diagonalize it. In this subsection we will describe the first task, and then describe the second one in the following.

We assume that we have a system of N nodes, which in the case of biomolecules may refer to atoms, amino acid residues, or another structural unit. These nodes are coupled via harmonic potentials in some fashion. Let us take as an example a simple Elastic Network Model, known as the Gaussian Network Model [2], in which pairs nodes closer than a certain cutoff distance are coupled with Hookean potentials with a uniform constant k. This kind of model is illustrated in Figure 2, where the maroon lines indicate the harmonic interactions used in the GNM. We must start with a data file containing the coordinates of each node, such as a PDB file of a protein structure, which will be discussed in subsection 3.6. Given the positions of all the nodes, which are typically the carbon alpha atoms of each amino acid residue, the students can write code in a programming language of their choice to construct the Hessian matrix as follows:

constructing the Hessian matrix for the Gaussian Network Model

- obtain list of coordinates for N nodes $X_i = (x_i, y_i, z_i)$
- set cutoff distance R
- define a distance function $\texttt{dist}(X_i, X_j)$ that returns the distance between two 3-dimensional vectors

- initialize N by N matrix H with 0 values
- **for** i from 1 to N

  for j from 1 to $i-1$

  if ($\texttt{dist}(X_i, X_j) < R$)

  $H(i,j) \leftarrow -k$

  $H(j,i) \leftarrow -k$

  $H(j,j) \leftarrow H(j,j) + k$

  end if statement

  end for loop

- **end** for loop

After writing their own program for building a Hessian, the student may be given the freedom to modify and improve the model. It has been demonstrated that using different values of force constants for residues which are covalently bonded and those which are not, and are placed in proximity by the process of protein folding, results in better agreement between predicted and experimental variances of fluctuation [6]. The students may experiment with choosing different parameter values for $k_{i,j}$ based on inter-residue distance, types of residues, types or number of atoms in contact, to name some possibilities. Then they may observe the effect the changes in parametrization makes on the calculations described in the next subsection.

3.5. Numerical calculations using normal modes

Once the Hessian matrix has been constructed, the next step is to perform diagonalization to find the normal modes and their frequencies. For a biology student learning mathematical methods, it is appropriate to use existing implementations of numerical diagonalization algorithms, which can accessed by built-in functions in computational platforms such as Mathematica or Matlab. In a course with an objective of learning numerical methods, the student may learn to implement numerical eigenvalue algorithms, for instance the QR algorithm, which is described in numerous sources, for instance in [10]. Let us postulate that we have found a proper diagonalization of the Hessian matrix H, consisting of the matrix U with eigenvectors $\vec{u}_i$ as the columns, and the diagonal matrix Λ with eigenvalues λ_i as the diagonal elements:

$$H = U\Lambda U^{-1}.$$

The eigenvectors need to be sorted by size of eigenvalues, in ascending order. We now possess the normal modes of the system sorted by frequency. There may be several eigenvectors with zero eigenvalues, corresponding to rigid-body motions. For instance, in a one-dimensional set of coupled oscillators, which are not coupled to any external object, there is a single rigid-body degree of freedom, translation along the line. This is the case for the distance-based Gaussian Network

Figure 3: The predicted directions and magnitudes of flexibility of calmodulin from the lowest frequency mode of a three-dimensional Elastic Network Model [7].

Model that we used above to construct the Hessian matrix because it has only a single dimension - distance between the nodes.

After excluding the zero-frequency eigenvectors, the remaining normal modes can be used for both qualitative and quantitative analysis. Many researchers focus on one or two normal modes with the lowest frequencies, and use them to predict the preferred magnitudes, and in the case of three-dimensional models, directions of flexibility for all parts of a macromolecule. Figure 3 illustrates the involvement of all residues of the protein calmodulin in the lowest-frequency normal mode. While there have been a number of cases of agreement between observed conformational changes in proteins and individual lowest-frequency modes, this approach lacks a solid theoretical foundation [13]. It may be instructive for students to plot the displacements predicted by lowest frequency modes for all the nodes, and to observe which parts of the molecule are most involved. The nodes which have zero or small displacements are often identified as "hinge" regions of protein structures, which have been found to be statistically likely to be located near the catalytic active centers of proteins [14].

A mathematically sound use of normal modes is to find the covariance structure of the fluctuations in the set of coupled oscillators. As described in subsection 3.3., the Hessian matrix has an inverse relationship to the variance-covariance matrix of the multidimensional Gaussian process describing random fluctuations in the harmonic potential. In order to compute the covariance matrix, however, one needs to find the pseudo-inverse, because the Hessian has zero eigenvalues. The pseudo-inverse can be calculated as a sum of contributions from all the normal modes indexed by k, with non-zero eigenvalues λ_k, and where $u_{i,k}$ are the i-th components of the k-th normal mode:

$$\Sigma_{i,j} = \langle \Delta x_i, \Delta x_j \rangle = \sum_k \frac{1}{\lambda_k} u_{i,k} u_{j,k}.$$

This formula generates the covariance between any two nodes i and j, including in particular the variance for mode i, if $i = j$. Notice that the contribution of each mode is inversely proportional to the eigenvalue, thus the lowest-frequency modes make the greatest contribution to the covariance calculation. This means that a subset of normal modes (whose number is equal to the number of nodes, which can easily range into thousands for large proteins) may be sufficient to obtain a good approximation of the variance-covariance matrix. This information can be used to quantify the degree of dynamic coupling between different parts of a protein structure.

3.6. Using Web resources for normal mode analysis

There are valuable resources online that allow students to obtain their own data, and even perform normal mode calculations, as long as they known how to interpret the results. Depending on the aims of the course, the instructor may prefer to have students using these resources for some, or all, of their computational work.

If one wants to study the flexibility of proteins, the most crucial resource is the Protein Data Bank, at `http://www.pdb.org/`. It is the general repository of all solved protein structures in the world, and contains around 50,000 structures which are available to all for free. A protein structure, in the form of a text file that contains the coordinates of the atoms (called a PDB file), is necessary as the starting point for normal mode calculations, as we saw above. Moreover, the protein data bank contains multiple structures of many proteins, determined under different conditions, which frequently show conformational changes, such as the one for Adenylate Kinase in Figure 1.

These conformational changes can be used to test and calibrate models of protein flexibility. To this end, Gerstein et al compiled pairs of structures that exhibit conformational changes, and created a library of conformational changes, classified by type [8]. This database, which may be found at `http://molmovdb.org/`, can be a useful source of information for students who want to test the predictions of normal mode models against experimentally observed conformational changes, and may suggest research projects on understanding the differences between different kinds of conformational changes.

There are also a number of Web-based servers for computing the normal modes of a given protein structures. The iGNM server, at `http://ignm.ccbb.pitt.edu`, provides normal mode calculations based on the Gaussian Network Model described above [15]. Another server, at `http://igs-server.cnrs-mrs.fr/elnemo`, performs calculations for a given structure based on a 3-dimensional elastic network model of Sanejouand, el al [11]. This model returns not only the predicted magnitudes of fluctuations for each node, but vectors of deviations, indicating direction of collective fluctuations in each normal mode.

3.7. Outline of a research project

I conclude with a sample research project that can get students started on generating their own normal modes of proteins. This will enable them to learn the linear algebra and normal mode concepts actively, and leave them the creative freedom to design their own connectivity models.

Below are the steps that students with basic programming skills can take in order to write their own normal mode analysis code.

- **Simple system.** Start by writing a code to generate a Hessian for two oscillators coupled by a spring, with the force dependent on the difference between the two positions, as in equation (3.2). Start without any external coupling ($k = 0$) and with only the coupling spring k_c, and find the eigenvectors and eigenvalues of the Hessian. Then add external coupling and observe that the zero eigenvalue becomes positive, and connect the frequencies of the two normal modes with the collective vibrational motions.

- **String of coupled oscillators.** Write a function that outputs a Hessian for a linear string of coupled oscillators. The Hessian is a tridiagonal matrix, with $2k_c$ on the diagonal and $-k_c$ on both off-diagonal neighbors; for simplicity you can set $k_c = 1$. The two ends may be tethered by external coupling k, eliminating zero-frequency modes. Ask the students to generate and plot the eigenvectors, which for large enough N will look like smooth sinusoidal curves with m maxima and minima for the m-th lowest frequency modes. These normal modes correspond to solutions of the continuous wave equation in the limit of large N, and can be connected with acoustical modes (harmonics) of a plucked string.

- **Distance-based connectivity matrix.** Write a function that takes in N 3-dimensional coordinates, and returns a connectivity matrix, of size N by N. Follow the pseudocode in section 3.5 to generate it. Then diagonalize this matrix, and arrange the normal modes (eigenvectors) in order from lowest to highest frequency. Use some test cases, e.g. a Pac-Man type set of coupled oscillators, and plot the lowest normal modes vs the oscillator index. The plot will illustrate the relative displacements of each mass in the set of coupled oscillators, although not the directions, since this is strictly a distance-based model. For a Pac-Man type shape, the lowest frequency mode will have a minimum at the hinge and maxima at the tips of the two "jaws".

- **Application to protein structure.** Using the code developed in the previous step, the students can now produce normal modes for a protein structure. Go to the PDB (`http://www.pdb.org/`) and find the protein structure you wish to analyze. For example, for adenylate kinase from *E. coil*, type PDB code 4AKE. Download the text file in PDB format containing the types of atoms and coordinates. Before using it, one needs to some processing of the PDB file, usually reducing it to only carbon α atoms (which are denoted CA in PDB format). Use your favorite text editor to select only lines with atom type CA; this will be your data input file. Read in the file into your program, such that the 7, 8 and 9th columns are read in as the x, y, and z coordinates. Now you can use the code from the last step to find the normal modes of the simplified structure.

- **Normal mode analysis.** Now that normal modes have been generated, there are several ways they can be analyzed for biophysical information. The first few (1-3) lowest frequency modes can be used to predict the predominant conformational changes in the protein structure. For instance, in Adenylate kinase mentioned above, simple distance based normal mode model

predicts very well the conformational transition between the open and closed states (shown in Figure 1). In order to quantify this, go to the PDB and find the structure of adenylate kinase in the closed state (PDB code 1ANK). As above, reduce it to the CA atoms, and find the distances between CA atoms in the open state (4AKE) and the closed state (1ANK). Then find the dot products between the normal modes, starting with the lowest frequency, and the vector of distances. The normal modes form an orthonormal basis of the N-dimensional vector space, and the dot products are the projection coefficients for the difference vector. The students should see that the lowest frequency mode accounts for the lion's share of the conformation change, with the inner product around 0.8, and that higher frequency normal modes quickly become irrelevant.

The second observation of interest is finding the "hinges" of the protein motion. Plot the lowest frequency normal modes for a protein structure, and note for which residue values the plot crosses zero (in deviation). These residues correspond topologically to the hinges in a Pac-Man like jaw opening motion, and are frequently the site of functional significance for a protein [14], such as the enzymatic active site in adenylate kinase.

- **Further directions.** Now that the students have developed some skills, they can develop their own creativity in assessing or improving distance-based normal mode models. Here are some ideas:

  Investigate the lowest-frequency modes in a protein family. How does the shape of the normal mode (comparison can be done by inner product) correspond to percent sequence identity between two proteins?

  Experiment with different elastic network models. What is the effect of including more or less information in the model? Read in every atom in a protein structure file, and compare the overall normal mode shape with that of only CA atoms, or one with every other CA atom, or every tenth CA atom? Comparing the normal modes is less straightforward when the dimension of vectors is different; one way is to discard points (atoms, oscillators) from the normal mode vector with more elements, and leave only the atoms that correspond to the ones in the smaller vector.

  Introduce directions into the model. It is possible to extend the simple distance-based model to incorporate the direction of motion. The simplest model, called Anisotropic Network Model is presented in [1]. The procedure results in a $3N$ by $3N$ Hessian matrix, and the normal modes have dimension $3N$ for N input atoms, in which the first 3 entries represent the x, y, and z displacements for the first atom, and so on. These types of models predict the direction of displacement as well as its magnitude, and the resultant lowest frequency modes can be compared with conformational transitions, like the one in adenylate kinase.

4. Conclusions

I have presented the mathematical modeling technique of normal mode analysis, in the context of modeling the interactions within biological macromolecules. The models provide an assortment of mathematical and computational topics that are important for biology students to learn, whether or not they are interested in molecular structures. Further, the ease of setting up and modifying the elastic network models allows the students creativity in choosing their own parameters and designing their own computational experiments. While these models are extremely simplified versions of reality, and their applicability to protein dynamics remains controversial [9], they provide a compelling example of applying a simple mathematical idea to model complex biological objects.

Acknowledgements

The author would like to thank George Phillips and Qiang Cui for insightful discussions, and two anonymous reviewers for constructive suggestions. I am also grateful to José Quintáns, Master of the Biological Sciences Collegiate Division at the University of Chicago for encouraging innovative curriculum design in mathematical modeling for biologists.

References

[1] A. Atilgan, S. Durell, R. Jernigan, M. Demirel, O. Keskin, and I. Bahar. *Anisotropy of fluctuation dynamics of proteins with an elastic network model*. Biophysical Journal, 80 (2001), 505–515.

[2] I. Bahar, A. R. Atilgan, and B. Erman. *Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential*. Folding and Design, 2 (1997), 173–181.

[3] I. Bahar and A. Rader. *Coarse-grained normal mode analysis in structural biology*. Current Opinion in Structural Biology, 15 (2005), 586–592.

[4] Q. Cui and I. Bahar. *Normal Mode Analysis: Theory and Applications to Biological and Chemical Systems*. Chapman and Hall/CRC, 1 ed., 2005.

[5] J. L. Dunn. *A pictorial visualization of normal mode vibrations of the fullerene (C60) molecule in terms of vibrations of a hollow sphere*. Journal of Chemical Education, 87 (2010), 819–822.

[6] D. A. Kondrashov, Q. Cui, and G. N. Phillips, Jr. *Optimization and evaluation of a coarse-grained model of protein motion using X-Ray crystal data*. Biophysical Journal, 91 (2006), 2760–2767.

[7] D. A. Kondrashov, A. W. Van Wynsberghe, R. M. Bannenl, Q. Cui, and G. N. Phillips, Jr. *Protein structural variation in computational models and crystallographic data*. Structure, 15 (2007), 169–177.

[8] W. G. Krebs, V. Alexandrov, C. A. Wilson, N. Echols, H. Yu, and M. Gerstein. *Normal mode analysis of macromolecular motions in a database framework: Developing mode concentration as a useful classifying statistic*. Proteins: Structure, Function, and Genetics, 48 (2002), 682–695.

[9] L. Orellana, M. Rueda, C. Ferrer-Costa, J. Lopez-Blanco, P. Chacon, and M. Orozco. *Approaching elastic network models to molecular dynamics flexibility*. Journal of Chemical Theory and Computation, 6 (2010), 2910–2923.

[10] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. Numerical recipes: The art of scientific computing. Cambridge University Press, Cambridge, 3rd ed, 2007.

[11] K. Suhre and Y. Sanejouand. *Elnemo: A normal mode web server for protein movement analysis and the generation of templates for molecular replacement*. Nucleic Acids Research, 32 (2004), W610–W614.

[12] M. M. Tirion. *Large-amplitude elastic motions in proteins from a single-parameter atomic analysis*. Physical Review Letters, 77 (1996), 1905–1915.

[13] A. W. Van Wynsberghe and Q. Cui. *Interpreting correlated motions using normal mode analysis*. Structure, 14 (2006), 1647–1653.

[14] L. Yang and I. Bahar. *Coupling between catalytic site and collective dynamics: A requirement for mechanochemical activity of enzymes*. Structure, 13 (2005), 893–904.

[15] L. Yang, X. Liu, C. J. Jursa, M. Holliman, A. Rader, H. A. Karimi, and I. Bahar. *iGNM: A database of protein functional motions based on gaussian network model*. Bioinformatics, 21 (2005), 2978 –2987.