Studying Protein Flexibility in a Statistical Framework: Tools and Databases for Analyzing Structures and Approaches for Mapping this onto Sequences

W. G. Krebs[1], J. Tsai[3], Vadim Alexandrov[1], Nathaniel.Echols[1], Jochen Junker[1], Ronald Jansen[1] and Mark Gerstein[1,2,4]

[1]Department of Molecular Biophysics & Biochemistry
[2]Department of Computer Science
Yale University
P.O. Box 208114
New Haven, CT 06520-8114, USA.

[3]University of Washington

[4]Corresponding Author

Version 20010925

Abstract

We describe how protein flexibility can be studied statistically in a database framework. The Database of Macromolecular Movements, which is accessible over the Internet at http://molmovdb.org, organizes a few hundred well-characterized motions on the basis of size and then packing, with the involvement of a well-packed interface in the motion being a key classifying feature. It also contains ~4000 putative motions from automatic comparisons on the whole PDB. Systematic literature searches suggest that the "universe" of studied motions is no more than twice this. We survey the computational tools employed in the database analysis. (i) Structure comparison is useful to align and superpose different conformations. (ii) Interpolation, which is implemented on a large-scale by the morph server, provides movie-like pathways between two superposed conformations and in the process generates many standardized statistics. (iii) Normal mode analysis provides readily interpretable information on the flexibility of a single conformation. (iv) And Voronoi volume calculations provide a rigorous basis for characterizing packing. Finally, we explain how the structural features in the motions database can be related to sequence, an important part of the overall process of transferring annotation to uncharacterized genomic data. This allows determination of a sequence propensity scale for amino acids to be in linkers, in general, or flexible hinges, in particular. Preliminary calculations show more proline and less alanine and tryptophan in linkers.

Introduction

When scientists first started to grapple with proteins they were thought of largely as a black boxes—as catalysts that act in some "magical" way to achieve catalysis. One of the main contributions of structural biology has been to illuminate this magic and show how proteins work in structural terms. Often this is through mechanical actions and movement of various domains and parts (Figures 1A and 1B). Thus, the conceptual framework of mechanical engineering may be as appropriate as the traditional, black-box statistical mechanical perspective to understanding these molecules. A wonderful example of is the work by Sigler et al. on GroEL. They discussed GroEL as a two-stroke engine[1] in analogy to the development by James Watt in 1769 of the double-action steam engine.

James Watt's major contribution was the development of valves allowed the steam to drive a piston in both directions, which allowed a ten-fold improvement in energy efficiency over the previous Neucomen single-stroke design—an improvement that finally made steam locomotion practical. In the case of GroEL, this two-stroke cycle is achieved through an anti-coordinated allosteric motion between the upper and lower rings (a different set of allosteric motions allows a coordinate binding of ATP within each ring). Future protein engineers might think of GroEL as a model to use in modifying the design of other protein ATPase 'engines' so as to convert them from the more common one-stroke cycle to a two-stroke design and thereby increase energy efficiency.

In general macromolecular motions are often the essential link between structure and function. They are also of great intrinsic interest because of their relationship to principles of macromolecular structure and stability. A rich literature in macromolecular motions exists[2-5]. Although studying individual protein motions provides the most information about the manner in which a specific protein operates, by systematizing and analyzing many of the instances of protein structures solved in multiple conformations it is possible to study motions in a database framework. This provides a statistical overview of motions, making it possible to sense broad patterns and trends as well as place an individual motion in perspective. This approach also encourages the development of standardized tools and approaches.

Previously, we developed a comprehensive scheme for classifying and systematizing protein motions[6,7]. This scheme is intended to be useful to those studying structure-function relationships (in particular, rational drug design[8]) and also those involved in large-scale protein or genome surveys. Boutonnet et al. also made a detailed attempt at the systematic classification of protein motions[9].

Numerous computational methods have been developed for the study of protein motions[6]. Among these are many computational methods from traditional biophysics (e.g., molecular dynamics, energy minimization) which also relate to problems involving protein folding and the analysis of static (i.e., non-moving) protein structures; these are well described in the literature[10-17].

Here we will describe how protein motions can be analyzed statistically in a database framework. First, we will describe a database of motions and then we will illustrate some overall statistical themes derived from it, particularly related to the prevalence of motions in proteins. Next, we will present some computational tools that are well suited towards studying protein motions in a database framework (i.e.

structural alignment, adiabatic mapping interpolation, normal mode analysis, and Voronoi packing calculations). Finally, looking towards the future, we will suggest how database analysis of motions can be extended to the vast new frontier of genomic sequences, through the identification of likely hinge residues in primary sequence.

A Database of Macromolecular Motions

A statistical survey of protein and nucleic acid motions is embodied in the Database of Macromolecular Motions (Figure 2), a comphrensive internet-accessible database[7] that attempts to classify all known instances of macromolecular motions on the basis of size and packing (Figure 3]). This database is accessible on the web at molmovdb.org and is tightly integrated with a number of other internet resources, such as the PDB[18], scop[19], CATH[20], Entrez[21], SPINE[22] and PartsList[23-26].

Attributes of a Motion

Each motion in the database is associated with a variety of information:

(i) Classification. A classification number gives the place of a motion in the size and packing classification scheme for motions described below.

(ii) 3D Structures. The identifiers have been made into hypertext link that link indirectly to the structure entries at PDB and other databases.

(iii) Literature references, cross-referenced through medline.

(iv) Standardized numeric values describing the motion, such as the maximum displacement (overall and of just backbone atoms), the degree of rotation around the hinge, and residues with large torsion angle changes when these numbers are available from the scientific literature. (The morph server, described below, attempts to automatically compute these values from the structures.).

(v) Annotation Level. The database is constructed so that each entry indicates the evidence behind its description and classification. For example, the classification might be based on careful manual analysis of two conformations, automatic output of a `conformation comparison program, inferred based on structure comparison, or inferred based on sequence comparison. A clear distinction was made between the carefully analyzed, “gold-standard” motions, such as lactoferrin, and the much more tentatively understood motion in a protein that is a sequence homologue of another protein which is structurally similar to lactoferrin. They indicated the evidence behind a motion through listing information about the experimental techniques used, telling whether or not the motion is inferred, and giving a standardized “annotation level.”

Size Classification

The classification scheme for proteins has a hierarchical layout shown in Figure 3. Proteins motions are first ranked in order of their size (subunit, domain, and fragments). Domain motions, such as those in hexokinase or citrate synthase[27,28], provide the most common examples of protein flexibility[29-31]. Usually, the motion of fragments smaller than domains refers to the motion of surface loops, such as the ones in triose phosphate isomerase or lactate dehydrogenase. It can also refer to the motion of secondary structures, such as of the helices in insulin[32-34].

Packing Classification

For fragment and domain protein motions the database systematizes the motions on the basis of the packing of atoms inside of proteins, which is a fundamental constraint on protein structure[29,34-42]. Interfaces between different parts of a protein are usually packed very tightly. Consequently, two basic mechanisms for protein motions, hinge and shear, are proposed depending on whether or not there is a continuously maintained interface preserved through the motion (Figure 3). A complete protein motion can be built up from a number of these basic motions. For the database, a motion is classified as "Shear" if it is predominately a shear motion and "Hinge" if it is predominately composed of hinge motions.

The shear mechanism basically describes the special kind of sliding motion a protein must undergo if it wants to maintain a well-packed interface; these constraints mean that individual shear motions are constrained to be very small. When no continuously maintained interface constrains the motion, a hinge motion occurs. Typically, these motions usually occur in proteins with two domains (or fragments) connected by linkers (i.e. hinges) that are relatively unconstrained by packing. The whole motion may be produced by a few large torsion angle changes.

Beyond hinge and shear, there are number of other possible classifications:

* A special mechanism that is clearly neither hinge nor shear accounts for the motion. An example of this sort of motion is what occurs in the immunoglobulin ball-and-socket joint[43], where the motion involves sliding over a continuously maintained interface (like a shear motion) but because the interface is smooth and not interdigitating the motion can be large (like a hinge).
* Motion involves a partial refolding of the protein. This usually results in dramatic changes in the overall structure.

Subunit motions are classified differently as allosteric, non-allosteric, or unclassifiable. Finally, large protein motions which cannot easily be classified as subunit motions are classified as complex movements. For example, the order-to-disorder transition that the headpiece domain undergoes when it binds DNA. Another example involves a molecule binding between two other domains in the protein, such as observed in the bacterial periplasmic binding proteins[44].

How many motions are there?

One basic question relates to the number of proteins motions and to what degree they divide up amongst the basic classification categories in the database. This can be answered on a number of levels, depending on our degree of knowledge about the motion. There are currently (21 September 2001) the following motions in the protein motions database and other public repositories:

(i) 120 manually classified and curated motions. There are 261 pairs of PDB identifiers that are associated with the best-studied (gold-standard) motions. These are motions for which evidence has been manually gathered from the scientific literature and compared to the structures. The vast majority of these have at least two solved x-ray structures.

(ii) 240 submitted morphs. These were contributed interactively by Internet users via a web form. They have annotation of a variable quality, depending on the person submitting the motion. Some of these explicitly reference two different PDB structures, though many use uploaded coordinates.

(iii) 441 PDB annotated entries. There are 441 entries in the PDB that explicitly mention the phrase "conformational change" in their comments section.

(iv) 3814 automatically found conformational change outliers. Wilson et al.[45] did a comprehensive set of structural alignments on version 1.39 of the scop database, which represented most the known structures as of the beginning of 1999. From this they found 4403 pairs of domains that had appreciable sequence similarity yet had great structural differences; these represented putative instances of conformational changes. 3814 of these could be standardized and processed by the morph server (see below).

(v) 13191 hits in PubMed. Searching the titles and abstracts provide san additional way to identify putative motions and get a sense of the full size the motions "universe." There are currently 13191 entries in the NCBI's PubMed database that contain phrases such as "conformational change" or "macromolecular motion." The increase of these terms over the years is diagrammed in Figure 5 and the search methodology is explained in the caption. Obviously this number contains quite a few false positives. One can estimate the fraction of false positive by examining their occurrence in a randomly selected subset of 100 articles. Doing this yields a false-positive rate of ~20%, implying that only about 10,000 of the hit represent real motions.

One can breakdown the gold-standard motions depending on their classification or experimental technique (Figure 4). Over 60% of the motions in the database are classified as domain motions, while the hinge mechanism is the most common mechanistic classification in the database, accounting for 45% of the entries. Reflecting the greater ease with which smaller motions can be studied experimentally, a greater percentage of fragment motions have structures for multiple conformations in the motion. Most of the fragment and domain motions in the database fall into the hinge or shear classification.

The most common method for study of protein motions involving a mechanical function is traditional x-ray crystallography[1,46], which was found to have contributed experimental data to nearly all of the protein motions in one comprehensive survey[7]. NMR[47], Time-resolved X-ray crystallography[48-50], and computational techniques such as molecular dynamics each contributed to less than 7% of the surveyed motions[7]. However, it is conceivable that one or more of these latter techniques may become considerably more important as methodological advances continue to be made.

Methods for protein structure comparison

The study of protein motions in a database framework rests on a number of techniques, which we discuss in the following sections. One of the most basic of these techniques is structure comparison, i.e., the comparison of two structures to determine which residues are analogous and then to superpose them based on these residues.

Sieve-fit Superposition and Screw Axis Orientation

If one has the correspondences between atoms in two structures (i.e. an alignment between an open and closed structure), one can use traditional "RMS superposition" to minimize the RMS difference between the atoms. However, there are some complexities associated with this. In a simple hinge motion, e.g. calmodulin, such an alignment fits the closed conformation symmetrically inside the open conformation. Amongst other things, the maximum Cα displacement computed from such a superposition is considerably underestimated from the common sense alignment, and an analysis or morph movie made with such an alignment would give the impression of a motion far more complicated than a simple opening of a hinge. To overcome these problems, one needs to do an iterated superposition or "sieve-fit"[42,51-53].

A comparison of the new position of the ending conformation following the last fit with its position following the "sieve-fit" procedure yields a geometric transformation whose screw axis is (approximately) the axis of the hinge motion[51]. If a significant hinge motion is present, one can use these transformations to align the Z-axis of the coordinate frame parallel to the hinge axis so that, when the motion is rendered, viewers will look down the screw-axis of the hinge motion.

Structural Alignment

When the proteins being compared have different sequences and one does not have an obvious alignment between the two sequences one has to first use pairwise structural alignment before superposition can be attempted. Structural alignment consists of establishing equivalences between the residues in two different proteins, as is the case with conventional sequence alignment. However, this equivalence is determined principally on the basis of the three-dimensional coordinates corresponding to each residue, not on the basis of the amino acid type. The general idea of structural alignment has been around since the first comparisons of the structures of myoglobin and hemoglobin[54]. Systematic structural alignment began with the analysis of heme binding proteins and dehydrogenases by Rossmann and colleagues[55,56].

Completely automatic methods have the advantage of speed and objectivity. However, the structural classifications produced by a computer are not always as understandable or reliable as those produced by humans. Furthermore, although manual classification is slow, if it is done correctly, it only has to be done once.

Because of their obvious utility, a large number of automated methods for protein structure comparison have been developed, using different representation of structures, definitions of similarity measure and optimization algorithms[57-72]. Among them, methods based on utilization of distance matrices (also called distance maps or distance plots)[73] Nishikawa & Ooi, 1974; Liebman, 1980; Sippl, 1982) for describing and comparing protein conformations were found quite useful for treating large structures. Some of these effectively compare the respective distance matrices of each structure, trying to minimize the difference in intra-atomic distances for selected aligned substructures[60,61,74]. Other methods[58,75] directly try to minimize the inter-atomic distances between two structures. A similar approach is taken in minimizing the "soap-bubble area" between two structures[68]. Yet other methods involve further techniques, such as geometric hashing or lattice fitting[59,66,69].

To understand these procedures, it is useful to compare structural alignment with the much more thoroughly studied methods for sequence alignment[76,77]. Both sequence and structure alignment methods produce an alignment that can be described as an ordered set of equivalent pairs (i,j) associating residue i in protein A with residue j in protein B. Both methods allow gaps in these alignments that correspond to non-sequential i (or j) values in consecutive pairs—i.e., one has pairs $(i, j \neq i)$. And both methods reach an alignment by optimizing a function that scores well for good matches and badly for gaps. The major difference between the methods is that the optimization used for sequence alignment is globally convergent, whereas that used for structural alignment is not. This is the case for sequence alignment because the optimum match for one part of a sequence is not affected by the match for any other part. Structural alignment fails to converge globally because the possible matches for different segments are tightly linked as they are part of the same rigid 3D structure. For this reason, the alignment found by a structural alignment algorithm can depend on the initial equivalences, whereas in sequence alignment there is no such dependence.

The lack-of-convergence problem has led to a large number of different approaches to structural alignment, the methods differing in how they attack the problem. However, no current algorithm can find the globally optimum solution all the time; the convergence problem remains unsolved in the general case. The methods also differ in the function they optimize (the equivalent of the amino acid substitution matrix used in sequence alignment) and how they treat gaps.

Multiple Structural Alignment

The next step after pairwise structural alignment is multiple structural alignment, simultaneously aligning three or more structures together. This is an essential first step in the construction of *consensus* structural templates, which aim to encapsulate the information in a family of structures. It can also form the nucleus for a large multiple sequence alignment—i.e., highly homologous sequences can be aligned to each structure in the multiple alignment. There are currently a number of approaches for this[72,78-80]. Most of them proceed by analogy to multiple sequence alignment[80-83] building up an alignment by adding one structure at a time to the growing consensus.

Since most new structures are similar structurally to ones reported previously they can be grouped into families, and with sufficient number of members in each family it becomes possible to summarize, statistically, the commonalities and differences within each family. A method for finding the atoms in a family alignment that have low spatial variance and those that have higher spatial variance has been developed[84,85]. It allows one to determine the "core" atoms that have the same relative position in all family members and the "non-core" atoms that do not.

HMMs for Structural Alignment

Recently, Hidden Markov Models (HMMs) applied to sequences were found to be highly useful for relating protein structures. In particular, they have been used for building the Pfam database of protein familes[86-88], for gene finding[89], for predicting secondary structure[90] and transmembrane helices[91]. An important property of HMMs is their ability to capture information about the degree of conservation at various positions in a sequence alignment and the varying degree to which insertions and deletions are permitted. This explains why HMMs can detect considerably more homologues compared to simple pairwise comparison[89,92]. Despite the fact that the recent attempts proved that linear HMMs can be useful for structural studies[90,91] none of the suggested schemes are fundamentally three-dimensional (coordinate dependent), since all of them are based on building a 1D HMM profile representing a sequence alignment and structural information only enters in the form of encoded symbols (i.e. *H* for helix and *E* for sheet). Adding in real 3D structure turns out to be non-trivial, as the structure is fundamentally different from the sequence not only in increased dimensionality, but also due to the transition from discrete to continuous representation. Efforts to build HMMs that explicitly represents a protein in terms of 3D coordinates are currently underway[93].

Interpolating between structures: the Morph Server

Following alignment and superposition of two structures, it is possible to characterize the extent to which the two conformations differ using a variety of straightforward analytical operations and statistical measures. Many of these metrics can be derived through trying to "intelligently" interpolate between the two structures. This is achieved through the morph server, which is associated with the database of macromolecular motions.

Adiabatic Mapping Interpolation

The morph server attempts to describe protein motions as a rigid-body rotation of a small "core" relative to a larger one, using a set of hinges. To ensure all statistics between any two motions are directly comparable, the motion is placed in a standardized coordinate system. Without special techniques, such as high temperature simulation or Brownian dynamics[94,95], normal dynamics simulations cannot approach the timescales of the large-scale motions in the database. Rather, a pathway interpolation is produced by two principal methods:

(i) Straight Cartesian interpolation. The difference in each atomic coordinate (between the known endpoint structures) is calculated and then divided into a number of evenly spaced steps.

(ii) Adiabatic Mapping. This is a modification of straight Cartesian interpolation, adding the addition of energy minimization after each interpolation step. This procedure produces interpolated frames with much more realistic geometry.

A criticism of adiabatic mapping, often made by researchers attempting to interpolate between a protein in an unfolded and native, folded state, is that the intermediates, although geometrically realistic, have somewhat higher energies than a theoretical analysis would indicate. However, from the standpoint of the server, which tackles the significantly easier problem of interpolating between two folded conformations, the intermediates' energies are not that unrealistic (they can serve at least as a reasonable upper bound). Moreover, the technique has the advantage of providing useful results back to the user in a reasonable amount of time.

Visualization and statistics

With the intermediate conformations morphed, the molecule is now visually rendered (Figure 2). In connection with this, Martz[96] has developed an external web site that provides a Chime-based interface to the interpolated images.

Users have already submitted hundreds of examples of protein motions to the server, producing a comprehensive set of statistics. Some examples of recent morphs include human interleukin 5[97], bc1 complex[98,99], glycerol kinase[100,101], and lactoferrin[102,103]. The server collects a number of statistics, include hinge angle rotation during motion. Of the ~200 motions submitted for analysis, the median motion has a maximum rotation of 9.5° over a range of 0 through 150° as computed by our algorithm, whereas the twelve motions culled from the scientific literature had an average rotation of 24° over a range of 5 through 148°. Similarly, the algorithms found a median maximum Cα displacement of 17 Å ranging from 0 to 81Å for the submitted motions, whereas eleven motions reported in the scientific literature average 12Å over a range of 1.5 through 60Å. Although most of the structures are very similar in sequence, the server has been able to accommodate sequence identity down to 8% for some motions.

Normal Mode Analysis

While the morph server can analyze motions when two or more solved conformations exist, in many cases a protein having a suspected motion will only have one conformation with a solved 3D structure. Given only one solved conformation, normal mode analysis is one of the best ways to understand and perhaps predict its flexibility.

For this kind of analysis, normal mode analysis has two advantages for large-scale database analysis over other techniques, such as molecular dyanimcs: (1) it requires very little CPU power (especially when certain realistic approximations are made), and thus is amendable to database screening techniques[104], and (2) it provides an intuitive conceptual model of protein motions in terms of frequencies and vibrations.

Widely used by spectroscopists for years[105], advances in computer technology made normal mode analysis of large molecules practical[106-115]. The concept of normal mode analysis is to find a set of basis vectors (normal modes) describing the molecule's concerted atomic motion and spanning the set of all $3N-6$ degrees of freedom. For very large molecules the lowest frequency normal modes of proteins are thought to correspond to the large-scale real-world vibrations of the protein[116], and can be used to deduce significant biological properties. Moreover, there is evidence to suggest[117-122] that proper, symmetric normal mode vibration of binding pockets is crucial to correct biological activity in some proteins.

The classical Lagrangian for the vibrations of a protein with *N* atoms is given by

$$L = T - V\,, \tag{1}$$

where *V* is the potential energy describing interactions among atoms, and *T* is the kinetic energy

$$T = \tfrac{1}{2}\sum_{i=1}^{N} m_i\left(\dot{x}_i^2 + \dot{y}_i^2 + \dot{z}_i^2\right), \tag{2}$$

where the dot notation has been used for derivatives with respect to time.
The above expression for *T* is can be rewritten by introducing mass-weighted Cartesian displacement coordinates. Let

$$q_1 = \sqrt{m_1}\cdot\left(x_1 - x_{1e}\right),\ldots,q_{3N} = \sqrt{m_N}\cdot\left(z_N - z_{Ne}\right), \tag{3}$$

in which the q_i coordinate is proportional to the displacement from the equilibrium value q_{ie}. Expanding potential energy in Taylor series and neglecting all terms with powers greater than two (harmonic approximation), potential energy will assume the form

$$V = \tfrac{1}{2}\sum_{i,j=1}^{3N} f_{ij}q_iq_j\,, \tag{4}$$

where f_{ij}, the force constants, are defined as the second derivatives of the potential energy function:

$$f_{ij} = \frac{\partial^2 V}{\partial q_i \partial q_j}\,. \tag{5}$$

Substituting $T = \tfrac{1}{2}\sum_{i=1}^{3N}\ddot{q}_i^2$ and $V = \tfrac{1}{2}\sum_{i,j=1}^{3N} f_{ij}q_iq_j$ in the Lagrange's equation of motion

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_i}\right)_{q_i} - \left(\frac{\partial L}{\partial q_i}\right)_{\dot{q}_i} = 0\,, \tag{6}$$

we obtain

$$\ddot{q}_i + \sum_{i,j=1}^{3N} f_{ij}q_iq_j = 0\,,\quad i\text{=1..3}N, \tag{7a}$$

or in matrix form

$$\ddot{\mathbf{q}} + \mathbf{Fq} = 0\,. \tag{7b}$$

This is the set of 3*N* coupled 2nd order differential equations with constant coefficients. It can be solved by assuming a solution of the form

$$q_i = A_i\cos(\omega t + \varphi)\,. \tag{8}$$

This substitution converts the set of differential equations into a set of 3*N* homogeneous linear equations:

$$\sum_{j=1}^{3N} f_{ij} A_j - \omega^2 A_i = 0\,, \tag{9a}$$

or

$$\mathbf{F}\vec{\mathbf{A}} - \omega^2 \vec{\mathbf{A}} = 0\,. \tag{9b}$$

This problem may now be solved with any eigenvector/eigenvalue solution method. One of the simplest is to attempt to diagonalize the matrix $\mathbf{F}$ and extract the eigenvalues from the diagonal. It turns out that six eigenvalues of $\mathbf{F}$ are zero for a non-linear molecule. This result can be expected from the fact that there are three degrees of freedom associated with the translation of the center of mass, and three with rotational motion of the molecule as a whole. Since, there is no restoring force acting on these degrees of freedom, their eigenvalues are zero.

Associated with each eigenvalue is a coordinate, called normal mode coordinate Q_i. The normal modes represent a set of coordinates related to the old one by an orthogonal linear transformation $\mathbf{U}$:

$$\mathbf{Q} = \mathbf{U}\mathbf{q}\,, \tag{10}$$

Such that the transformation matrix $\mathbf{U}$ diagonalizes $\mathbf{F}$:

$$\mathbf{U}\mathbf{F}\mathbf{U}^T = \mathbf{\Lambda}\,\text{(diagonal).} \tag{11}$$

This transformation has a deep impact on the resulting form of the differential equations:
(7b) transforms to

$$\ddot{\mathbf{Q}} + \mathbf{\Lambda}\mathbf{Q} = 0\,, \tag{12}$$

but since $\mathbf{\Lambda}$ is diagonal, the equations (12) are effectively decoupled:

$$\ddot{Q}_1 + \lambda_1 Q_1 = 0\,, \ldots, \ddot{Q}_{3N} + \lambda_{3N} Q_{3N} = 0\,, \tag{13}$$

and the system therefore behaves like a set of 3N independent harmonic oscillators, each oscillating without interaction with the others.

It is of considerable importance to examine the nature of the above solutions. It is evident from Eq. (8) that each atom is oscillating about its equilibrium position with the same frequency and phase for a given solution ω_k. In other words, each atom reaches its position of maximum displacement at the same time, and each atom passes through its equilibrium position at the same time. A mode of vibration having all these characteristics is called a *normal mode* of vibration, and its frequency is known as a *normal mode frequency*.

Tools for Quantification of packing: Voronoi polyhedra

Packing clearly is an essential component of a motion's classification. Often this concept is discussed loosely and vaguely by crystallographers analyzing a particular protein structure—for instance, "Asp23 is packed against Gly38" or "the interface between domains appears to be tightly packed." One can systematize and quantify the discussion of packing in the context of the motions database through the use of particular geometric constructions called Voronoi polyhedra.

Nearly a century ago, Voronoi developed the method to construct polyhedra as a novel application of quadratic equations[123]. Bernal and Finney used them to study the structure of liquids in the 1960s[124]. However, despite the general utility of these polyhedra, their application to proteins was limited by a serious methodological difficulty. While the Voronoi construction is based around partitioning space amongst a collection of "equal" points, all protein atoms are not equal: some are clearly larger than others (e.g. sulfur versus oxygen). Richards found a solution to this problem and first applied the Voronoi methods to proteins in 1974[125]. He has, subsequently, reviewed their use in this application[38,40]. Richard's solution was to allocate space based proportionally to the size of an atom's atomic radii. The resulting Voronoi-like polyhedra were no longer an equal partition of space, but were weighted by an atom's size. However, as an additional level of complexity, atoms usually include their bonded hydrogens, since these are not usually resolved in the solutions to crystal structures. This united atom model has posed a problem in finding the correct radii to use in Voronoi as well as other applications[126,127]. In a detailed analysis of organic crystals and protein structures, we have develop a standard set of protein atom radii for united atom models[41,127,128], which are shown in Table 1.

Voronoi polyhedra are a useful way of partitioning space amongst a collection of atoms. The simplest method for calculating volumes in a Voronoi-like manner is to put all atoms in the system on a grid. Then go to each grid-point (i.e. voxel) and add its volume to the atom center closest to it. This is prohibitively slow for a real protein structure, but it can be made somewhat faster by randomly sampling grid-points[129]. More classic approaches to calculating Voronoi volumes have two parts: (1) for each atom find the vertices of the polyhedron around it and (2) systematically collect these vertices to draw the polyhedron and calculate its volume. In the classic Voronoi construction (Figure 8), each atom is surrounded by a unique limiting polyhedron such that all points within an atom's polyhedron are closer to this atom than all other atoms. Points equidistant from two atoms are on a plane; those equidistant from three atoms are on a line, and those equidistant from four centers form a vertex. One can use this last fact to easily find all the vertices associated with an atom. With the coordinates of four atoms, it is straightforward to solve for possible vertex coordinates using the equation of a sphere. One then checks whether this putative vertex is closer to these four atoms than any other atom; if so, it is a vertex.

In the procedure just outlined, all the atoms are considered equal, and the dividing planes are positioned midway between atoms (Figure 8). As mentioned above, this method of partition, called bisection, is not physically reasonable for proteins, which have atoms of obviously different size (such as oxygen and sulfur). It chemically misallocates volume, giving more to the smaller atom. Currently, two principal methods of re-positioning the dividing plane have been proposed to make the partition more physically reasonable: method B[125] and the radical-plane method[130]. Both methods depend on the radii of the atoms in contact (R1 and R2) and the distance between the atoms (D).

As Richards originally showed[131] and many have shown recently[132-137], Voronoi procedure is particularly well suited for analyzing the packing of the protein interior. The Voronoi procedure fails at a protein's surface, since atoms do not have neighbors and only incomplete polyhedra can be built. Unlike the surface, protein interiors all have neighbors, and the construction of Voronoi polyhedra is able to allocate all space amongst this collection of atoms. There are no gaps as there would be if one, say, simply drew spheres around the atoms. Thus, the volume of cavities or defects between atoms are included in their Voronoi volume, and one finds that the packing efficiency is inversely proportional to the size of the polyhedra. This indirect measurement of cavities contrasts with other types of calculations that measure the volume of cavities explicitly. Moreover, since protein interiors are tightly packed, fitting together like a jig-saw puzzle, the various types of protein atoms occupy well-defined amounts of space. This fact has made the calculation of standard volumes for atoms and residues in proteins a worthwhile proposition using Voronoi constructs[127,128,136,138]. It was shown that the residue volumes derived from the previously mentioned radii set using a Voronoi procedure in Table 2. Comparing these standard protein volumes to those calculated along the interface of a domain motion can be use to analyze the quality of packing, as has been done in a similar analysis of protein crystals[136]. Such an analysis provides another property of a domain motion to use in its characterization.

Additional techniques for studying protein motions

We have described computational techniques most suitable for a high-throughput analysis of thousands of proteins motions within a database. Over the years, many other techniques have been developed for analyses of individual protein motions, mostly derived from classical molecular mechanics approaches[2,4,5,139-142]. Many of these require much greater computational resources than the methods described here, and so are better suited to detailed study of an individual molecule rather surveys of a whole database.

Table 3 presents a summary of many of the computational techniques. Molecular Dynamics (MD), Energy Minimization (EM), and Normal Mode Analysis (NMA) are arguably the most widely used of the techniques presented[143]. There are many variants on these techniques that do not differ substantially in terms of computational cost. Adiabatic mapping, used in the morph server, is essentially a form of energy minimization. It should be emphasized that there are significant differences in the tractability of the various techniques. For instance, for constructing a protein interpolation in the morph server, molecular dynamics requires six orders of magnitude more processing power than simple energy minimization.

Relating Protein Motions to Genome Sequences

Genome sequencing has vastly expanded the amount of information available for bioinformatic analyses. However, much of the information in genomes is raw and uncharacterized from the point of view of protein structure and function [144-146]. One of the current challenges is take the information about the few relatively well-characterized proteins, such as those in the macromolecular motions database, and extrapolate this to uncharacterized genome sequences. In general this process is dubbed annotation transfer [45,147,148].

In relation to macromolecular motions, one of the most useful calculations we can do is to develop models for predicting the location of the flexible linkers that typically serve as the hinges in protein motions. A first step in accomplishing this is to determine the amino acid propensities of interdomain linkers. Below two propensity scales for amino acids to be in linkers in general or in flexible hinges in particular were calculated using structural data from the database of macromolecular motions [149].

It may be possible to predict protein domains in protein sequences of unknown structure using information about the amino acid composition of linker sequences. For example, a profile of flexible linker regions might be used to predict the location of domain hinges for the structural annotation of genome sequences. A tool to achieve this successfully would be quite useful in the context of gene-finding in genomic sequences [150].

Propensities for Linkers in General

Flexible as well as inflexible linkers are included in the first method of analysis. We have arbitrarily defined in this method a linker sequence as a 16-residue region centered on the peptide bond linking two domains.

The analysis of the amino acid composition of linker sequences is an example of deducing sequence information from structural information. The location of protein domains and other structural information can be found in SCOP [26,151], which contains several databases of amino acid sequences of protein domains. The PDB40 database provided by SCOP was used to create a database of linker sequences. The PDB40 database consists of a subset of proteins in the Protein Data Bank (PDB) with known structures selected so that, when aligned, no two proteins in the subset show a sequence identity of 40% or greater. Thus, the data set is not biased towards protein structures listed multiple times in the PDB. From the 1,500 protein sequences in the PDB40 database it was possible to extract 234 linker sequences, thus reflecting that only a small fraction of proteins contain multiple domains and therefore linker regions.

Table 4 shows a profile of the amino acid composition at each of the sixteen positions in the linker sequence. The residue-specific amino acid composition can be summarized in the average amino acid composition of the whole linker sequences (Figure 9). The linker sequences can be regarded as a random sample of the sequences in the PDB40 database and thus the statistical significance of this sample can be determined. In particular, the probability $P^n(k)$ that a particular amino acid occurs k times among the n amino acids in a sequence sample is given by the familiar binomial distribution:

$$P^n(k) = \binom{n}{k} p^k (1-p)^{n-k} . \tag{13}$$

where p is the probability that the amino acid occurs in the PDB40 database (n = 234 for the distribution of every *single* of the sixteen specific linker positions and n = 234 x 16 for the distribution of the linker *average*). Accordingly, the cumulative distribution function $CDF^n(k)$, representing the probability that the amino acid occurs less than k times, is then given by:

$$CDF^n(k) = \sum_{i=0}^{k} P^n(i) . \tag{14}$$

Consequently, if o and e are the observed and expected counts, then a two-sided P-value is given by $1\text{-}CDF^n(e+|o\text{-}e|) + CDF^n(e\text{-}|o\text{-}e|)$. This is the probability that the number of amino acid counts in a random subset of the PDB40 would be either smaller or greater than the expected value by a difference $|o\text{-}e|$. The two-sided P-values are shown in figure 10 for the average linker compositions across all 16 positions. The results imply, with better than 98% confidence, that linker regions are proline-rich and alanine- and trypthophan-poor. In particular, the statistical evidence that linkers are proline-rich is unusually strong and is significant at better than the hundredth-of-a-percent level. No particular trends could be seen after roughly grouping the amino acids according to the attributes hydrophobic, charged, and polar (Table 5 and Figure 10) following the classification of Branden and Tooze[152] The frequencies of the remaining amino acids in linkers are not statistically different from the database as a whole at the 5% significance level. P-values for amino acids at each of the six-teen linker positions are shown in Table 5.

Towards Propensities for Flexible Linkers

A variant of their procedure involves focusing just on linkers that are known to be flexible. The Database of Macromolecular Motions contains residue selections for known protein hinge regions (i.e., flexible linkers) that have been found in the scientific literature. These sequences were manually verified to be true flexible linker regions, and thus this database is a potential “gold standard” free from algorithmic biases that can be used as a starting point in the development of propensity scales and other research leading towards algorithmic techniques. By expanding these residue selections slightly with a predetermined protocol and extracting the corresponding sequences from the PDB, a series of sequences of known flexible linkers can be obtained. A FASTA search with a suitable cutoff (e.g., e-value 0.001) can then be performed on known linker sequences to obtain a series of near homologues (Table 7). It is then possible to arrange these homologues into a multiple alignment (via the CLUSTALW) program[153,154] and the multiple alignment can be fused into a variety of consensus pattern representations, such as Hidden Markov Models or simply consensus sequences[155-159]. A sample multiple alignment for the hinge in calmodulin was performed (Table 7) and a number of consensus sequences generated (Table 6). It is possible to average the amino acid composition over all the different hinges and different positions within a hinge to give a single composition vector for flexible hinges. Finally, by comparing this latter quantity to the overall amino acid composition or that of just linkers a preliminary scale of amino acid propensity in flexible linkers may be obtained (Table 8). This can be compared with the scale of amino acid propensities in linkers as obtained by the procedure previously described (Table 4).

Conclusions

We have described how protein flexibility can be studied in a database framework. The database of macromolecular motions contains thousands of motions, with varying levels of annotation. We survey a number of the tools that underlie the statistics in the database (i.e. structural alignment, adiabatic mapping interpolation, normal mode analysis, and Voronoi packing calculations). Finally, looking towards the future, we suggest how database analysis of motions can be extended to the vast new frontier of genomic sequences, through the identification of likely hinge residues in primary sequence.

We expect that the number of macromolecular motions will greatly increase in the future, making a database of motions somewhat increasingly valuable. Our reasoning behind this conjecture is as follows: The number of new structures continues to go up at a rapid rate (nearly exponential). However, the increase in the number of folds is much slower and is expected to level off much more in the future as the we find more and more of the limited number of folds in nature, estimated to be as low as 1000. Each new structure solved that has the same fold as one in the database represents a potential new motion—i.e. it is often a structure in a different liganded state or a structurally perturbed homologue. Thus, as we find more and more of the finite number of folds, crystallography and NMR will increasingly provide information about the variability and mobility of a given fold, rather than identifying new folding patterns.

Databases potentially represent a new paradigm for scientific computing. In a highly schematized cartoon view, scientific computing traditionally involved big calculations on fast computers. The aim in these often was prediction based on first principles—e.g. prediction of protein folding based on molecular dynamics. These calculations naturally emphasized the processor speed of the computer. In contrast, the new “database paradigm” focuses on small, inter-connected information sources on many different computers. The aim is communication of scientific information and the discovering of unexpected relationships in the data—e.g. the finding that heat shock protein looks like hexokinase. In contrast to their more traditional counterparts, these calculations are more dependent on disk-storage and networking rather than raw CPU power.

Acknowledgements

The authors gratefully acknowledge the financial support of the Keck Foundation and the National Science Foundation (Grant DBI-9723182). Numerous people have also either contributed entries or information to the database and morph server or have given us feedback on what the user community wants. The authors also wish to thank Informix Software, Inc. for providing a grant of its database software.

References

1. Xu, Z. & Sigler, P. B. GroEL/GroES: structure and function of a two-stroke folding machine. *J Struct Biol* **124**, 129-41 (1998).
2. Isralewitz, B., Gao, M. & Schulten, K. Steered molecular dynamics and mechanical functions of proteins. *Curr Opin Struct Biol* **11**, 224-30. (2001).
3. Young, M., Kirshenbaum, K., Dill, K. A. & Highsmith, S. Predicting conformational switches in proteins. *Protein Sci* **8**, 1752-64. (1999).
4. Shaknovich, R., Shue, G. & Kohtz, D. S. Conformational activation of a basic helix-loop-helix protein (MyoD1) by the C-terminal region of murine HSP90 (HSP84). *Mol Cell Biol* **12**, 5059-68. (1992).
5. Dixon, M. M., Nicholson, H., Shewchuk, L., Baase, W. A. & Matthews, B. W. Structure of a hinge-bending bacteriophage T4 lysozyme mutant, Ile3-- >Pro. *J Mol Biol* **227**, 917-33. (1992).
6. Krebs, W. G. & Gerstein, M. The morph server: a standardized system for analyzing and visualizing macromolecular motions in a database framework. *Nucleic Acids Res* **28**, 1665-1675 (2000).
7. Gerstein, M. & Krebs, W. A Database of Macromolecular Movements. *Nucl. Acids Res* **26**, 4280 (1998).
8. Kuntz, I. D. Structure-Based Strategies for Drug Design and Discovery. *Science* **257**, 1078-1082 (1992).
9. Boutonnet, N. S., Rooman, M. J. & Wodak, S. J. Automatic analysis of protein conformational changes by multiple linkage clustering. *J. Mol. Biol.* **253** (1995).
10. Tsai, J., Levitt, M. & Baker, D. Hierarchy of structure loss in MD simulations of src SH3 domain unfolding. *J Mol Biol* **291**, 215-25. (1999).
11. Tang, Y. Z., Chen, W. Z. & Wang, C. X. Molecular dynamics simulations of the gramicidin A-dimyristoylphosphatidylcholine system with an ion in the channel pore region. *Eur Biophys J* **29**, 523-34 (2000).
12. Van Belle, D., De Maria, L., Iurcu, G. & Wodak, S. J. Pathways of ligand clearance in acetylcholinesterase by multiple copy sampling. *J Mol Biol* **298**, 705-26. (2000).
13. Wlodek, S. T., Shen, T. & McCammon, J. A. Electrostatic steering of substrate to acetylcholinesterase: analysis of field fluctuations. *Biopolymers* **53**, 265-71. (2000).
14. Daggett, V. & Levitt, M. Realistic simulations of native-protein dynamics in solution and beyond. *Annu Rev Biophys Biomol Struct* **22**, 353-80 (1993).
15. Berneche, S. & Roux, B. Molecular dynamics of the KcsA K(+) channel in a bilayer membrane. *Biophys J* **78**, 2900-17. (2000).
16. Gilson, M. K. *et al.* Open "Back Door" in a Molecular Dynamics Simulation of Acetylcholinesterase. *Science* **263**, 1276-1278 (1994).
17. Wriggers, W. & Schulten, K. Investigating a back door mechanism of actin phosphate release by steered molecular dynamics. *Proteins* **35**, 262-73. (1999).
18. Berman, H., M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E. The Protein Data Bank. *Nucleic Acids Res.* **28**, 235-242 (2000).
19. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* **247**, 536-40. (1995).
20. Orengo, C. A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B., Thornton, J.M. CATH--a hierarchic classification of protein domain structures. *Structures* **5**, 1093-1108 (1997).
21. Hogue, C. W., Ohkawa, H. & Bryant, S. H. A dynamic look at structures: WWW-Entrez and the Molecular Modeling Database. *Trends Biochem Sci* **21**, 226-9. (1996).
22. Bertone, P. *et al.* SPINE: an integrated tracking database and data mining approach for identifying feasible targets in high-throughput structural proteomics. *Nucleic Acids Res* **29**, 2884-98. (2001).

23. Qian, J. *et al.* PartsList: a web-based system for dynamically ranking protein folds based on disparate attributes, including whole-genome expression and interaction information. *Nucleic Acids Research* (2001).

24. Schuler, G. D., Epstein, J. A., Ohkawa, H. & Kans, J. A. Entrez: Molecular Biology Database and Retrievel System. *Meth. Enz.* **266**, 141-162 (1996).

25. Epstein, J. A., Kans, J. A. & Schuler, G. D. WWW Entrez: A Hypertext Retrieval Tool for Molecular Biology. *2nd Ann. Int. WWW Conf.*, (in press) (1994).

26. Murzin, A., Brenner, S. E., Hubbard, T. & Chothia, C. SCOP: A Structural Classification of Proteins for the Investigation of Sequences and Structures. *J. Mol. Biol.* **247**, 536-540 (1995).

27. Remington, S., Wiegand, G. & Huber, R. Crystallographic refinement and atomic models of two different forms of citrate synthase at 2.7 and 1.7 Å resolution. *J. Mol. Biol.* **158**, 111-152 (1982).

28. Bennett, W. S., Jr & Steitz, T. A. Glucose induced conformational change in yeast hexokinase. *Proc. Natl. Acad. Sci. U.S.A.* **75**, 4848-4852 (1978).

29. Gerstein, M., Lesk, A. M. & Chothia, C. Structural Mechanisms for Domain Movements. *Biochemistry* **33**, 6739-6749 (1994).

30. Bennett, W. S. & Huber, R. Structural and Functional Aspects of Domain Motion in Proteins. *Crit. Rev. Biochem* **15**, 291-384 (1984).

31. Janin, J. & Wodak, S. Structural domains in proteins and their role in the dynamics of protein function. *Prog. Biophys. Mol. Biol.* **42**, 21-78 (1983).

32. Abad-Zapatero, C., Griffith, J. P., Sussman, J. L. & Rossman, M. G. Refined Crystal Structure of Dogfish M_4 Apo-lactate Dehydrogenase. *J. Mol. Biol.* **198**, 445-67 (1987).

33. Wierenga, R. K. *et al.* The crystal structure of the “open” and the “closed” conformation of the flexible loop of trypanosomal triosephosphate isomerase. *Proteins* **10**, 93 (1991).

34. Chothia, C., Lesk, A. M., Dodson, G. G. & Hodgkin, D. C. Transmission of conformational change in insulin. *Nature* **302**, 500-505 (1983).

35. Richards, F. M. & Lim, W. A. An analysis of packing in the protein folding problem. *Quart. Rev. Biophys.* **26**, 423-498 (1994).

36. Harpaz, Y., Gerstein, M. & Chothia, C. Volume Changes on Protein Folding. *Structure* **2**, 641-649 (1994).

37. Levitt, M., Gerstein, M., Huang, E., Subbiah, S. & Tsai, J. Protein Folding: the Endgame. *Ann. Rev. Biochem.* **66**, 549-579 (1997).

38. Richards, F. M. Calculation of Molecular Volumes and Areas for Structures of Known Geometry. *Methods in Enzymology* **115**, 440-464 (1985).

39. Richards, F. M. Areas, Volumes, Packing, and Protein Structure. *Ann. Rev. Biophys. Bioeng.* **6**, 151-76 (1977).

40. Gerstein, M. & Richards, F. in *International Tables for Crystallography* (eds. Rossman, M. & Arnold, E.) 531-539 (Kluwer, Dordrecht, 2001).

41. Tsai, J., Voss, N. & Gerstein, M. Determining the minimum number of types necessary to represent the sizes of protein atoms. *Bioinformatics* **in press** (2001).

42. Lesk, A. M. & Chothia, C. Mechanisms of Domain Closure in Proteins. *J. Mol. Biol.* **174**, 175-91 (1984).

43. Lesk, A. M. & Chothia, C. Elbow Motion in the immunoglobulins involves a molecular ball and socket joint. *Nature* **335**, 188-190 (1988).

44. Vyas, N. K., Vyas, M. N. & Quiocho, F. A. Comparison of the periplasmic receptors for L-arabinose, D-glucose, and D-ribose — structural and functional similarity. *J. Biol. Chem.* **266**, 5226-5237 (1991).

45. Wilson, C. A., Kreychman, J. & Gerstein, M. Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores [In Process Citation]. *J Mol Biol* **297**, 233-49 (2000).

46. Chik, J. K., Lindberg, U. & Schutt, C. E. The structure of an open state of beta-actin at 2.65 A resolution. *J Mol Biol* **263**, 607-23 (1996).

47. Volkman, B. F., Lipson, D., Wemmer, D. E. & Kern, D. Two-state allosteric behavior in a single-domain signaling protein. *Science* **291**, 2429-33. (2001).
48. Oka, T. *et al.* Time-resolved x-ray diffraction reveals multiple conformations in the M- N transition of the bacteriorhodopsin photocycle. *Proc Natl Acad Sci U S A* **97**, 14278-82. (2000).
49. Genick, U. K. *et al.* Structure of a protein photocycle intermediate by millisecond time- resolved crystallography. *Science* **275**, 1471-5 (1997).
50. Schlichting, I. *et al.* Time-resolved X-ray crystallographic study of the conformational change in Ha-Ras p21 protein on GTP hydrolysis. *Nature* **345**, 309 (1990).
51. Gerstein, M., Lesk, A. & Chothia, C. in *Protein Motions* (ed. Subbiah, S.) (in press) (R G Landes, Austin, TX, 1995).
52. Gerstein, M. & Chothia, C. H. Analysis of Protein Loop Closure: Two Types of Hinges Produce One Motion in Lactate Dehydrogenase. *J. Mol. Biol.* **220**, 133-149 (1991).
53. Gerstein, M. & Altman, R. Average core structures and variability measures for protein families: Application to the immunoglobulins. *J. Mol. Biol.* **251**, 161-175 (1995).
54. Perutz, M. F. *et al. Nature* **185**, 416-422 (1960).
55. Rossmann, M. G. & Argos, P. A comparison of the heme binding pocket in globins and cytochrome b5. *J Biol Chem* **250**, 7525-32. (1975).
56. Argos, P. & Rossmann, M. G. Structural comparisons of heme binding proteins. *Biochemistry* **18**, 4951-60. (1979).
57. Remington, S. J. & Matthews, B. W. A systematic approach to the comparison of protein structures. *J Mol Biol* **140**, 77-99. (1980).
58. Satow, Y., Cohen, G. H., Padlan, E. A. & Davies, D. R. Phosphocholine binding immunoglobulin Fab McPC603. An X-ray diffraction study at 2.7 A. *J Mol Biol* **190**, 593-604. (1986).
59. Artymiuk, P., Mitchell, E., Rice, D. & Willett, P. Searching techniques for databases of protein structures. *J. Inform Sci.* **15**, 287-298 (1989).
60. Taylor, W. R. & Orengo, C. A. Protein structure alignment. *J Mol Biol* **208**, 1-22. (1989).
61. Sali, A. & Blundell, T. L. Definition of general topological equivalence in protein structures. A procedure involving comparison of properties and relationships through simulated annealing and dynamic programming. *J Mol Biol* **212**, 403-28. (1990).
62. Vriend, G. & Sander, C. Detection of common three-dimensional substructures in proteins. *Proteins* **11**, 52-8 (1991).
63. Russell, R. B. & Barton, G. J. Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. *Proteins* **14**, 309-23. (1992).
64. Grindley, H. M., Artymiuk, P. J., Rice, D. W. & Willett, P. Identification of tertiary structure resemblance in proteins using a maximal common subgraph isomorphism algorithm. *J Mol Biol* **229**, 707-21. (1993).
65. Holm, L. & Sander, C. Structural alignment of globins, phycocyanins and colicin A. *FEBS Lett* **315**, 301-6. (1993).
66. Godzik, A. & Skolnick, J. Flexible algorithm for direct multiple alignment of protein structures and sequences. *Comput Appl Biosci* **10**, 587-96. (1994).
67. Feng, Z. K. & Sippl, M. J. Optimum superimposition of protein structures: ambiguities and implications. *Fold Des* **1**, 123-32 (1996).
68. Falicov, A. & Cohen, F. E. A surface of minimum area metric for the structural comparison of proteins. *J Mol Biol* **258**, 871-92. (1996).
69. Gibrat, J. F., Madej, T. & Bryant, S. H. Surprising similarities in structure comparison. *Curr Opin Struct Biol* **6**, 377-85. (1996).
70. Cohen, G. ALIGN: A program to superimpose protein coordinates, accounting for insertations and deletion. *J. Appl. Crystallogr.* (1997).
71. Gerstein, M. & Levitt, M. Comprehensive Assessment of Automatic Structural Alignment against a Manual Standard, the Scop Classification of Proteins. *Protein Science* **7**, 445-456 (1998).
72. Levitt, M. & Gerstein, M. A Unified Statistical Framework for Sequence Comparison and Structure Comparison. *Proceedings of the National Academy of Sciences USA* **95**, 5913-5920 (1998).

73. Phillips, D. C., Rivers, P. S., Sternberg, M. J. E., Thornton, J. M. & Wilson, I. A. *Biochem. Soc. Trans.* **5**, 642-647 (1977).

74. Holm, L. & Sander, C. Protein structure comparison by alignment of distance matrices. *J Mol Biol* **233**, 123-38. (1993).

75. Cohen, G. H. ALIGN: A program to superimpose protein coordinates, accounting for insertions and deletions. *J. Appl. Cryst.*, (in press) (1997).

76. Doolittle, R. *Of Urfs and Orfs* (University Science Books, Mill Valley, CA, 1987).

77. Gribskov, M. & Devereux, J. *Sequence analysis primer* (Oxford University Press, New York, 1992).

78. Russell, R. B. & Barton, G. J. The limits of protein secondary structure prediction accuracy from multiple sequence alignment. *J Mol Biol* **234**, 951-7. (1993).

79. Sali, A. & Blundell, T. L. The definition of general topological equivalence in protein structures: A procedure involving comparison of properties and relationships through simulated annealing and dynamic programming. *J. Mol. Biol.* **212**, 403-428 (1990).

80. Taylor, W. R., Flores, T. P. & Orengo, C. A. Multiple protein structure alignment. *Protein Sci* **3**, 1858-70. (1994).

81. Taylor, W. R. Multiple sequence alignment by a pairwise algorithm. *Comput Appl Biosci* **3**, 81-7. (1987).

82. Taylor, W. R. A flexible method to align large numbers of biological sequences. *J Mol Evol* **28**, 161-9. (1988).

83. Taylor, W. R. Hierarchical method to align large numbers of biological sequences. *Methods Enzymol* **183**, 456-74 (1990).

84. Gerstein, M. & Altman, R. A Structurally Invariant Core for the Globins. *CABIOS* **11**, 633-644 (1995).

85. Schmidt, R., Gerstein, M. & Altman, R. LPFC: An Internet Library of Protein Family Core Structures. *Prot. Sci.* **6**, 246-248 (1997).

86. Eddy, S. R. Hidden Markov models. *Curr Opin Struct Biol* **6**, 361-5. (1996).

87. Durbin, R., Eddy, S., Krogh, A. & Mitchison, G. *Biological sequence analysis: probalistic models of proteins and nucleic acids* (Cambridge University Press, New York, 1998).

88. Bateman, A. *et al.* The Pfam protein families database. *Nucleic Acids Res* **28**, 263-6. (2000).

89. Reese, M. G., Kulp, D., Tammana, H. & Haussler, D. Genie--gene finding in Drosophila melanogaster. *Genome Res* **10**, 529-38. (2000).

90. Bystroff, C., Thorsson, V. & Baker, D. HMMSTR: a hidden Markov model for local sequence-structure correlations in proteins. *J Mol Biol* **301**, 173-90. (2000).

91. Sonnhammer, E. L., von Heijne, G. & Krogh, A. A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc Int Conf Intell Syst Mol Biol* **6**, 175-82 (1998).

92. Park, J. *et al.* Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J Mol Biol* **284**, 1201-10. (1998).

93. Alexandrov, V. & Gerstein, M. Using 3D Hidden Markov Models that explicitly represent spatial coordinates to model and compare protein structures. *in preparation.* (2001).

94. Joseph, D., Petsko, G. A. & Karplus, M. Anatomy of Conformational Change: Hinged 'Lid' Motion of the Triosephosphosphate Isomerase Loop. *Science* **249**, 1425-1428 (1990).

95. Wade, R. C., Davis, M. E., Luty, B. A., Madura, J. D. & McCammon, J. A. Gating of the active site of triose phosphate isomerase: Brownian dynamics simulations of flexible peptide loops in the enzyme. *Biophys. J.* **64**, 9-15 (1993).

96. Martz, E. (URL: http://www.umass.edu/microbio/chime/explorer/index.htm, 1999).

97. Verschelde, J. L. *et al.* Analysis of three human interleukin 5 structures suggests a possible receptor binding mechanism. *FEBS Lett* **424**, 121-6 (1998).

98. Crofts, A. R. *et al.* Pathways for proton release during ubihydroquinone oxidation by the bc(1) complex. *Proc Natl Acad Sci U S A* **96**, 10021-10026 (1999).

99. Crofts, A. R. & Berry, E. A. Structure and function of the cytochrome bc1 complex of mitochondria and photosynthetic bacteria. *Curr Opin Struct Biol* **8**, 501-9 (1998).

100. Bystrom, C. E., Pettigrew, D. W., Branchaud, B. P., P, O. B. & Remington, S. J. Crystal structures of Escherichia coli glycerol kinase variant S58-->W in complex with nonhydrolyzable ATP analogues reveal a putative active conformation of the enzyme as a result of domain motion. *Biochemistry* **38**, 3508-18 (1999).

101. Feese, M. D., Faber, H. R., Bystrom, C. E., Pettigrew, D. W. & Remington, S. J. Glycerol kinase from Escherichia coli and an Ala65-->Thr mutant: the crystal structures reveal conformational changes with implications for allosteric regulation. *Structure* **6**, 1407-18 (1998).

102. Thompson, A. B. *et al.* Aerosolized beclomethasone in chronic bronchitis. Improved pulmonary function and diminished airway inflammation. *Am Rev Respir Dis* **146**, 389-95 (1992).

103. Sykes, J. A., Thomas, M. J., Goldie, D. J. & Turner, G. M. Plasma lactoferrin levels in pregnancy and cystic fibrosis. *Clin Chim Acta* **122**, 385-93 (1982).

104. Krebs, W., Alexandrov, V., Wilson, C. & Gerstein, M. Normal Mode Analysis of Macromolecular Motions in a Database Framework: Developing Mode Concentration as a Useful Classifying Statistic. *in peer-review.* (2001).

105. Wilson, E. B., Decius, J. C. & Cross, P. C. *Molecular Vibrations* (McGraw-Hill, New York, 1955).

106. Ma, J., Sigler, P. B., Xu, Z. & Karplus, M. A dynamic model for the allosteric mechanism of GroEL. *J Mol Biol* **302**, 303-13. (2000).

107. Ma, J. & Karplus, M. The allosteric mechanism of the chaperonin GroEL: a dynamic analysis. *Proc Natl Acad Sci U S A* **95**, 8502-7. (1998).

108. Hayward, S., Kitao, A. & Berendsen, H. J. Model-free methods of analyzing domain motions in proteins from simulation: a comparison of normal mode analysis and molecular dynamics simulation of lysozyme. *Proteins* **27**, 425-37. (1997).

109. van der Spoel, D., de Groot, B. L., Hayward, S., Berendsen, H. J. & Vogel, H. J. Bending of the calmodulin central helix: a theoretical study. *Protein Sci* **5**, 2044-53. (1996).

110. Duncan, B. S. & Olson, A. J. Approximation and visualization of large-scale motion of protein surfaces. *J Mol Graph* **13**, 250-7. (1995).

111. Levitt, M., Sander, C. & Stern, P. S. Protein normal-mode dynamics: trypsin inhibitor, crambin, ribonuclease and lysozyme. *J Mol Biol* **181**, 423-47. (1985).

112. Brooks, B. & Karplus, M. Normal modes for specific motions of macromolecules: Application to the hinge-bending mode of lysozyme. *Proc. Natl. Acad. Sci. USA* **82**, 4995-4999 (1985).

113. Brooks, B. & Karplus, M. Harmonic dynamics of proteins: Normal modes and fluctations in bovine pancreatic trypsin inhibitor. *P.N.A.S. U.S.A.* **80**, 6571 (1983).

114. Levy, R. M. Computer simulations of macromolecular dynamics: models for vibrational spectroscopy and X-ray refinement. *Ann N Y Acad Sci* **482**, 24-43 (1986).

115. Levy, R. M., Rojas, O. d. l. L. & Friesner, R. A. Quasi-harmonic method for calculating vibrational spectra from classical simulations on multidimensional anharmonic potential surfaces. *Jour. Phys. Chem.* **88**, 4233 (1984).

116. Levy, R., Perahia, D. & Karplus, M. Molecular dynamics of an ff-helical polypeptide: temperature dependance and deviation from harmonic behavior. *Proc. Natl. Acad. Sci. USA* **79**, 1346-1350 (1982).

117. Miller, D. W. & Agard, D. A. Enzyme specificity under dynamic control: a normal mode analysis of alpha-lytic protease. *J Mol Biol* **286**, 267-78 (1999).

118. Marques, O. & Sanejouand, Y. H. Hinge-bending motion in citrate synthase arising from normal mode calculations. *Proteins* **23**, 557-60 (1995).

119. Thomas, A., Hinsen, K., Field, M. J. & Perahia, D. Tertiary and quaternary conformational changes in aspartate transcarbamylase: a normal mode study. *Proteins* **34**, 96-112 (1999).

120. Thomas, A., Field, M. J. & Perahia, D. Analysis of the low-frequency normal modes of the R state of aspartate transcarbamylase and a comparison with the T state modes. *J Mol Biol* **261**, 490-506 (1996).

121. Thomas, A., Field, M. J., Mouawad, L. & Perahia, D. Analysis of the low frequency normal modes of the T-state of aspartate transcarbamylase. *J Mol Biol* **257**, 1070-87 (1996).

122. Hinsen, K. Analysis of domain motions by approximate normal mode calculations. *Proteins* **33**, 417-29 (1998).

123. Voronoi, G. F. Nouveles applications des paramétres continus á la théorie de formes quadratiques. *J. Reine Angew. Math.* **134**, 198-287 (1908).

124. Bernal, J. D. & Finney, J. L. Random close-packed hard-sphere model II. Geometry of random packing of hard spheres. *Disc. Faraday Soc.* **43**, 62-69 (1967).

125. Richards, F. M. The Interpretation of Protein Structures: Total Volume, Group Volume Distributions and Packing Density. *J. Mol. Biol.* **82**, 1-14 (1974).

126. Li, A. J. & Nussinov, R. A set of van der Waals and coulombic radii of protein atoms for molecular and solvent-accessible surface calculation, packing evaluation, and docking. *Proteins* **32**, 111-27 (1998).

127. Tsai, J., Taylor, R., Chothia, C. & Gerstein, M. The packing density in proteins: standard radii and volumes. *J Mol Biol* **290**, 253-66. (1999).

128. Tsai, J. & Gerstein, M. Voronoi Calculations of Macromolecular Volumes: Sensitivity Analysis and Parameter Defense. *Bioinformatics* **in press** (2001).

129. Gerstein, M., Tsai, J. & Levitt, M. The volume of atoms on the protein surface: calculated from simulation, using Voronoi polyhedra. *J Mol Biol* **249**, 955-66. (1995).

130. Gellatly, B. J. & Finney, J. L. Calculation of Protein Volumes: An Alternative to the Voronoi Procedure. *J. Mol. Biol.* **161**, 305-322 (1982).

131. Richards, F. M. Packing Defects, Cavities, Volume Fluctuations, and Access to the Interior of Proteins. Including Some General Comments on Surface Area and Protein Structure. *Carlsberg. Res. Commun.* **44**, 47-63 (1979).

132. Duyckaerts, C. & Godefroy, G. Voronoi tessellation to study the numerical density and the spatial distribution of neurones. *J Chem Neuroanat* **20**, 83-92. (2000).

133. Fleming, P. J. & Richards, F. M. Protein packing: dependence on protein size, secondary structure and amino acid composition. *J Mol Biol* **299**, 487-98. (2000).

134. Kussell, E., Shimada, J. & Shakhnovich, E. I. Excluded volume in protein side-chain packing. *J Mol Biol* **311**, 183-93. (2001).

135. Likic, V. A. & Prendergast, F. G. Structure and dynamics of the fatty acid binding cavity in apo rat intestinal fatty acid binding protein. *Protein Sci* **8**, 1649-57. (1999).

136. Pontius, J., Richelle, J. & Wodak, S. J. Deviations from standard atomic volumes as a quality measure for protein crystal structures. *J Mol Biol* **264**, 121-36. (1996).

137. Quillin, M. L. & Matthews, B. W. Accurate calculation of the density of proteins. *Acta Crystallogr D Biol Crystallogr* **56**, 791-4. (2000).

138. Tsai, J. & Gerstein, M. Calculations of Protein Volumes: Sensitivity Analysis and Parameter Database. *Bioinformatics* **in press** (2001).

139. Thomas, P. D. & Dill, K. A. Statistical potentials extracted from protein structures: how accurate are they? *J Mol Biol* **257**, 457-69Methods Enzymol 1997;277:556-571 (1996).

140. Wodak, S. J., De, C. M. & Janin, J. Computer studies of interactions between macromolecules. *Prog. Biophys. Mol. Biol.* **49**, 29-63 (1987).

141. Karplus, M. & McCammon, J. A. The dynamics of proteins. *Sci. Am.* **254**, 42-51 (1986).

142. Friesner, R. A. & Dunietz, B. D. Large-Scale ab Initio Quantum Chemical Calculations on Biological Systems. *Acc Chem Res* **34**, 351-8. (2001).

143. Schlick, T. Ways & Means: time-trimming tricks for dynamic simulations: splitting force updates to reduce computational work. *Structure* **9**, R45-R53 (2001).

144. Gerstein, M. A structural census of genomes: comparing bacterial, eukaryotic, and archaeal genomes in terms of protein structure. *J Mol Biol* **274**, 562-76 (1997).

145. Gerstein, M. Patterns of protein-fold usage in eight microbial genomes: a comprehensive structural census. *Proteins* **33**, 518-34. (1998).

146. Teichmann, S., Chothia, C., Gerstein, M. Advances in structural genomics. *Curr. Opin. Struc. Biol.* **9**, 390-399 (1999).

147. Hegyi, H. & Gerstein, M. The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. *J Mol Biol* **288**, 147-64. (1999).

148. Hegyi, H. & Gerstein, M. Annotation Transfer for Genomics: Measuring Functional Divergence in Multi-domain Proteins. *Genome Res.* **in press** (2001).

149. Gerstein, M. B., Jansen, R., Johnson, T., Park, B. & Krebs, W. in *Rigidity theory and applications* (eds. Thorpe, M. F. & Duxbury, P. M.) 401-442 (Kluwer Academic/Plenum press, 1999).

150. Harrison, P. M., Echols, N. & Gerstein, M. B. Digging for dead genes: an analysis of the characteristics of the pseudogene population in the Caenorhabditis elegans genome. *Nucleic Acids Res* **29**, 818-30. (2001).

151. Hubbard, T. J. P., Murzin, A. G., Brenner, S. E. & Chothia, C. SCOP: a structural classification of proteins database. *Nucleic Acids Res* **25**, 236-9 (1997).

152. Branden, C. & Tooze, J. *Introduction to Protein Structure* (Garland Publishing Incorporated, New York, 1991).

153. Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nuc. Acid. Res.* **22**, 4673-4680 (1994).

154. Higgins, D. G., Thompson, J. D. & Gibson, T. J. Using CLUSTAL for multiple sequence alignments. *Methods Enzymol* **266**, 383-402 (1996).

155. Sonnhammer, E. L., Eddy, S. R., Birney, E., Bateman, A. & Durbin, R. Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Res* **26**, 320-2 (1998).

156. Krogh, A., Brown, M., Mian, I. S., Sjölander, K. & Haussler, D. Hidden Markov Models in Computational Biology: Applications to Protein Modelling. *J. Mol. Biol.* **235**, 1501-1531 (1994).

157. Eddy, S. R. Hidden Markov models. *Curr. Opin. Struc. Biol.* **6**, 361-365 (1996).

158. Eddy, S. R., Mitchison, G. & Durbin, R. Maximum Discrimination Hidden Markov Models of Sequence Consensus. *J. Comp. Bio.* **9**, 9-23 (1994).

159. Baldi, P., Chauvin, Y. & Hunkapiller, T. Hidden Markov Models of Biological Primary Sequence Information. *Proc. Natl. Acad. Sci.* **91** (1994).

160. Bairoch, A. & Boeckmann, B. The Swiss-Prot Protein-Sequence Data-Bank. *Nucl. Acids Res.* **20**, 2019-2022 (1992).

161. Holm, L. & Sander, C. The FSSP database of structurally aligned protein fold families. *Nuc. Acid Res.* **22**, 3600-3609 (1994).

162. Abola, E., Sussman, J., Prilusky, J. & Manning, N. Protein Data Bank archives of three-dimensional macromolecular structures. *Meth. Enz.* **277**, 556-571 (1997).

163. Orengo, C. A., Jones, D. T. & Thornton, J. M. Protein superfamilies and domain superfolds. *Nature* **372**, 631-634 (1994).

164. Altman, R. B., Abernethy, N. F. & Chen, R. O. Standardized representations of the literature: combining diverse sources of ribosomal data. *Ismb* **5**, 15-24 (1997).

165. Chen, R. O., Felciano, R. & Altman, R. B. RIBOWEB: linking structural computations to a knowledge base of published experimental data. *Ismb* **5**, 84-7 (1997).

166. Bairoch, A., Bucher, P. & Hofmann, K. The prosite database, its status in 1995. *Nucleic Acids Research* **24**, 189-196 (1996).

Tables

Table 1. Summary of ProtOr Type Set. This table gives ProtOr parameters for the various atom types used to model and compute the volumes of the amino acids.

Atom Type	num (173)	Vol. (Å^3)	Radii (Å)	Comments	Protein Atoms
C3H0s	20	8.72	1.61	carbonyl carbons with branching (mainchain carbonyls from residues with a C_β, so no gly carbon)	ALA_C,ARG_C,ASN_C,ASP_C,CSS_C,CYS_C,GLN_C,GLU_C,HIS_C,ILE_C,LEU_C, LYS_C,MET_C,PHE_C,PRO_C,SER_C,THR_C,TRP_C,TYR_C,VAL_C
C3H0b	13	9.70	1.61	carboxyl and carbonyl carbons w/o branching (side chain and glycine's) and aromatic carbons w/o hydrogen	ARG_CZ,ASN_CG,ASP_CG,GLN_CD,GLU_CD,GLY_C,HIS_CG,PHE_CG,TRP_CD2, TRP_CE2,TRP_CG,TYR_CG,TYR_CZ
C4H1s	18	13.17	1.88	aliphatic carbons with one hydrogen and branching from all three heavy atom bonds	ARG_CA,ASN_CA,ASP_CA,CSS_CA,CYS_CA,GLN_CA,GLU_CA,HIS_CA,ILE_CA, LEU_CA,LYS_CA,MET_CA,PHE_CA,SER_CA,THR_CA,TRP_CA,TYR_CA,VAL_CA
C4H1b	6	14.35	1.88	aliphatic carbons with one hydrogen and no branching through at least one heavy atom bond	ALA_CA,ILE_CB,LEU_CG,PRO_CA,THR_CB,VAL_CB
C3H1s	8	20.44	1.76	small aromatic carbons with one hydrogen	HIS_CD2,HIS_CE1,PHE_CD1,TRP_CD1,TYR_CD1,TYR_CD2,TYR_CE1,TYR_CE2
C3H1b	8	21.28	1.76	big aromatic carbons with one hydrogen	PHE_CD2,PHE_CE1,PHE_CE2,PHE_CZ,TRP_CE3,TRP_CH2,TRP_CZ2,TRP_CZ3
C4H2s	21	23.19	1.88	aliphatic carbons with two hydrogens, small	ARG_CB,ARG_CD,ARG_CG,ASN_CB,ASP_CB,GLN_CB,GLN_CG,GLU_CB,GLU_CG, GLY_CA,HIS_CB,LEU_CB,LYS_CB,LYS_CD,LYS_CG,MET_CB,PHE_CB,PRO_CD, SER_CB,TRP_CB,TYR_CB
C4H2b	7	24.26	1.88	aliphatic carbons with two hydrogens, big	CSS_CB,CYS_CB,ILE_CG1,LYS_CE,MET_CG,PRO_CB,PRO_CG
C4H3u	9	36.73	1.88	aliphatic carbons with three hydrogens, i.e. methyl groups	ALA_CB,ILE_CD1,ILE_CG2,LEU_CD1,LEU_CD2,MET_CE, THR_CG2,VAL_CG1,VAL_CG2
N3H0u	1	8.65	1.64	imide nitrogens (only member is Pro N)	PRO_N
N3H1s	20	13.62	1.64	amide nitrogens with one hydrogen (all other mainchain N's)	ALA_N,ARG_N,ASN_N,ASP_N,CSS_N,CYS_N,GLN_N,GLU_N,GLY_N,HIS_N,ILE_N, LEU_N,LYS_N,MET_N,PHE_N,SER_N,THR_N,TRP_N,TYR_N,VAL_N
N3H1b	4	15.72	1.64	amide nitrogens with one hydrogen (on sidechains)	ARG_NE,HIS_ND1,HIS_NE2,TRP_NE1
N3H2u	4	22.69	1.64	all amide nitrogens with 2 hydrogens (only on sidechains)	ARG_NH1,ARG_NH2,ASN_ND2,GLN_NE2
N4H3u	1	21.41	1.64	amide nitrogen charged, with 3 hydrogens	LYS_NZ
O1H0u	27	15.91	1.42	all oxygens in carboxyl or carbonyl groups (no distinction made between oxygens in carboxyl group)	ALA_O,ARG_O,ASN_O,ASN_OD1,ASP_O,ASP_OD1,ASP_OD2,CSS_O,CYS_O, GLN_O,GLN_OE1,GLU_O,GLU_OE1,GLU_OE2,GLY_O,HIS_O,ILE_O,LEU_O,LYS_O, MET_O,PHE_O,PRO_O,SER_O,THR_O,TRP_O,TYR_O,VAL_O
O2H1u	3	17.98	1.46	all hydroxyl atoms	SER_OG,THR_OG1,TYR_OH
S2H0u	2	29.17	1.77	sulfurs with no hydrogens	CSS_SG,MET_SD
S2H1u	1	36.75	1.77	sulfurs with one hydrogen	CYS_SG

Table 2. ProtOr Residue Volumes. This table gives the volume of the various amino acids as computed by ProtOr using the parameter set given in Table 2. Note that reduced cysteine (CYS) was considered distinct from disulfide bonded cysteine (CSS).

Amino Acid	ProtOr Volume/Å^3
GLY	63.8
ALA	89.3
VAL	138.2
LEU	163.1
ILE	163.0
PRO	121.3
MET	165.8
PHE	190.8
TYR	194.6
TRP	226.4
SER	93.5
THR	119.6
ASN	122.4
GLN	146.9
CYS	112.8
CSS	102.5
HIS	157.5
GLU	138.8
ASP	114.4
ARG	190.3
LYS	165.1

Table 3. Computational Techniques for Studying Protein Motions. This table was based in part on a figure in Schlick[143].

Technique	**Pros**	**Cons**	**CPU Complexity**
Molecular dynamics (MD)	Continuous actions	Expensive; short time span	10 picoseconds = weeks for 50,000 atoms
Targeted MD (TMD)	Connection between two states; useful for ruling ut steric clashes	Not necessarily physical	Same as MD for each step
Continuum salvation	Mean-force potential approximates environment and reduces model's cost; useful information on ionic atmosphere and intermolecular associations	Approximate	Technique-dependent; can be as expensive as MD, but number of variables is reduced
Brownian dynamics (BD)	Large-scale and long-time motion	Approximate hydrodynamics; limited to systems with small relative inertia	Days for long DNA (1000s of base pairs)
Monte Carlo (MC)	Large-scale sampling; useful statistics	Move definitions are difficult	Hours of a million configurations
Minimization	Valuable equilibrium information; experimental constraints can be incorporated	No dynamic information	Minutes to hours for biomolecules
Stochastic Path Approach	Filtering of high-frequency motion; approximate long-time trajectories	Expensive (global optimization of entire trajectory)	1 picosecond approximate trajectory (1000 simulated annealing steps) = 1 day on 100 processors for 25,000 atoms
Normal mode analysis	Fast with interesting statistics, but potential unrealistic; may have large memory requirements	No dynamic information	Seconds to minutes to hours depending on problem and implementation

Table 4. Profile of the amino acid composition in linker sequences for every single linker position in detail compared with the PDB40 averages. A linker has been arbitrarily defined as the 16 residue region centered around the peptide bond (between positions 8 and 9) linking two domains. Positions where the amino acid frequency is less than the PDB40 average have a gray background. (MANUSCRIPT NOTE: YELLOW ENTRIES MAY BE BOLDED INSTEAD IN JOURNAL PROOF.)

																	PDB40 average
A	8.6	7.8	4.7	5.6	6.0	8.6	9.5	5.6	4.7	6.5	5.6	7.3	6.9	9.1	9.5	9.9	8.4
V	6.0	8.2	8.2	6.0	8.2	5.6	9.1	6.0	8.2	4.7	6.0	4.7	7.3	9.1	5.2	8.6	7.0
F	4.7	3.9	6.5	3.5	2.6	2.6	6.0	2.6	4.7	3.0	4.3	6.0	5.2	4.3	4.3	5.6	4.0
P	3.9	6.5	6.0	6.0	5.2	9.1	6.9	10.8	9.1	10.3	9.9	6.0	8.6	2.6	4.7	3.5	4.7
M	4.7	1.3	1.3	2.6	2.6	0.0	1.7	1.7	4.3	3.0	1.3	1.3	2.2	1.7	3.0	3.0	2.2
I	5.6	3.5	7.3	6.5	3.9	6.0	3.9	3.5	5.2	6.9	4.7	2.6	4.7	8.6	5.6	6.0	5.6
L	11.6	9.1	11.2	6.0	16.4	7.3	4.3	6.5	8.2	3.5	7.3	5.2	7.3	6.5	10.3	7.8	8.5
D	4.7	6.5	6.0	3.9	6.0	4.7	5.6	8.6	4.3	3.9	3.5	7.3	6.9	7.3	4.3	5.6	6.0
E	5.2	5.2	3.9	6.5	4.7	4.7	7.8	4.7	6.5	4.3	6.5	9.1	7.3	5.2	8.6	5.6	6.3
K	5.2	6.5	3.9	5.6	5.2	6.9	4.7	4.7	6.0	7.8	3.9	6.5	5.2	5.2	3.0	7.8	5.9
R	5.2	3.9	4.7	9.1	6.5	5.2	5.2	5.6	5.6	4.7	6.0	5.2	5.2	4.7	3.0	4.3	4.8
S	7.8	6.0	5.2	6.9	6.5	8.2	6.9	6.5	3.5	6.0	9.5	7.8	4.3	3.9	8.6	4.7	6.0
T	4.7	5.6	3.0	5.6	6.5	9.5	6.9	6.0	6.5	11.2	7.3	6.5	6.0	4.7	8.2	3.5	5.8
Y	2.2	3.9	6.5	3.0	3.5	2.2	2.6	3.5	2.2	3.9	2.6	2.2	3.0	3.5	3.5	4.3	3.7
H	1.7	3.5	3.0	3.5	3.5	2.6	3.5	2.2	2.2	0.9	1.7	2.2	1.7	2.6	1.3	2.2	2.2
C	1.7	2.6	0.9	1.3	1.7	2.6	0.4	2.2	0.9	1.3	4.7	1.7	1.7	3.9	0.4	0.9	1.7
N	4.7	3.9	3.5	6.5	3.0	4.3	2.6	3.0	5.6	5.2	3.5	6.5	3.9	6.0	3.0	5.6	4.6
Q	3.9	5.2	3.5	5.2	2.6	0.9	3.0	2.2	3.5	4.7	3.5	2.2	6.5	4.3	4.3	4.7	3.8
W	1.3	0.9	0.9	2.6	0.4	0.9	0.4	0.9	0.4	1.3	0.0	1.3	0.4	0.9	2.2	0.9	1.5
G	6.0	6.0	9.9	4.3	5.2	8.2	9.1	13.4	8.2	6.9	8.2	8.6	5.6	6.0	6.9	5.6	7.8
X	0.4	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	

Table 5. P-values for the profile of the amino acid composition of linker sequences for every single position in the linkers. P-values less than 0.05 are represented by a gray background. The low P-values for proline in positions 6 to 11 are most conspicuous. The classification according to the attributes hydrophobic, charged, and polar (Branden and Tooze[152]) does not provide a satisfactory explanation for the observed levels of amino acids (see also Figure 10). (MANUSCRIPT NOTE: YELLOW ENTRIES MAY BE BOLDED INSTEAD IN JOURNAL PROOF.)

A	.908	.728	4e-2	.125	.196	.908	.562	.125	4e-2	.293	.125	.561	.415	.729	.562	.416	hydrophobic
V	.577	.481	.481	.577	.481	.417	.224	.577	.481	.184	.577	.184	.841	.224	.285	.338	
F	.598	.911	.059	.666	.276	.276	.126	.276	.598	.449	.836	.126	.393	.836	.836	.235	
P	.573	.207	.346	.346	.737	2e-3	.114	5e-5	2e-3	1e-4	3e-4	.346	4e-3	.134	.971	.385	
M	1e-2	.366	.366	.717	.717	2e-2	.637	.637	3e-2	.433	.366	.366	.961	.637	.433	.433	
I	.990	.155	.267	.585	.257	.793	.257	.155	.772	.408	.571	4e-2	.571	5e-2	.990	.793	
L	.084	.754	.136	.186	3e-5	.541	2e-2	.280	.882	6e-3	.541	.071	.541	.280	.312	.705	
D	.442	.750	.966	.185	.966	.442	.821	.089	.296	.185	.108	.389	.556	.389	.296	.821	charged
E	.476	.476	.127	.936	.327	.327	.384	.327	.936	.211	.936	.092	.545	.476	.158	.653	
K	.638	.730	.194	.842	.638	.538	.457	.457	.945	.243	.194	.730	.638	.638	.061	.243	
R	.793	.530	.974	2e-3	.240	.793	.793	.575	.575	.974	.389	.793	.793	.974	.215	.742	
S	.269	.990	.599	.578	.774	.166	.578	.774	.101	.990	2e-2	.269	.283	.176	.095	.425	polar
T	.498	.897	.069	.897	.673	2e-2	.485	.886	.673	5e-4	.328	.673	.886	.498	.121	.127	
Y	.234	.864	2e-2	.619	.872	.234	.402	.872	.234	.864	.402	.234	.619	.872	.872	.612	
H	.619	.237	.455	.237	.237	.740	.237	.939	.939	.166	.619	.939	.619	.740	.354	.939	
C	.997	.336	.345	.647	.997	.336	.139	.634	.345	.647	2e-2	.997	.997	2e-2	.139	.345	
N	.942	.597	.404	.193	.251	.820	.143	.251	.500	.710	.404	.193	.597	.326	.251	.500	
Q	.937	.281	.804	.281	.359	2e-2	.562	.206	.804	.460	.804	.206	3e-2	.684	.684	.460	
W	.810	.459	.459	.193	.197	.459	.197	.459	.197	.810	.055	.810	.197	.459	.452	.459	
G	.324	.324	.233	5e-2	.139	.823	.482	1e-3	.823	.621	.823	.643	.218	.324	.621	.218	
X	.717	.717	.752	.752	.752	.752	.752	.752	.717	.752	.752	.752	.752	.752	.752	.752	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	

Table 6. Example of protein flexible linker consensus sequences extracted from the Macromolecular Movements Database. The database contains residue selections for known hinge regions (flexible linkers) culled from the scientific literature. Sixteen of these residue selections were then “grown” slightly in both directions according to a fixed protocol. Each selection was assigned a linker ID, which is based either on a PDB ID or on the macromolecular movements database motion ID plus possible an optional additional numeric suffix to identify the specific residue selection used. A FASTA search with a cutoff of 0.01 was then performed on each sequence to obtain near homologues. The consensus sequence corresponding to each linker ID is given here.

Linker ID	Linker Consensus Sequence
4cln	MARKMKDTDSE
6ldh	AGARQQEGESRLNLVQRNVNIFKF
adenkin1	VPFEVI
adenkin2	LRLTA
adenkin3	GEPLIQRDDDKE
adenkin4	AYHAQTE
anxbreat	MKGAGT
anxtrp1	YEAGELKWG
anxtrp2	EETIDRET
dt	LFQVVHNS
enolase	GASTGIY
enolase2	SDKS
lfh_hinge1	QTHY
lfh_hinge2	RVPS
ras	AGQEEYSAMRDQYMR
tbsv	PQPTNTL

Table 7. Example of FASTA results. This table gives an example of sequences that might be obtained from a FASTA run on a known flexible linker sequence. In this case, the output of one FASTA run on the OWL database using the flexible linker region from Calmodulin (4cln) with a cutoff (e-value) of 0.001

OWL	**ID**
CALN_CHICK	MARKMKDTDSE
MUSCAMC	MARKMKDTDSE
CALM_PATSP	MARKMKDTDSE
CALM_PYUSP	MARKMKDTDSE
CALM_METSE	MARKMKDTDSE
CALM_STIJA	MARKMKDTDSE
CALM_HUMAN	MARKMKDTDSE
CALM_DROME	MARKMKDTDSE
HSCAM3X1	MARKMKDTDSE
CALM_EMENI	MARKMKDTDSE
CALM_NEUCR	MARKMKDTDSE
CALM_ELEEL	MAKKMKDTDSE
NEUCLMDLN	MARKMKDTDSE
SSO4B01	MARKMKDTDSE
CALL_ARBPU	MARKMKETDSE
CALM_PLECO	MARKMRDTDSE
CALL_HUMAN	MARKMKDTDNE
CALS_CHICK	MARKMRDSDSE
CALM_PHYIN	MARKMKDTDSE
CALM_PNECA	MARKMKDVDSE
CALM_TRYBB	MARKMQDSDSE
CALM_TRYCR	MARKMQDSDSE
S53019	MARKMKDTDSE
TRBCMRSG	MARKMQDSDSE
CALM_HORVU	MARKMKDTDSE
JC1033	MARKMKDTDSE
CAL1_PETHY	MARKMKDTDSE
CAL6_ARATH	MARKMKDTDSE

Table 8. Preliminary Flexible Linker Propensity Scale. A FASTA search with a cutoff of 0.01 was performed on sixteen flexible linker sequences, as described in the text. Amino acid frequency in the flexible linker sequences and their near homologues obtained in the FASTA search were tabulated and divided by the amino acid sequence frequency in the PDB to obtain the preliminary propensities given in this table. (The high propensity shown for methionine may be an artifact arising from methionine's presence as the first residue in many proteins.)

Residue	**Propensity**
A	1.3268
C	0.1097
D	1.1684
E	1.4702
F	0.5624
G	1.2972
H	0.4806
I	0.4462
K	1.0519
L	0.5303
M	2.6603
N	0.7729
P	0.4051
Q	1.8076
R	1.8013
S	0.8269
T	0.9002
V	0.6865
W	0.308
Y	1.3375

Figures

Figure 1A. Protein chemistry elucidated through the motion of the protein's moving parts. This figure shows the use of the "flickerbook" feature of the morph server associated with the motions database. Each boxed image of the protein represents a frame in the movie of the motion produced by the server. Frames are sequential in time, from the bottom to the top of the page, and then left to right. This particular flickerbook is a movie of the apical domain motion of GroEL.

Readers can photocopy this figure, cut along the edges of the boxes to produce ten still frames, and then bind these ten frames into a booklet (using, say, a stapler) to produce a "flickerbook." The movie may then be visualized by rapidly flipping the pages of the flickerbook to create the illusion of motion. Flickerbooks represent a low-tech means of displaying protein movies when Internet access to the server is not available.

The high-tech means of seeing this movie and other movies is to access the website for the motions database, http://www.molmovdb.org. (This particular movie has 71095-15408 as its movie ID; it is referenced under the movies for "GroEL.")

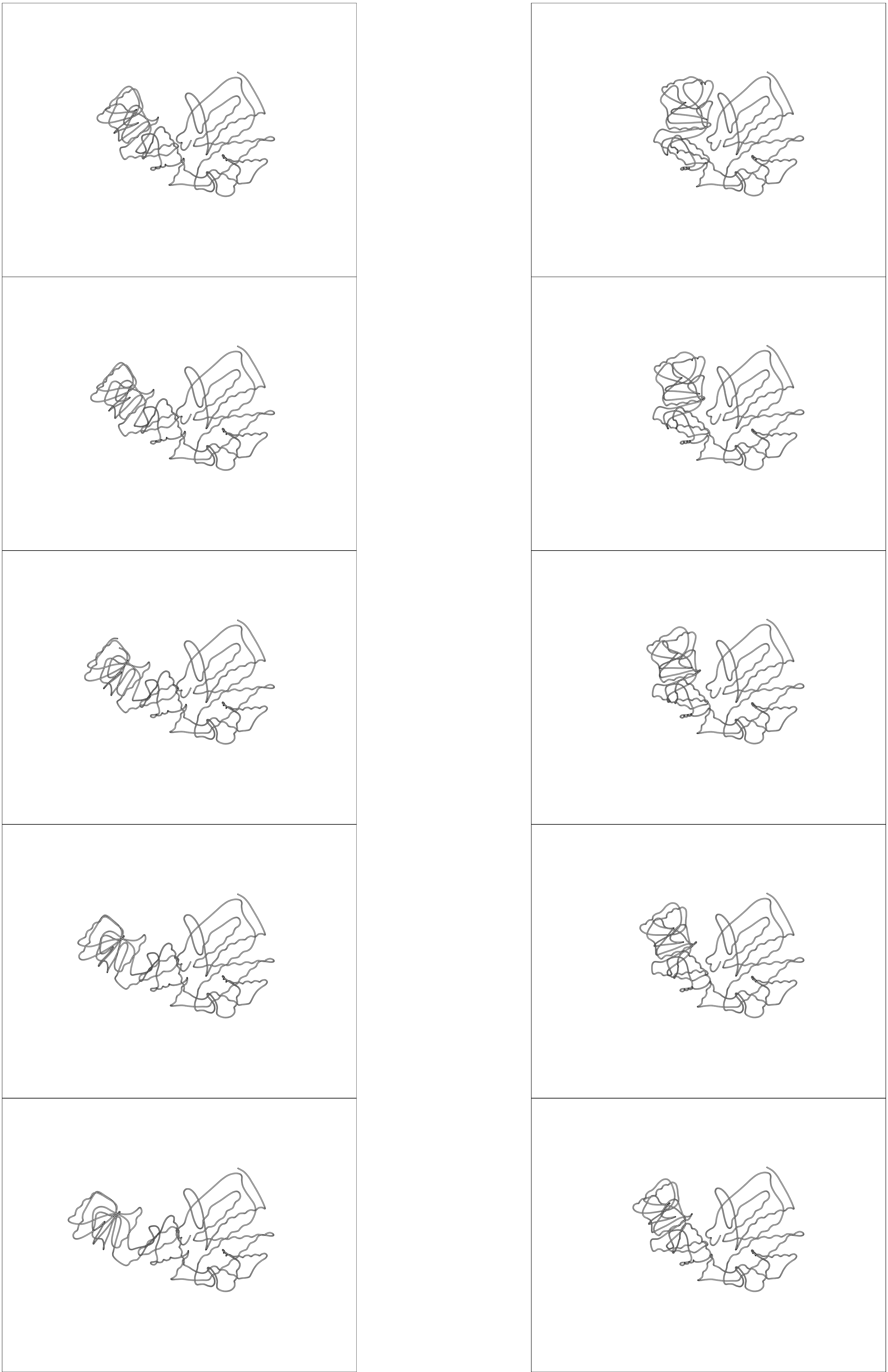


Figure 1B. This is a flickerbook of NtrC, a nitrogen-sensing regulatory protein. The motion and NMR determination of both structures are described in Volkman et al.[47] This particular movie is available for viewing viewing from within the online text of the article on the *Science* magazine website (http://www.sciencemag.org). It is also available for viewing as movie ID 7kern from the morph server website (URL: http://www.molmovdb.org/molmovdb/cgi-bin/morph.cgi?ID=7kern).

Normally, web users can generate flickerbooks like this (as well as Internet-accesible protein movies) by supplying morph server component of the motion database (http://www.molmovdb.org) with the PDB IDs or solved structures of the conformations. However, in this case, NMR provided additional experimental information on changes in protein secondary structure as well as a more precise identification of the mobile atoms. Because this experimental information is not normally readily available, at the time the online interface to the morph server did not provide an easy means for users to input this surplus information to the server's morphing engine. For this reason, this particular movie actually represents a custom morph in which the extra experimental data was manually fed into the server by its authors.

Readers without Internet access to the database can view this movie using only scissors and a stapler by following the instructions in the caption to Figure 1A.

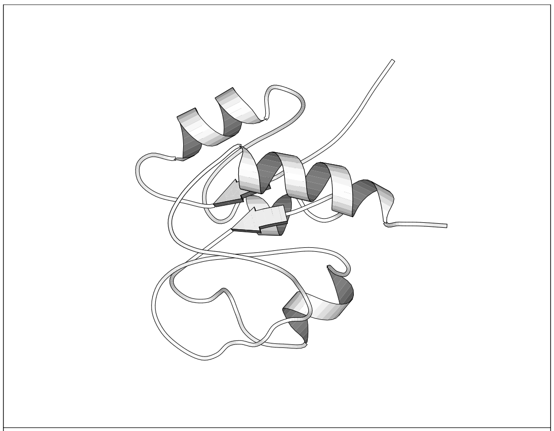
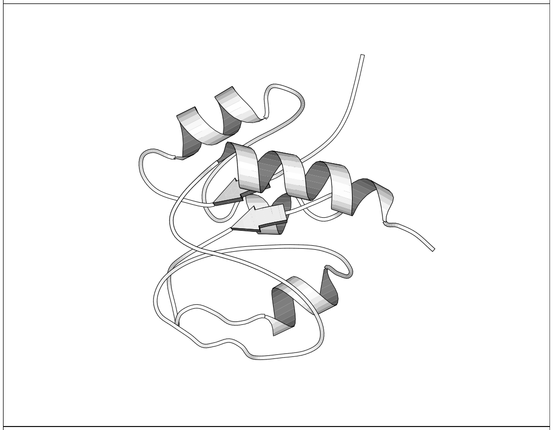
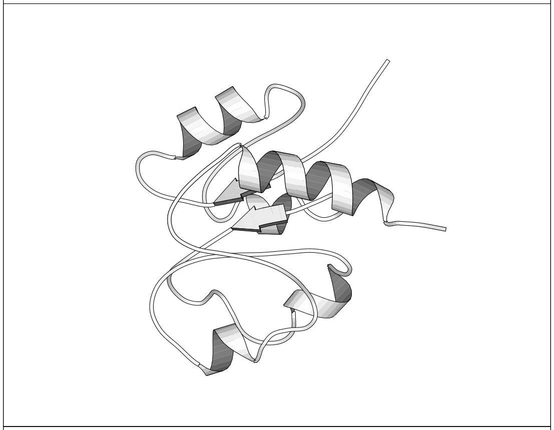
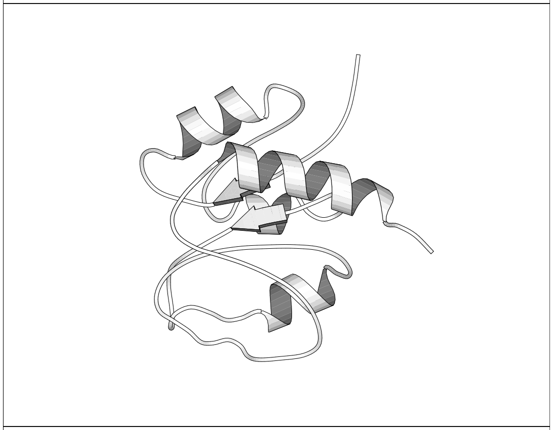
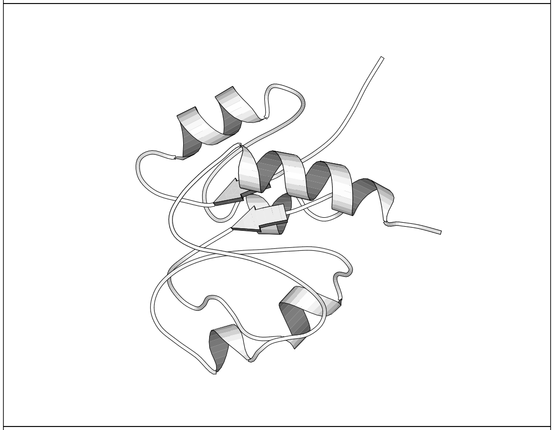
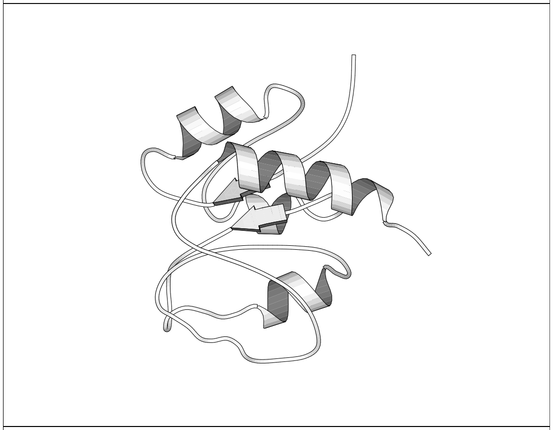
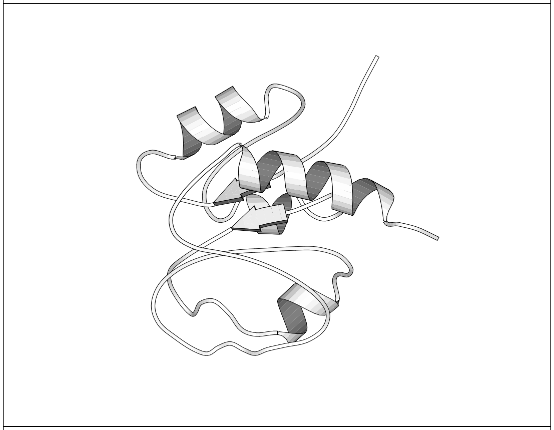
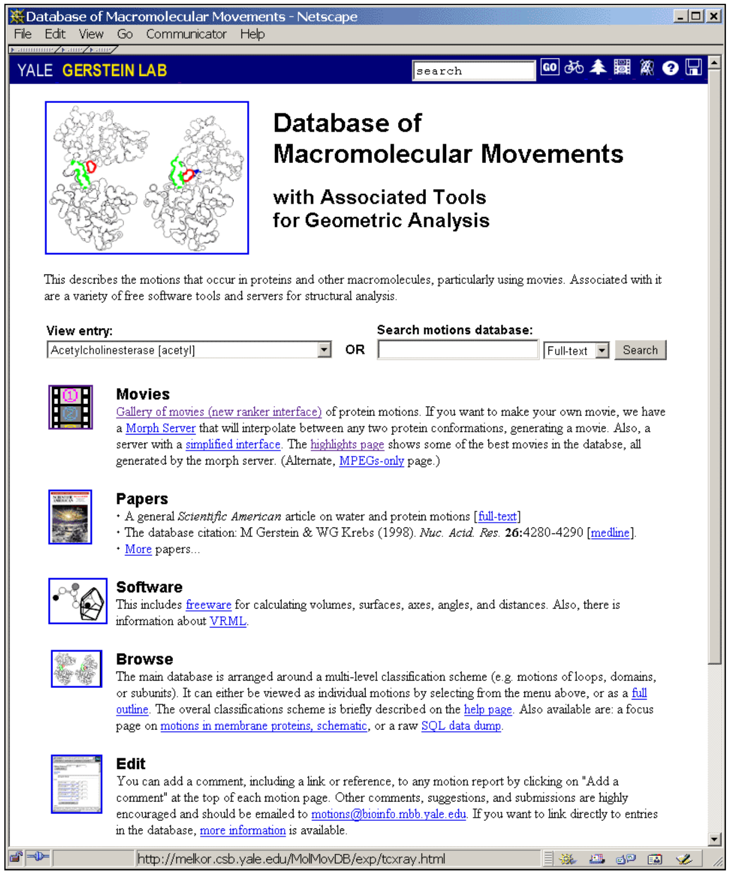








Figure 2. The Motions Database on the Web. LEFT shows the World Wide Web "home page" of the database. One can type keywords in the small box at the top to retrieve entries. MIDDLE shows a new 'ranker' interface to the motions database. Movies (and their associated motions entries) can be sorted on the basis of dozens of useful statistics, including the size of the motion in angstroms, rotation of the motion around the hinge in degrees, date of submission, as well as energy statistics associated with the interpolated pathway. RIGHT shows a protein 'morph' (animated representation) for calmodulin referenced by the database, along with the start of the database entry. Graphics and movies are accessed by clicking on an entry page. The main URL for the database is http://www.molmovdb.org. Beneath this are pages listing all the current movies, graphics illustrating the use of VRML to represent endpoints, and an automated submission form to add entries to the database. The database has direct links to the PDB for current entries (http://www.pdb.bnl.gov); the obsolete database (http://pdbobs.sdsc.gov) for obsolete entries; scop (http://scop.mrc-lmb.cam.ac.uk); Entrez/PubMed (http://www.ncbi.nlm.nih.gov/PubMed/medline.html); and LPFC (http://smi-web.stanford.edu/projects/helix/LPFC). Through these links one can easily connect to other common protein databases such Swiss-Prot, Pro-Site, CATH, RiboWeb, and FSSP[24,160-166].

(MANUSCRIPT NOTE: Because this Figure may be hard to reproduce, we have also included the LEFT and MIDDLE Figures as full-page figures on the two pages following the Figure.)


Database of Macromolecular Movements - Netscape
File Edit View Go Communicator Help
YALE GERSTEIN LAB
search
GO

Database of Macromolecular Movements

with Associated Tools for Geometric Analysis

This describes the motions that occur in proteins and other macromolecules, particularly using movies. Associated with it are a variety of free software tools and servers for structural analysis.

View entry: Acetylcholinesterase [acetyl] **OR** **Search motions database:** Full-text Search

Movies

Gallery of movies (new ranker interface) of protein motions. If you want to make your own movie, we have a Morph Server that will interpolate between any two protein conformations, generating a movie. Also, a server with a simplified interface. The highlights page shows some of the best movies in the databse, all generated by the morph server. (Alternate, MPEGs-only page.)

Papers

- A general *Scientific American* article on water and protein motions [full-text]
- The database citation: M Gerstein & WG Krebs (1998). *Nuc. Acid. Res.* **26:**4280-4290 [medline].
- More papers...

Software

This includes freeware for calculating volumes, surfaces, axes, angles, and distances. Also, there is information about VRML.

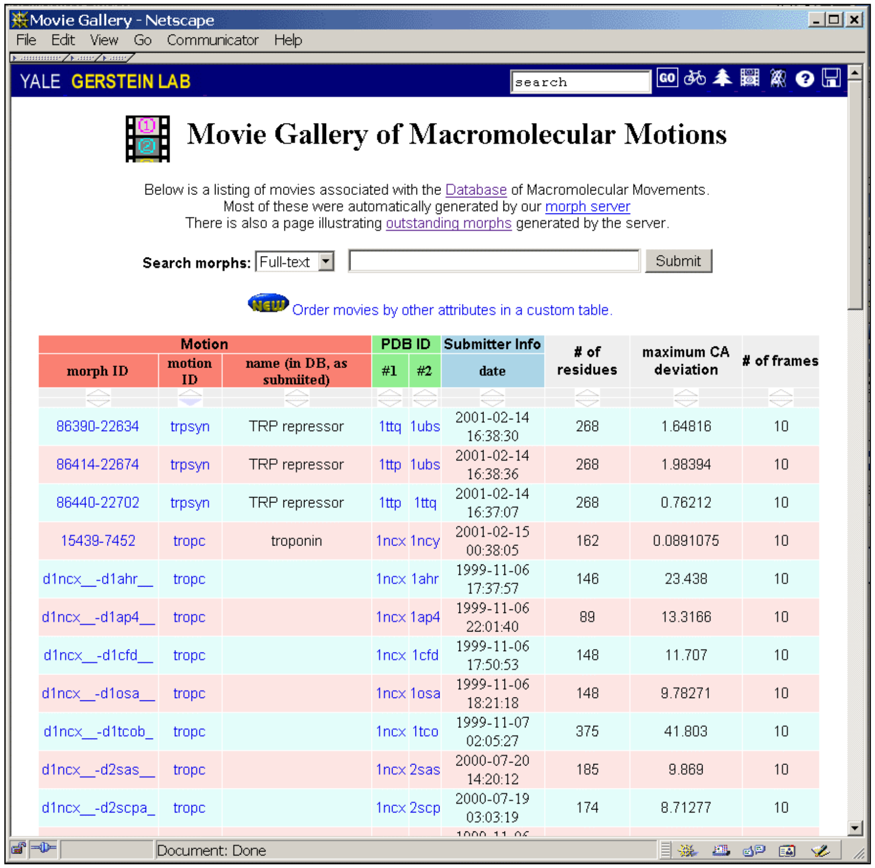
Browse

The main database is arranged around a multi-level classification scheme (e.g. motions of loops, domains, or subunits). It can either be viewed as individual motions by selecting from the menu above, or as a full outline. The overal classifications scheme is briefly described on the help page. Also available are: a focus page on motions in membrane proteins, schematic, or a raw SQL data dump.

Edit

You can add a comment, including a link or reference, to any motion report by clicking on "Add a comment" at the top of each motion page. Other comments, suggestions, and submissions are highly encouraged and should be emailed to motions@bioinfo.mbb.yale.edu. If you want to link directly to entries in the database, more information is available.

http://melkor.csb.yale.edu/MolMovDB/exp/tcxray.html



Movie Gallery of Macromolecular Motions

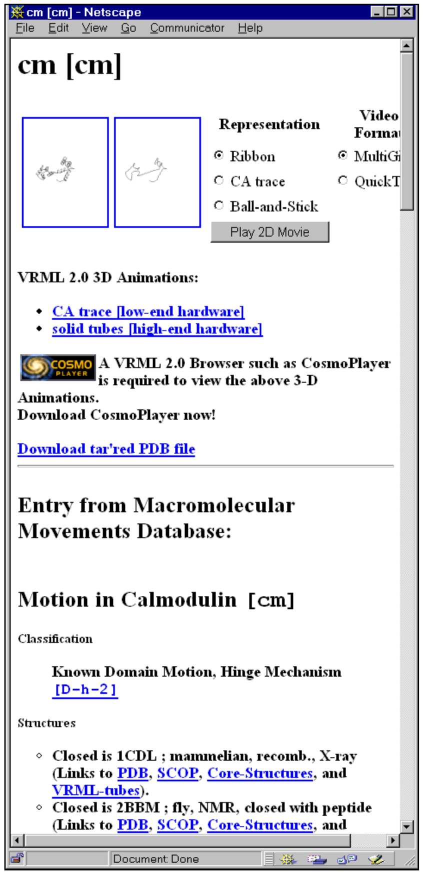
Below is a listing of movies associated with the Database of Macromolecular Movements.
Most of these were automatically generated by our morph server.
There is also a page illustrating outstanding morphs generated by the server.

Search morphs: Full-text Submit

Order movies by other attributes in a custom table.

Motion			PDB ID		Submitter Info	# of residues	maximum CA deviation	# of frames
morph ID	motion ID	name (in DB, as submiited)	#1	#2	date			
86390-22634	trpsyn	TRP repressor	1ttq	1ubs	2001-02-14 16:38:30	268	1.64816	10
86414-22674	trpsyn	TRP repressor	1ttp	1ubs	2001-02-14 16:38:36	268	1.98394	10
86440-22702	trpsyn	TRP repressor	1ttp	1ttq	2001-02-14 16:37:07	268	0.76212	10
15439-7452	tropc	troponin	1ncx	1ncy	2001-02-15 00:38:05	162	0.0891075	10
d1ncx__-d1ahr__	tropc		1ncx	1ahr	1999-11-06 17:37:57	146	23.438	10
d1ncx__-d1ap4__	tropc		1ncx	1ap4	1999-11-06 22:01:40	89	13.3166	10
d1ncx__-d1cfd__	tropc		1ncx	1cfd	1999-11-06 17:50:53	148	11.707	10
d1ncx__-d1osa__	tropc		1ncx	1osa	1999-11-06 18:21:18	148	9.78271	10
d1ncx__-d1tcob_	tropc		1ncx	1tco	1999-11-07 02:05:27	375	41.803	10
d1ncx__-d2sas__	tropc		1ncx	2sas	2000-07-20 14:20:12	185	9.869	10
d1ncx__-d2scpa_	tropc		1ncx	2scp	2000-07-19 03:03:19	174	8.71277	10

Document: Done



cm [cm]

Representation

● Ribbon
○ CA trace
○ Ball-and-Stick

Play 2D Movie

Video Forma

● MultiGi
○ QuickT

VRML 2.0 3D Animations:

- CA trace [low-end hardware]
- solid tubes [high-end hardware]

A VRML 2.0 Browser such as CosmoPlayer is required to view the above 3-D Animations.
Download CosmoPlayer now!

Download tar'red PDB file

Entry from Macromolecular Movements Database:

Motion in Calmodulin [cm]

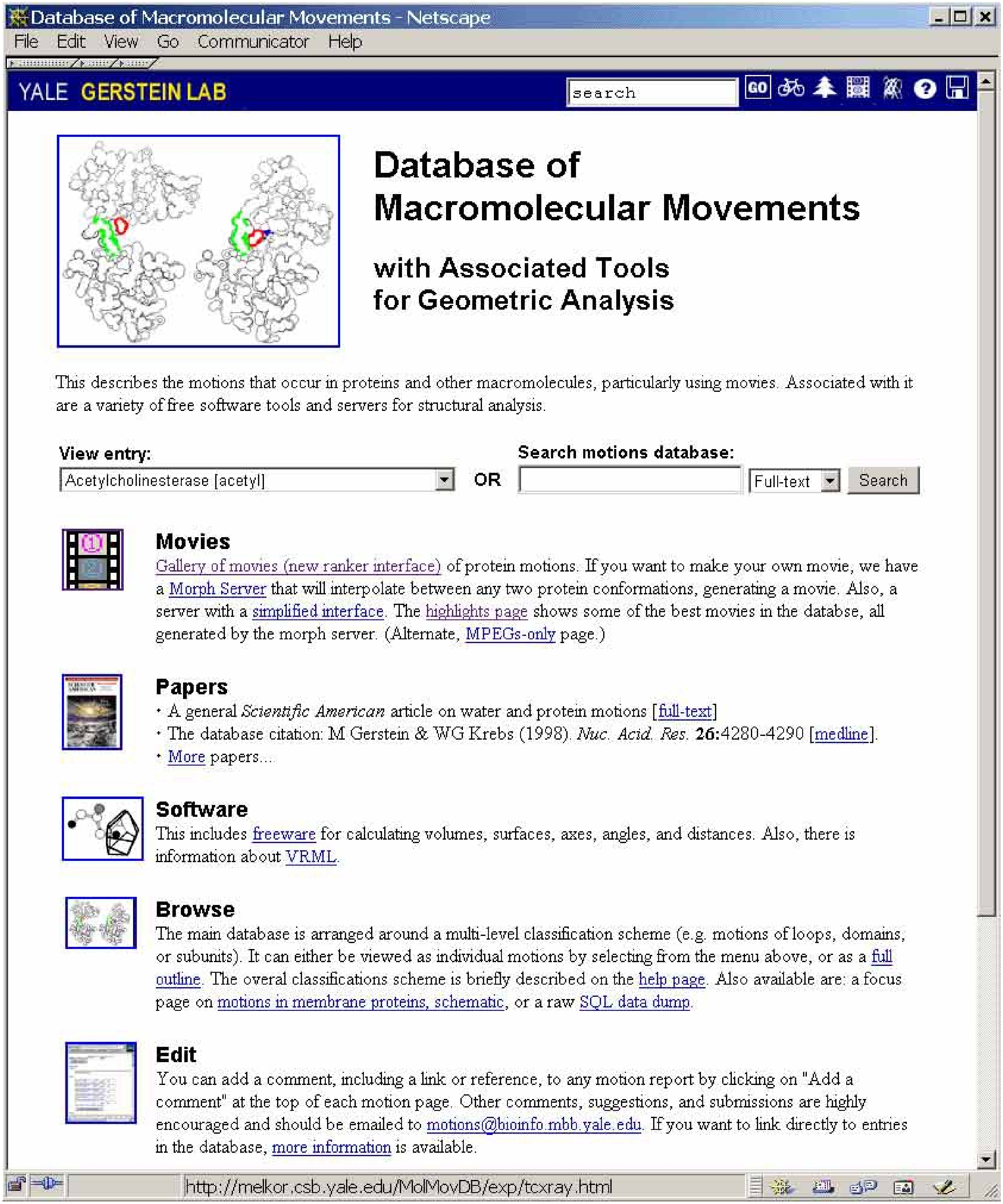
Classification

Known Domain Motion, Hinge Mechanism [D-h-2]

Structures

- **Closed is 1CDL ; mammelian, recomb., X-ray (Links to PDB, SCOP, Core-Structures, and VRML-tubes).**
- **Closed is 2BBM ; fly, NMR, closed with peptide (Links to PDB, SCOP, Core-Structures, and**

Document: Done


Database of Macromolecular Movements - Netscape
File Edit View Go Communicator Help
YALE GERSTEIN LAB
search
GO
http://melkor.csb.yale.edu/MolMovDB/exp/tcxray.html

Database of Macromolecular Movements

with Associated Tools for Geometric Analysis

This describes the motions that occur in proteins and other macromolecules, particularly using movies. Associated with it are a variety of free software tools and servers for structural analysis.

View entry: Acetylcholinesterase [acetyl] **OR** **Search motions database:** Full-text Search

Movies

Gallery of movies (new ranker interface) of protein motions. If you want to make your own movie, we have a Morph Server that will interpolate between any two protein conformations, generating a movie. Also, a server with a simplified interface. The highlights page shows some of the best movies in the database, all generated by the morph server. (Alternate, MPEGs-only page.)

Papers

- A general *Scientific American* article on water and protein motions [full-text]
- The database citation: M Gerstein & WG Krebs (1998). *Nuc. Acid. Res.* **26:**4280-4290 [medline].
- More papers...

Software

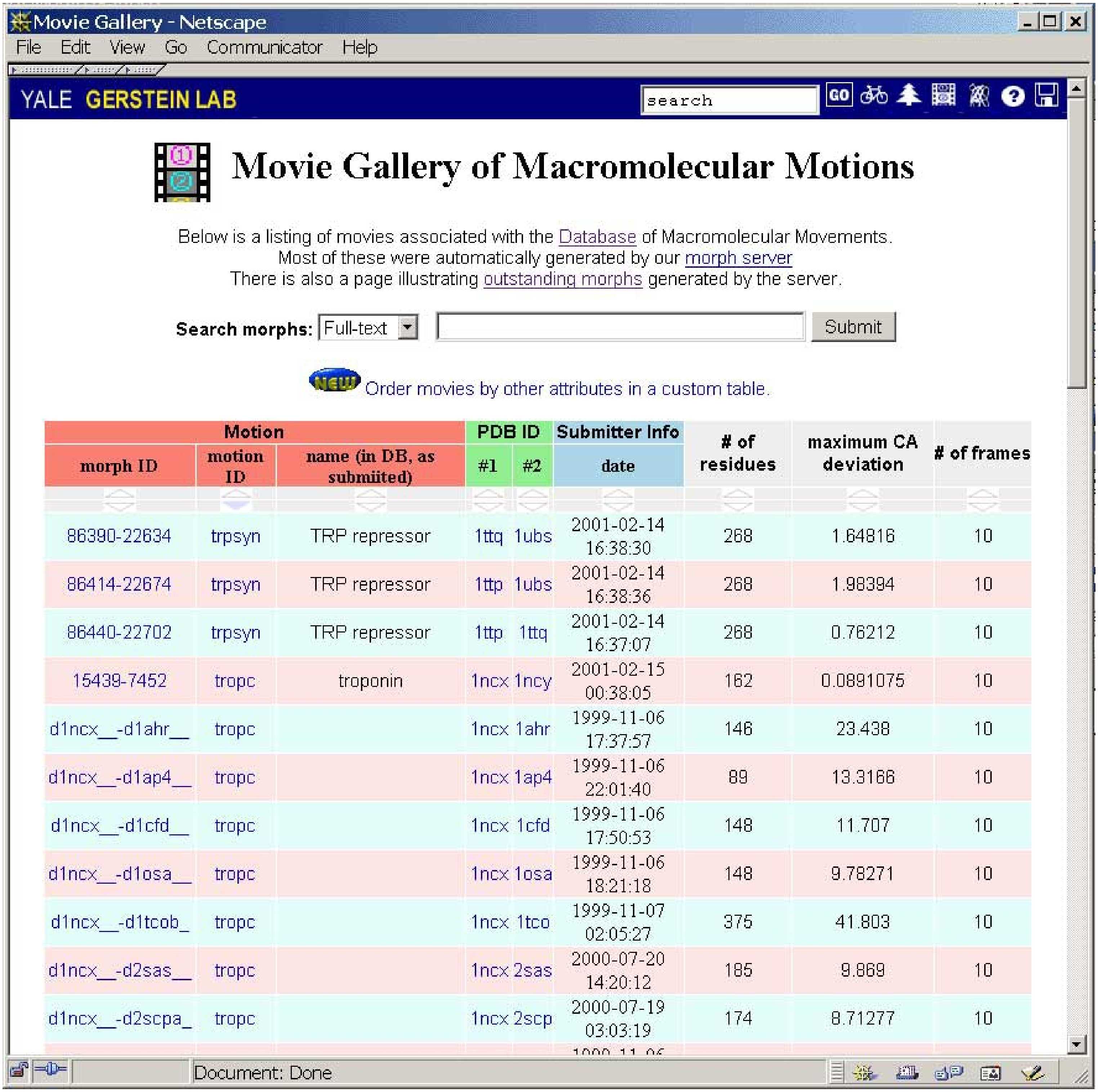
This includes freeware for calculating volumes, surfaces, axes, angles, and distances. Also, there is information about VRML.

Browse

The main database is arranged around a multi-level classification scheme (e.g. motions of loops, domains, or subunits). It can either be viewed as individual motions by selecting from the menu above, or as a full outline. The overal classifications scheme is briefly described on the help page. Also available are: a focus page on motions in membrane proteins, schematic, or a raw SQL data dump.

Edit

You can add a comment, including a link or reference, to any motion report by clicking on "Add a comment" at the top of each motion page. Other comments, suggestions, and submissions are highly encouraged and should be emailed to motions@bioinfo.mbb.yale.edu. If you want to link directly to entries in the database, more information is available.



Movie Gallery of Macromolecular Motions

Below is a listing of movies associated with the Database of Macromolecular Movements.
Most of these were automatically generated by our morph server
There is also a page illustrating outstanding morphs generated by the server.

Search morphs: Full-text ▾ Submit

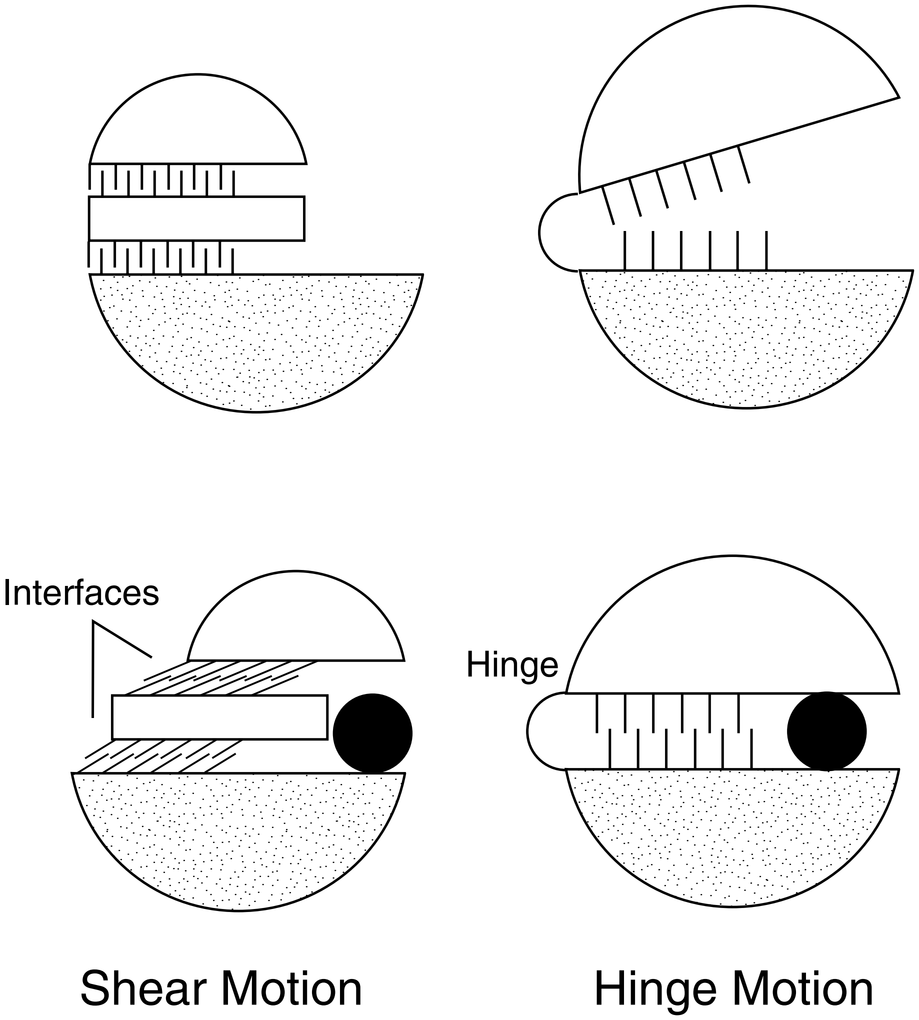
Order movies by other attributes in a custom table.

Motion			PDB ID		Submitter Info	# of residues	maximum CA deviation	# of frames
morph ID	motion ID	name (in DB, as submiited)	#1	#2	date			
86390-22634	trpsyn	TRP repressor	1ttq	1ubs	2001-02-14 16:38:30	268	1.64816	10
86414-22674	trpsyn	TRP repressor	1ttp	1ubs	2001-02-14 16:38:36	268	1.98394	10
86440-22702	trpsyn	TRP repressor	1ttp	1ttq	2001-02-14 16:37:07	268	0.76212	10
15439-7452	tropc	troponin	1ncx	1ncy	2001-02-15 00:38:05	162	0.0891075	10
d1ncx__-d1ahr__	tropc		1ncx	1ahr	1999-11-06 17:37:57	146	23.438	10
d1ncx__-d1ap4__	tropc		1ncx	1ap4	1999-11-06 22:01:40	89	13.3166	10
d1ncx__-d1cfd__	tropc		1ncx	1cfd	1999-11-06 17:50:53	148	11.707	10
d1ncx__-d1osa__	tropc		1ncx	1osa	1999-11-06 18:21:18	148	9.78271	10
d1ncx__-d1tcob_	tropc		1ncx	1tco	1999-11-07 02:05:27	375	41.803	10
d1ncx__-d2sas__	tropc		1ncx	2sas	2000-07-20 14:20:12	185	9.869	10
d1ncx__-d2scpa_	tropc		1ncx	2scp	2000-07-19 03:03:19	174	8.71277	10

Figure 3. Schematic Showing the Overall Classification Scheme for Motions. LEFT, the database is organized around a hierarchical classification scheme, based on size (fragment, domain, subunit) and then packing (hinge or shear). Currently, the hierarchy also contains a third level for whether or not the motion is inferred. RIGHT is a schematic showing the difference between shear (sliding) and hinge motions. It is important to realize that the hinge-shear classification in the database is only "predominate" so that a motion classified as shear can contain a newly formed interface and one classified as hinge can have a preserved interface across which there is motion. (Adapted from Gerstein and Krebs (1998)[7].)

Motion

Number Known Forms	Size of Motion	Mechanism of Motion	Examples	#
2 forms	Fragment	Hinge	TIM, LDH, TGL	14
		Shear	Insulin	3
		Unclassifiable	MS2 Coat	3
	Domain	Hinge	LF, ADK, CM	16
		Shear	CS, TrpR, AAT	8
		Refold	Serpin, RT	3
		Special	Ig elbow	1
		Unclassifiable	TBP, EF-tu	3
	Subunit	Allosteric	PFK, Hb, GP	4
		Non-allosteric	Ig VL-VH	2
		Unclassifiable		
1 form	Fragment	Hinge		
		Shear		
		Unclassifiable	bR	1
	Domain	Refold		
		Hinge	LF~TF,SBP	10
		Shear	HK~PGK,HSP	4
		Special		
		Unclassifiable	Myosin	4
	Subunit	Allosteric		
		Non-allosteric		
		Unclassifiable	PCNA, GroEL	3



Shear Motion

Hinge Motion

Figure 4. Classification Statistics. Approximately one third of protein motions that have been studied are in the database (MIDDLE RIGHT). The Database itself is divided into three categories: automatically found in the PDB, user-submitted, and the extensively studied “gold standard” motions that were manually curated from the scientific literature (TOP RIGHT). The “gold standard” set is further classified into subcategories on the basis of packing (TOP LEFT), size of motion (MIDDLE LEFT), known versus suspected motions based on the number of solved conformations (BOTTOM LEFT), and the experimental techniques used to study each motion (BOTTOM RIGHT). The latter numbers sum to more than 100% because some motions were studied by more than one technique.

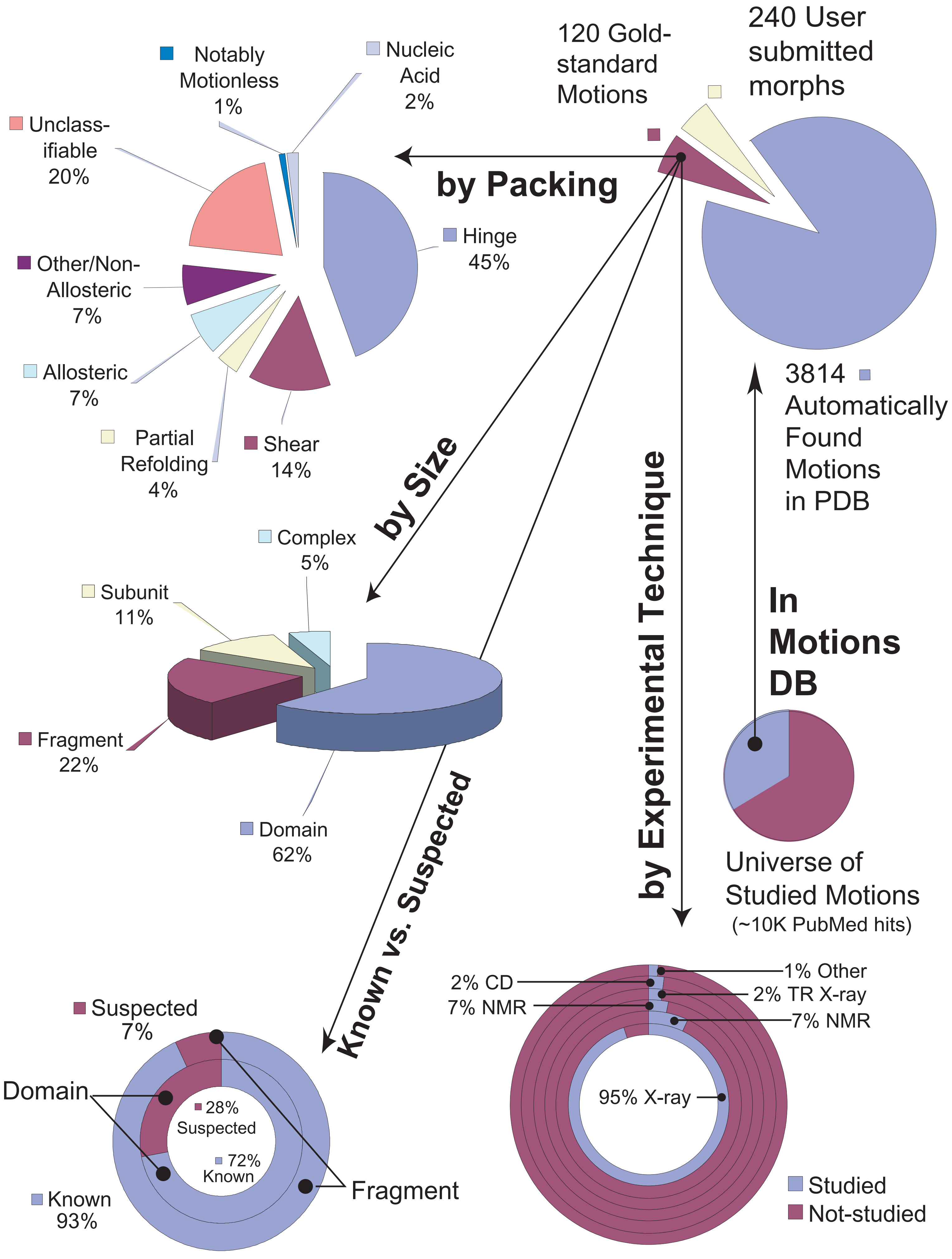

Notably Motionless 1%
Nucleic Acid 2%
Unclass-ifiable 20%
Hinge 45%
Other/Non-Allosteric 7%
Allosteric 7%
Partial Refolding 4%
Shear 14%
120 Gold-standard Motions
240 User submitted morphs
by Packing
3814 Automatically Found Motions in PDB
by Size
Complex 5%
Subunit 11%
Fragment 22%
Domain 62%
by Experimental Technique
In Motions DB
Universe of Studied Motions
(~10K PubMed hits)
Known vs. Suspected
Suspected 7%
Domain
28% Suspected
72% Known
Known 93%
Fragment
2% CD
7% NMR
1% Other
2% TR X-ray
7% NMR
95% X-ray
Studied
Not-studied

Figure 5. This figure shows how the usage of various terms and phrases associated with protein motion have increased every year in the literature. To construct this graph, various searches were done with the NCBI's PubMed database. The graph distinguishes between various quoted and not quoted searches. In total there were 13191 hits in the PubMed database relating to protein motions.

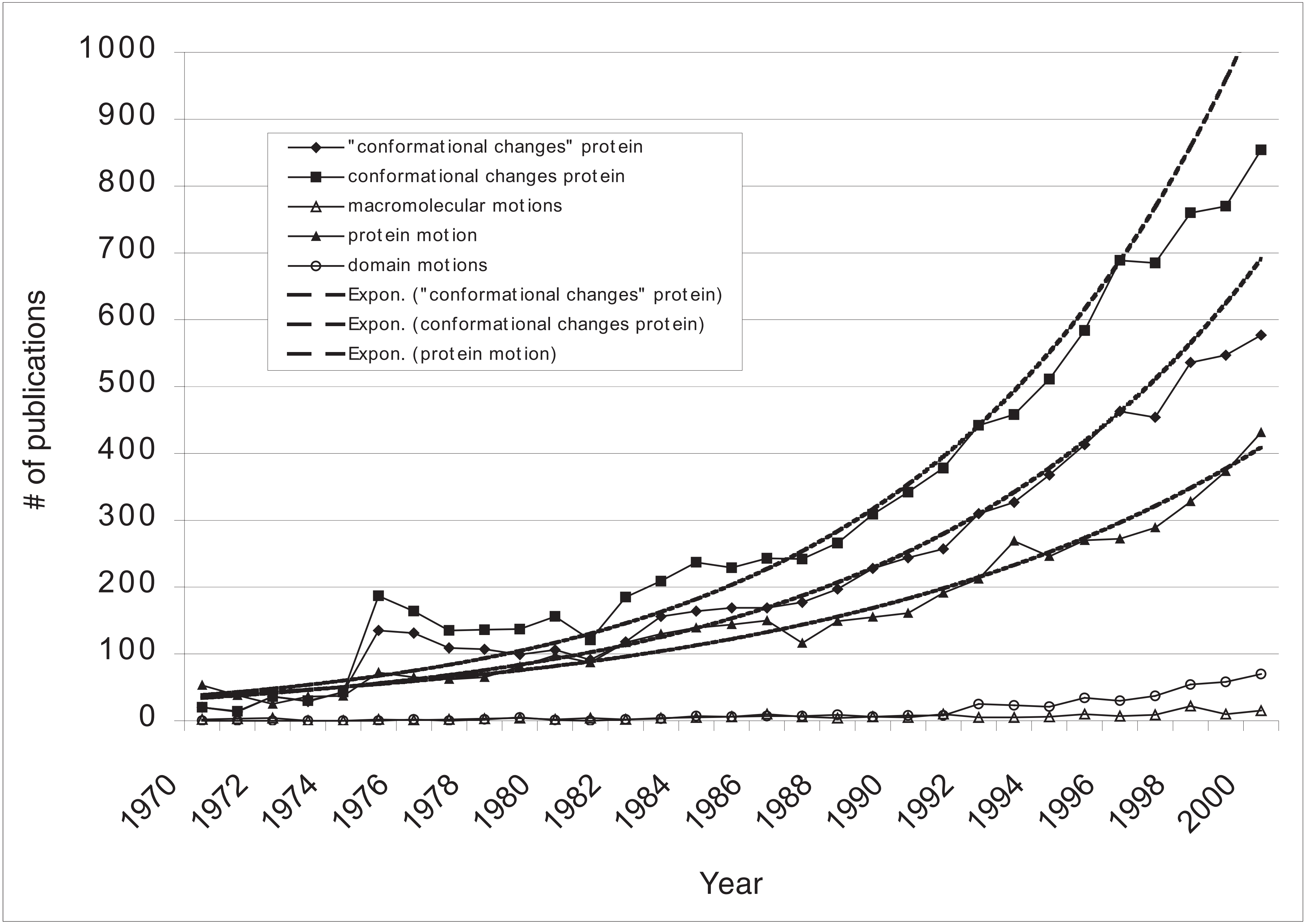

1000
900
800
700
600
500
400
300
200
100
0
# of publications
1970
1972
1974
1976
1978
1980
1982
1984
1986
1988
1990
1992
1994
1996
1998
2000
Year
"conformational changes" protein
conformational changes protein
macromolecular motions
protein motion
domain motions
Expon. ("conformational changes" protein)
Expon. (conformational changes protein)
Expon. (protein motion)

Figure 6. This figure illustrates figures generated by a new set of Web tools associated with normal mode analysis that the user may request on any protein for which a PDB structure file is available. Panel B performs a normal mode flexibility analysis on the structure. Regions that are more flexible are colored in red, while less flexible regions are colored in blue. Panel A gives similar information, using experimental b-factors supplied in the PDB file, if available. Panel C, shows the parts of the protein that actually move, as calculated from comparison of the starting and ending PDB structures for the motion. Areas that move are colored in red, while areas that remain stationary are colored in blue. The user may compare these three panels to deduce structural information. For example, hinge locations involved in the motion may be deduced, as these are highly flexible regions (as identified by panels A and B) located near the moving domains (show in red in panel C).

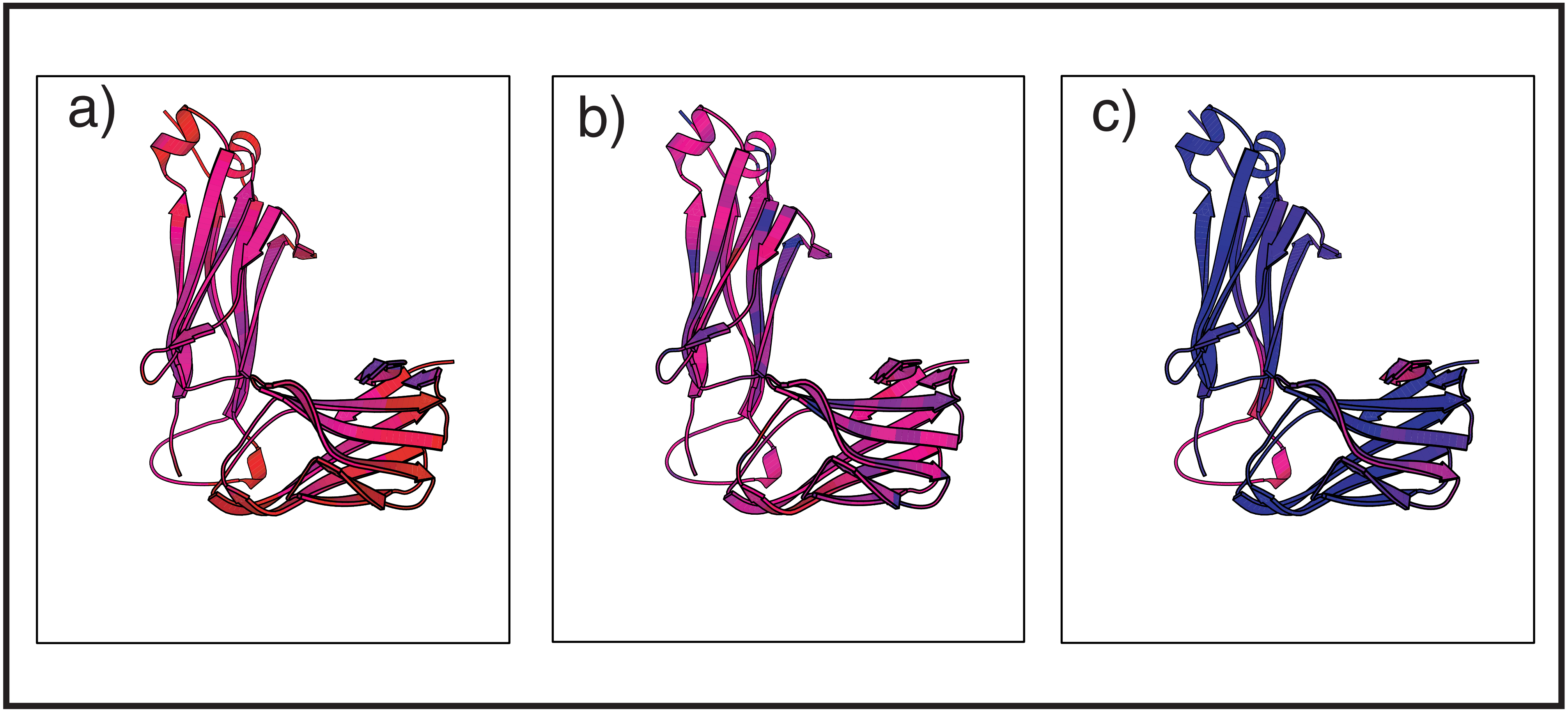

a)
b)
c)

Figure 7. Normal Modes. TOP LEFT: This figure briefly summarizes basic normal mode concepts. (It was inspired from Peter Steinbach's web illustration, http://cmm.info.nih.gov/intro_simulation/node26.html.) TOP RIGHT: This is a cartoon of one of the low-frequency normal modes of bacteriorhodopsin. This particular normal mode is approximately perpendicular to the cell membrane. BOTTOM: This panel illustrates the concept of the normalized dot-product which is sometimes useful in statistical calculations on normal mode vectors and their relationship to experimental displacement vectors.

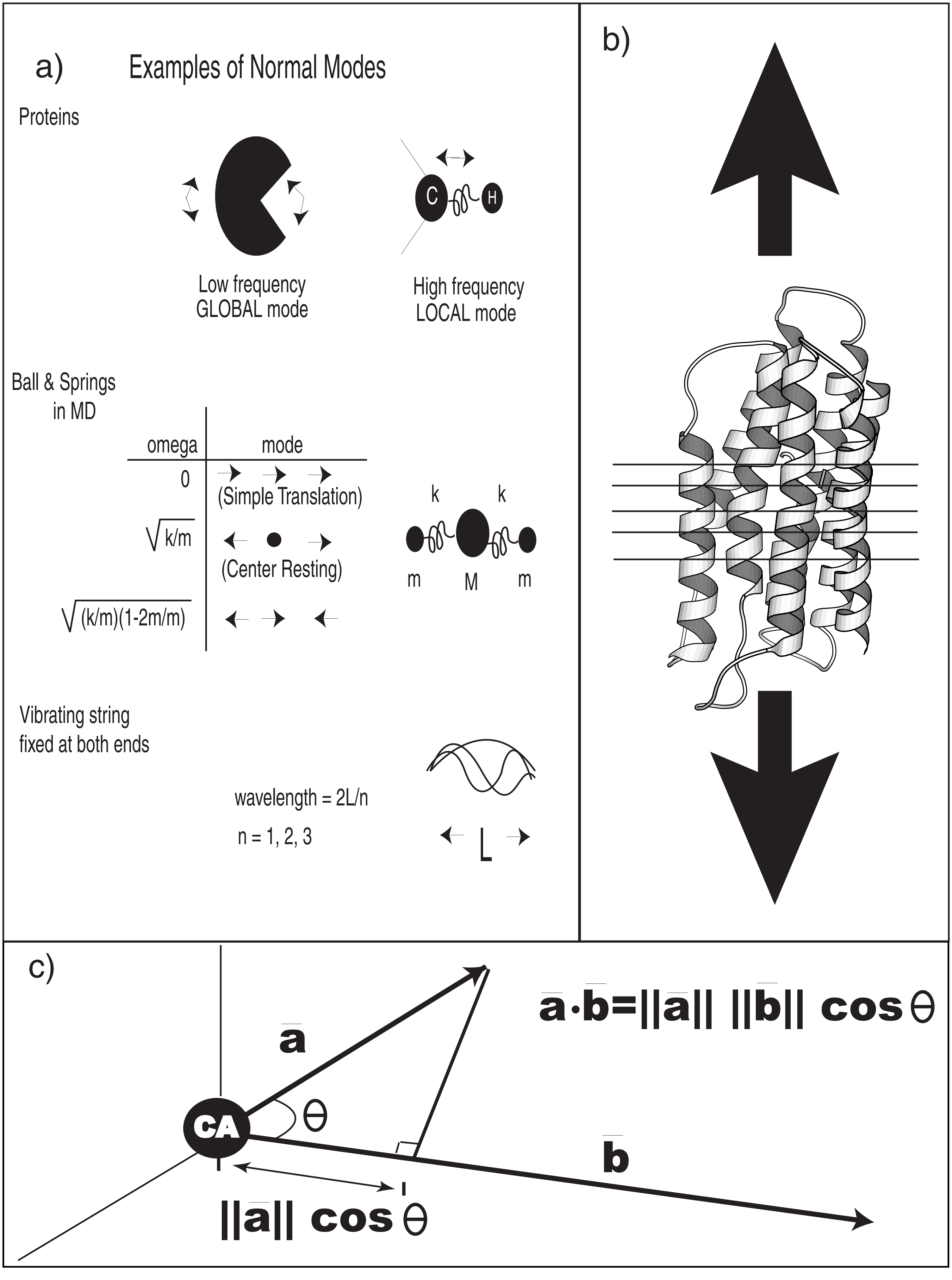

a)
Examples of Normal Modes
Proteins
Low frequency
GLOBAL mode
C
H
High frequency
LOCAL mode
Ball & Springs
in MD
omega
mode
0
(Simple Translation)
$\sqrt{k/m}$
(Center Resting)
$\sqrt{(k/m)(1-2m/m)}$
k
k
m
M
m
Vibrating string
fixed at both ends
wavelength = 2L/n
n = 1, 2, 3
L
b)
c)
$\bar{a}$
$\bar{a} \cdot \bar{b} = ||\bar{a}||\ ||\bar{b}|| \cos\theta$
θ
CA
$\bar{b}$
$||\bar{a}|| \cos\theta$

Figure 8. Voroni Polyhedra. LEFT: Two representative Voronoi polyhedra from 1CSE (subtilisin). On the left is shown the polyhedron around the sidechain hydroxyl oxygen (OG) of a serine. On right is shown the six polyhedra around the atoms in a Phe ring. RIGHT: The Voronoi Polyhedra Construction. A schematic showing the construction of a Voronoi polyhedron in 2-dimensions. The asymmetry parameter is defined as the ratio of the distances between the central atom and the farthest and nearest vertex.

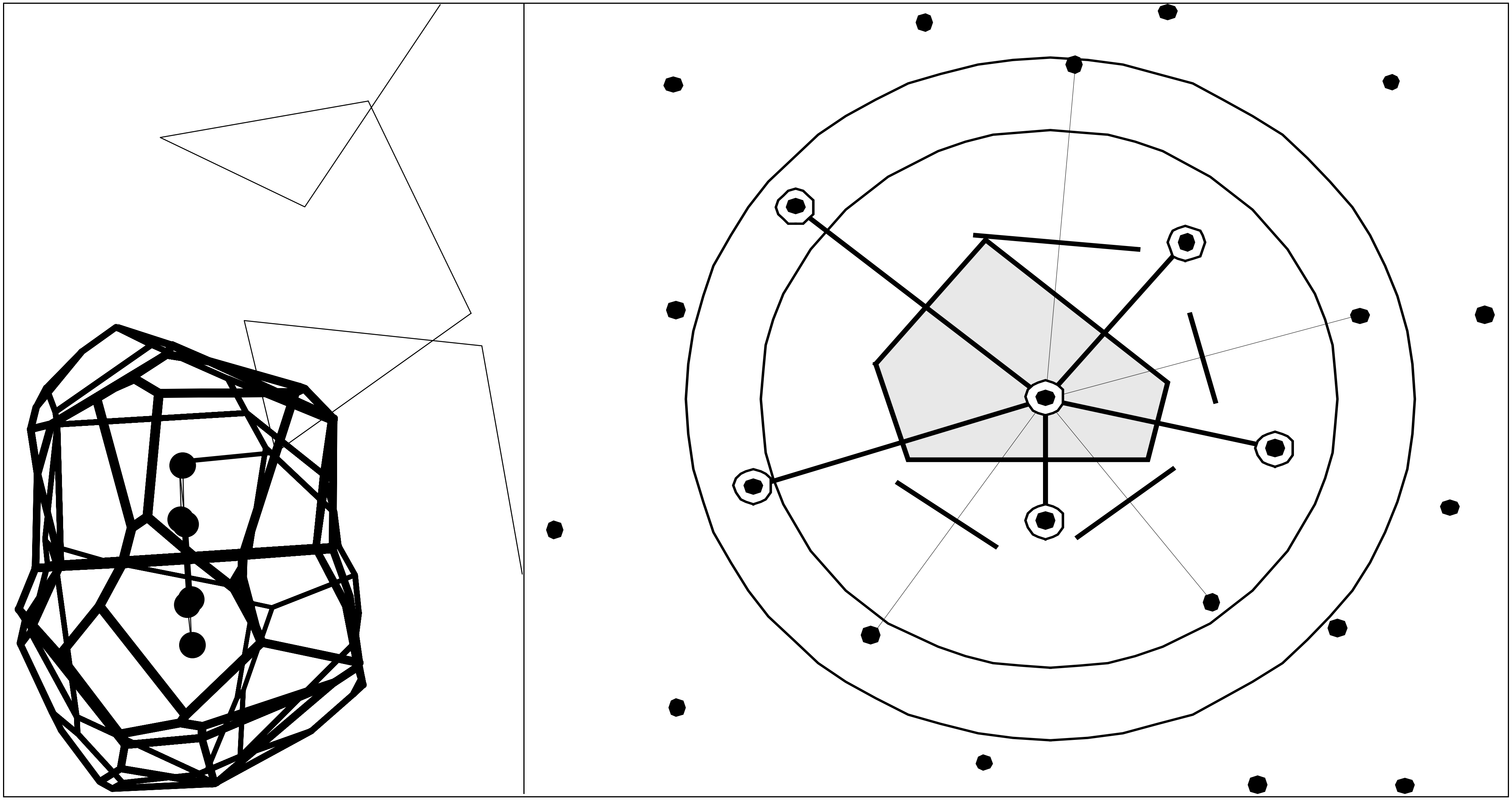


Figure 9. Comparison of the average amino acid composition in linker sequences and proteins in general (as represented by the PDB40 database).

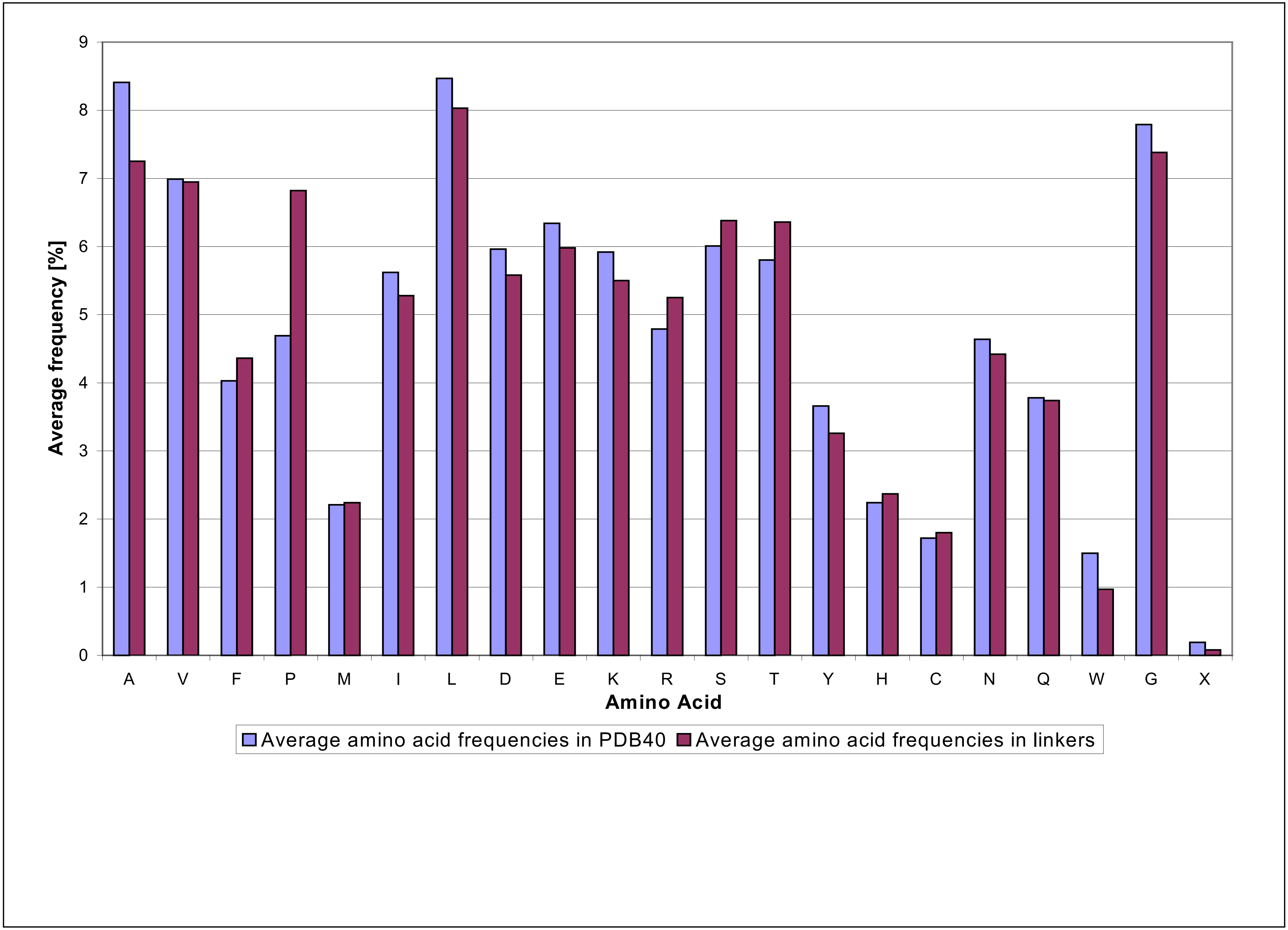
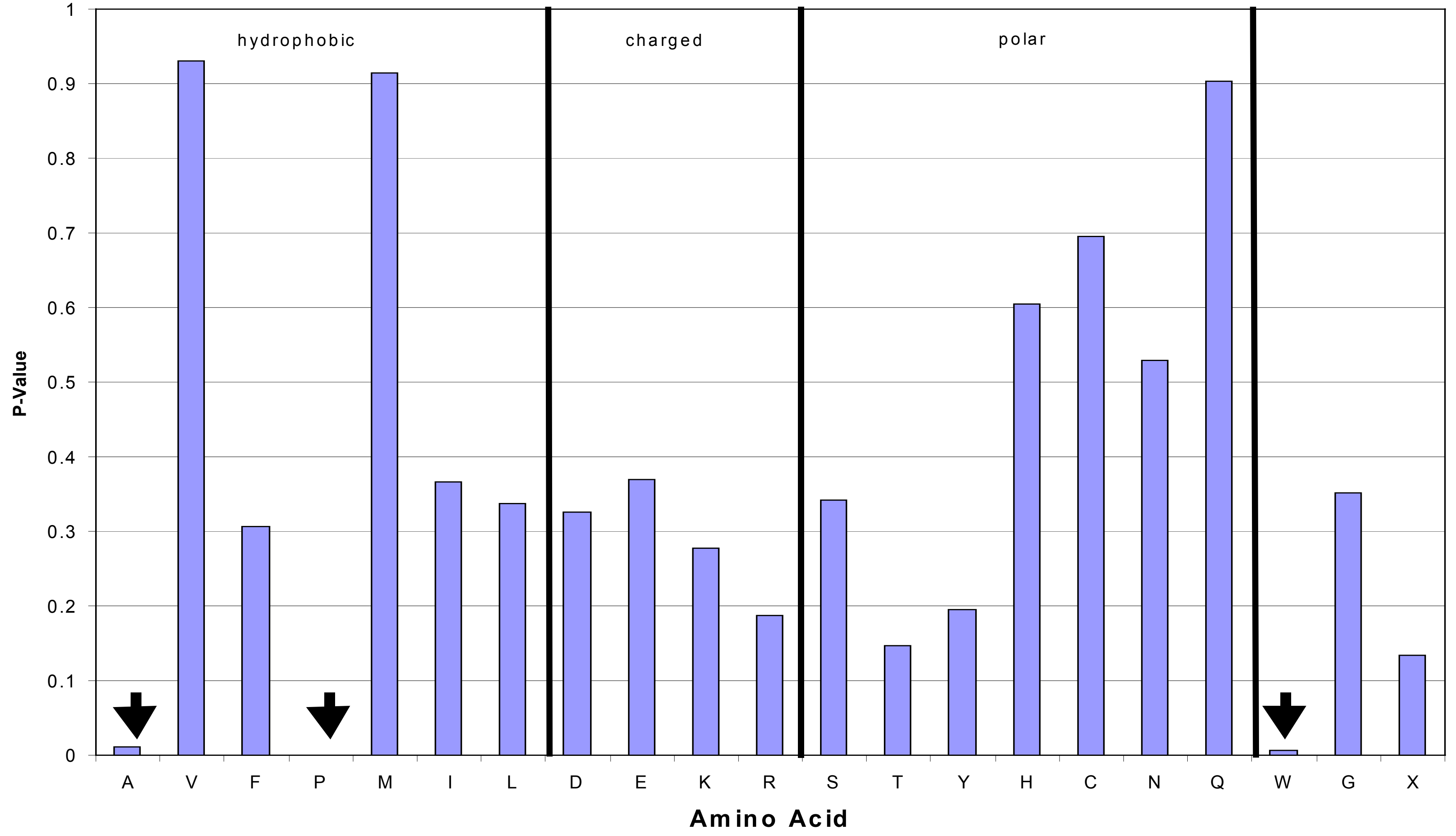

9
8
7
6
5
4
3
2
1
0
Average frequency [%]
A
V
F
P
M
I
L
D
E
K
R
S
T
Y
H
C
N
Q
W
G
X
Amino Acid
Average amino acid frequencies in PDB40
Average amino acid frequencies in linkers

Figure 10: P-values for the average amino acid compositons in linker sequences. The P-values of alanine, proline, and tryptophan are close to zero. The difference between the content of these amino acids in linkers and protein sequences in general (as represented by the PDB40 database) is statistically significant at better than 98% confidence.


hydrophobic
charged
polar
P-Value
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
A
V
F
P
M
I
L
D
E
K
R
S
T
Y
H
C
N
Q
W
G
X
Amino Acid