Nucleic Acids Research, 2000, Vol. 28, No. 8 **1665–1675**

SURVEY AND SUMMARY

The morph server: a standardized system for analyzing and visualizing macromolecular motions in a database framework

Werner G. Krebs and Mark Gerstein*

Department of Molecular Biophysics and Biochemistry, Yale University, PO Box 208114, New Haven, CT 06520, USA

Received September 30, 1999; Revised and Accepted February 2, 2000

ABSTRACT

The number of solved structures of macromolecules that have the same fold and thus exhibit some degree of conformational variability is rapidly increasing. It is consequently advantageous to develop a standardized terminology for describing this variability and automated systems for processing protein structures in different conformations. We have developed such a system as a 'front-end' server to our database of macromolecular motions. Our system attempts to describe a protein motion as a rigid-body rotation of a small 'core' relative to a larger one, using a set of hinges. The motion is placed in a standardized coordinate system so that all statistics between any two motions are directly comparable. We find that while this model can accommodate most protein motions, it cannot accommodate all; the degree to which a motion can be accommodated provides an aid in classifying it. Furthermore, we perform an adiabatic mapping (a restrained interpolation) between every two conformations. This gives some indication of the extent of the energetic barriers that need to be surmounted in the motion, and as a by-product results in a 'morph movie'. We make these movies available over the Web to aid in visualization. Many instances of conformational variability occur between proteins with somewhat different sequences. We can accommodate these differences in a rough fashion, generating an 'evolutionary morph'. Users have already submitted hundreds of examples of protein motions to our server, producing a comprehensive set of statistics. So far the statistics show that the median submitted motion has a rotation of ~10° and a maximum Cα displacement of 17 Å. Almost all involve at least one large torsion angle change of >140°. The server is accessible at http://bioinfo.mbb.yale.edu/MolMovDB

INTRODUCTION

Solved structures and related structural information on proteins is growing at an exponential rate. This is due chiefly to continuous technological progress in X-ray crystallography, NMR spectroscopy and computer technology. As researchers solve structures at an ever increasing rate, there occurs an obvious need for processing techniques to relate such structures to one another, beyond classification or structural alignment. Protein motions, as an essential link between structure and function, are an obvious area of relationships between protein structures in the databases. Motion is intimately related to the way a structure fulfills a particular function. Protein motions are involved in a wide variety of basic functions, including regulation, transport of metabolites, formation of large assemblies, and cellular locomotion. Examples can be found throughout nature, from local conformational changes involved in the binding of ligands (1) that occur in enzymatic reactions to the complex rearrangement of covalent bonds (2).

Obviously, one of the best ways to represent and communicate protein motions is through 'movies', especially when they are made available over the web. There have been a number of previous efforts in this area. Vonrhein *et al.* (3,4) made a custom movie of calmodulin and placed it on the Web. Similar work has been done by Sawaya *et al.*, who created movies of crystal structures of polymerase beta (5). Ray-traced 3D molecular dynamics simulation of acetylcholinesterase from mutagenesis data have also been made available (6–8). More recently, movies from molecular dynamics simulations of protein folding (plp group) (9,10) have been available on the Internet. Xu *et al.* used the techniques of normal mode analysis to produce a morph movie of GroEL from structural data (11–14).

In this paper we present a perspective on how protein motions can be put into standardized, consistent terms. We develop a simple model for protein motions involving rigid-body motion of parts, apply our model to actual cases and measure how well it fits. Our approach is embodied in an integrated Web server that provides tools to compare solved conformations of proteins involved in motion, generates statistics to characterize and classify them into a database and automatically makes a morph movie to represent them. In addition, the server presents a database linking protein motions with custom movies of

*To whom correspondence should be addressed. Tel: +1 203 432 6105; Fax: +1 360 838 7861; Email: mark.gerstein@yale.edu

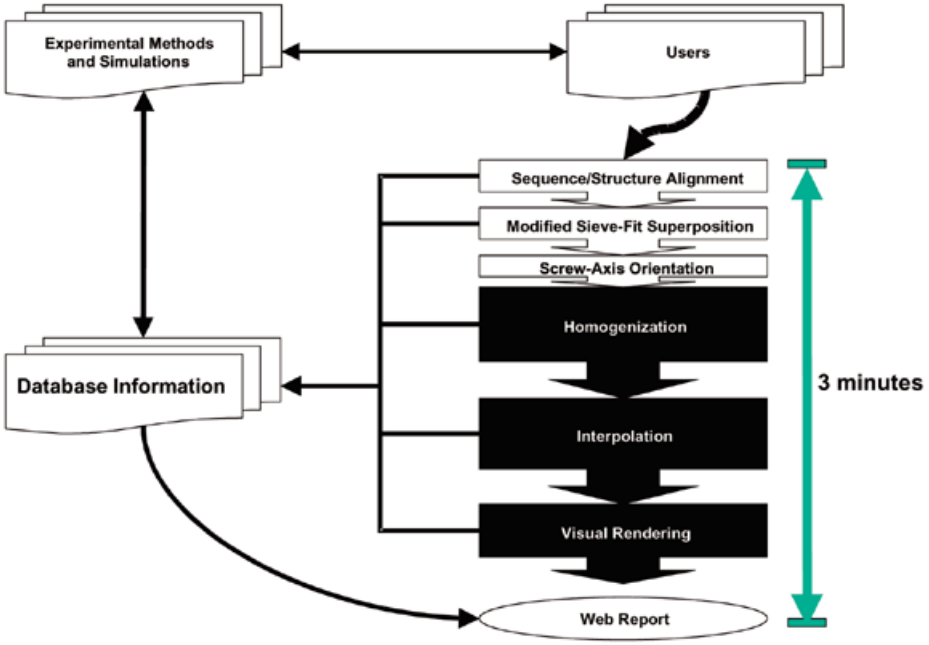

Figure 1. Diagram of our approach. The information flow from databases, through the server, and then back again to databases is broken down into its component steps. Experimental data in the PDB and other databases is converted into a motion entry in the Database of Macromolecular Motions, from whence a morph movie is generated and statistics are collected. These results are subsequently stored in the Database of Macromolecular Motions. The interaction of the server with the peripheral parts in the figure ('Database Information', 'Experimental Methods and Simulations' and 'Users') is largely under users' control, although we are developing automated tools to generate comparisons automatically from databases such as SCOP. The results of a comparison are both returned to the user and referenced in the Database of Macromolecular Motions, hence the arrow back to 'Database Information'. The web report extracts information both from server results and from pre-existing information in the Database of Macromolecular Motions, if any, hence the arrow from 'Database Information' to 'Web Report'.

motions available at other sites, along with our own morphs generated automatically by the server upon request by members of the Internet community. Our server and database have been used by Internet users to analyze a number of recent structures including human interleukin 5 (15), bc1 complex (16,17), glycerol kinase (18,19) and lactoferrin (20,21). It has also been used as a source of raw data in visualization tools (22) and in relation to other biological databases (23). The Web server is accessible at: http://bioinfo.mbb.yale.edu/MolMovDB/morph . It is integrated with the Database of Macromolecular Motions (24,25) and is also connected with a variety of tools for aligning protein folds and studying their occurrence in genomes (26–29).

INFORMATION FLOW

The best way to understand our approach is in terms of the 'information flow' diagrammed in Figure 1. One starts by submitting two or more conformations of a given protein to the server. Given one conformation, a number of online tools and databases, such as PDB, FSSP, SCOP, CATH, CE and VAST can suggest a second conformation. Then, through a variety of transformations, the server classifies the motion in the database and produces an appealing movie.

Data sources

Solved conformations analysis as performed by the server's tools requires two kinds of information: (i) 3D atomic coordinates of protein conformations as solved structure files (such as those at the PDB) and, more importantly, (ii) information relating two or more of these solved structures, thus selecting them for analysis. Such information, for instance, could come from the SCOP Database (30,31), from automated searching of databases for proteins related by structure or sequence or from a simple user input form on the Web. A selection scheme is important because the number of ordered pairs of PDB structures is rather large (more than $10\,000^2$). Figure 2 diagrams the server in the larger context of data sources.

Alignment

Once a string of structures has been given to the server, the first step is to establish equivalence (an alignment) between residues in the various proteins. This is necessary because the protein structures compared, while sharing some evolutionary or structural similarity, will, in general, not share the same amino acid sequence. Consequently, an alignment is necessary.

Because the server may be asked to simultaneously compare more than two sequences, an algorithm capable of simultaneously aligning multiple sequences (or structures) and potentially building an evolutionary tree must be used. For this purpose, we have chosen the AMPS algorithm (32–34). In cases in which sequence alignment is inappropriate, such as for highly diverged homologs, we use the technique of structure alignment (28,29). The latter method relies primarily on the use of 3D coordinates (i.e., solved PDB structures of proteins) to produce a sequence alignment otherwise analogous to an alignment produced purely from sequence information. As a result, the structural method is able to generate meaningful sequence alignments from both highly related proteins and completely unrelated proteins sharing similar structural features due to

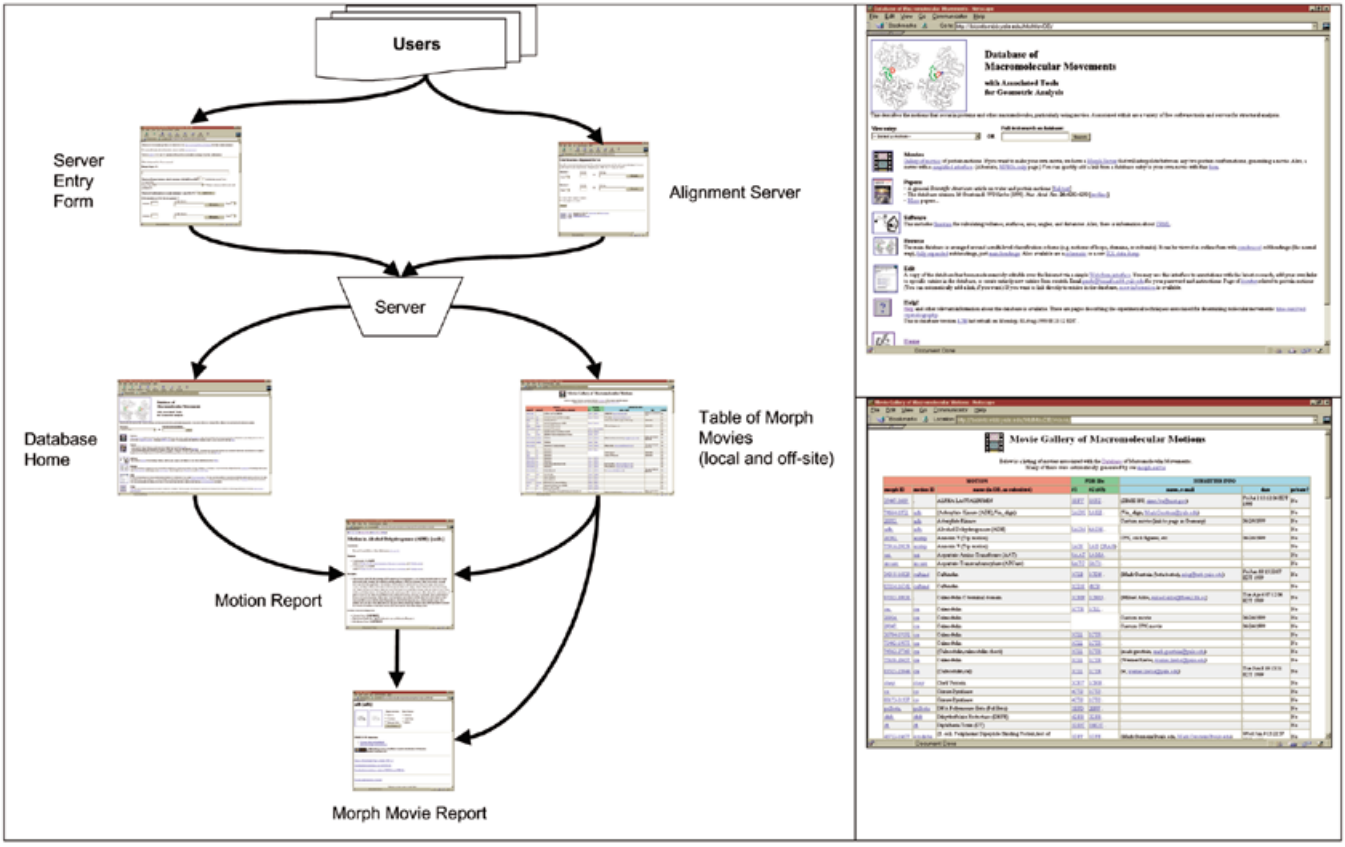

Figure 2. (**Left**) Here, the information flow may be visualized as a series of linked Web pages. Users submit new motions to the server via either the Server Submis sion Form or via a simplified interface through the Structural Alignment Server's submission form. The query is processed by the server. If the morph operation is successful, the new morph is added to the Table of Morph Movies (which links off-site URLs as well). This table has links to both the morph's report form (from which the morph may be viewed) and also the associated motion entry in the Database of Macromolecular Motions if the motion has one. An entry is also created in the motion's entry in the Database of Macromolecular Motions, linking the motion's report to the report for the morph movie. (**Top right**). This is a blow up of the main page of the database from the left side of Figure 2. The entry page of the Database of Macromolecular Motions, http://bioinfo.mbb.yale.edu/MolMovDB is shown above. Users may jump from this to entries on specific motions, many of which link morph movies, or to a table of morphs. (**Bottom right**). This is a blow up of Online Table of Morphs from the left side of Figure 2. Screen shot of the on-line table of morphs Web page at http://bioinfo.mbb.yale.edu/MolMovDB/morphs . In addition to linking to the Web report page for the morph, each entry links to the corresponding database motion entry (if applicable) and provides information on the PDB IDs used to generate the morph movie, along with the information on the submitting user. This table also references off-site morph URLs, and thus functions as a comprehensive database of protein morphs available on the Internet.

convergent evolution. Sequence alignment is used unless sequence similarity is below a user-defined cut-off, at which point structure alignment is used. The choice of approach (sequence or structural alignment) may also be forced by the user upon morph submission.

Superposition

One of the major aims of the server is to collect standardized statistics on the proteins involved in motions. Standardized statistics, such as maximum rotation or maximum Cα displacement, are computed with respect to a specific superposition and reference frame, and so the superposition algorithm is central to any conformational analysis tool.

The output of the alignment procedure establishes residue equivalencies that are used in an intelligent superposition of the structures onto one another. Traditional 'all-atom' RMS superposition minimizes the RMS difference between Cα atoms in the open and closed conformations. In a simple hinge motion, e.g., calmodulin, such an alignment fits the closed conformation symmetrically inside the open conformation (Fig. 3). Amongst other things, the maximum Cα displacement computed from such a superposition is considerably underestimated from the common sense alignment, and the morph movie gives the impression of motion far more complicated than a simple opening of a hinge. Instead, we perform the superposition with a modified 'sieve-fit' procedure (35,36). The procedure is iterative. On each iteration the remaining Cα atoms are superimposed by a standard RMS fit, and then the pair of corresponding Cα atoms furthest apart are eliminated. This is repeated until approximately half of the atoms in the protein have been eliminated. Previously described uses of the 'sieve-fit' procedure (36,37) used some sort of cut-off value to determine when to stop the procedure, typically RMS deviation. No single RMS deviation cut-off value has consistently worked well. However, we have found that by stopping the procedure after approximately half the atoms have been discarded, one of the 'domains' thus selected generally corresponds approximately to a superset or a subset of a real domain in the structure, and is thus well suited for performing the subsequent axes transformations.

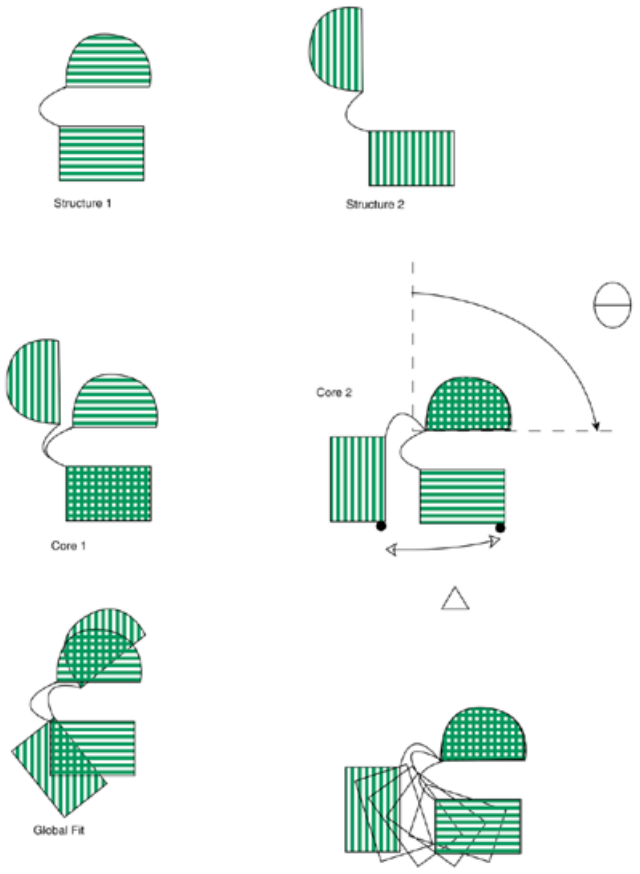

Figure 3. Superposition of a calmodulin-like protein undergoing a hinge motion. Structures 1 and 2 indicate the closed and open conformations, respectively. Compare 'Global Fit', the superposition produced by a traditional least-squares fit of the structures, to 'Core1' and 'Core 2', the two possible superpositions produced by sieve-fitting. The final panel depicts how a morph movie might appear using the 'Core 2' superposition.

Orientation and hinge location

To locate the screw-axis, a 'fit–refit' procedure, as described by Lesk and Chothia (38) is used. Following superposition of the starting and ending conformations, we only consider the set of eliminated atoms. We perform an RMS-fit of that set between the starting and ending conformations; the server performs the new superposition (arbitrarily) on the ending conformation. A comparison of the new position of the ending conformation following this latest fit with its position following the 'sieve-fit' procedure yields a geometric transformation whose screw axis is (approximately) the axis of the hinge motion, i.e., the location of the hinge, as has been published elsewhere (39). Straightforward calculations allow characterization of the angle of rotation around the hinge axis.

If a significant hinge motion is present, the software uses 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. The longest moment of the protein (long axis) is rotated (optionally) so that it is parallel to the Y-axis. Finally, the coordinate frame is translated so that the centroid of the initial conformation is in the center of the field of view.

The software also attempts to locate putative hinge regions using a simple and relatively fast algorithm. The algorithm looks for a persistent transition between the two domains identified by the program. The algorithm constructs a search window, initially with 24 residues. It examines each position along the peptide backbone in this window. If there is a persistent transition (i.e., one-half of the algorithm's search window belonging to one 'domain' and the other half to the other), a hinge is detected. If the program fails to find any hinges along the backbone chain, the window size is reduced by two, and the procedure is repeated until the window size has been shrunk down to 12 residues, at which point the program reports failure. Empirically, this crude but computationally inexpensive algorithm successfully finds many hinge regions, such as the hinge region for calmodulin, which agree well with published residue selections. In other cases, the algorithm comes close, identifying a residue selection that borders on a hinge. Hinges may be displayed graphically via a 'hinge movie' identifying the putative hinge region or regions in red.

In related work, Wriggers *et al.* presented techniques to identify protein domains and common hinges using an adaptive least-squares fitting technique (40); the user is presented with a number of options (spatial connectivity maintenance, significant structural difference filters) to ensure optimal hinge finding. For the remote user's convenience, our own hinge finder is at present fully automatic and presents no options to the user. It may be advantageous for us to provide such options in the future so that the user can override and improve on the putative hinge initially selected by our algorithm, although this would partially defeat our efforts at standardization. Maiorov *et al.* (41) has developed a system which detects hinges by large-scale sampling of torsion angle space; this technique, while presumably more accurate, is also much more computationally expensive then our current technique. It may be useful for us to give the user the option of using alternate hinge finding engines in the future.

To illustrate the putative hinge finder, a frame from one such 'hinge movie' is given in Figure 4, with the putative hinge identified in black. Superposition, orientation and hinge-finding are relatively fast steps, requiring a fraction of a second of computer time on our server.

Homogenization

We have modified the X-PLOR package (42) to homogenize the stored coordinates. This problem is non-trivial (43,44). The initial, solved intermediate and final conformations are parsed by X-PLOR and examined for missing non-hydrogen coordinates. These are filled in using energy minimization with the known coordinates of the molecule fixed at their solved positions. If these missing coordinates are available in another solved conformation, the coordinates from the superimposed and rotated conformation are used as an initial guess as to their likely positions. As written, filling-in of missing non-hydrogen coordinates is necessary for the energy minimization subsystems to work robustly with a large number of PDB files. It also ensures homogenized output of PDB files, which is required by the visual rendering subsystem.

Interpolation

The next step is in the dominion of what we refer to as the 'interpolation engine.' Once the structures have been homogenized

Figure 4. Putative hinge movie. A frame from a 'hinge movie' of ras protein (PDB ID 4Q21–6Q21 morph intermediate frame) showing the putative hinge regions as identified by the server. The server identifies 71:82 and 118:129 as putative hinge regions in the motion, here shown in black.

in terms of solved atomic coordinates, interpolation may proceed. Under command of the script, the custom X-PLOR interpolation function is repeatedly called, each time evenly reducing the distance between the current structure and the final structure. When more than two solved conformations are present, the distance between the current structure and a solved intermediate conformation is evenly reduced instead. Each step is followed by a round of energy minimization to correct molecular stereochemistry and enforce rules of chemical reality on the structure. To ensure that the final frames are as accurate as possible, the solved endpoint structures are used for these. When solved intermediates are present, these are inserted as frames at regular intervals. The entire process takes only a few minutes to produce 10 frames running on a 500 MHz Intel Pentium III workstation running Linux.

There are many possible interpolation strategies, and a number of tradeoffs between accuracy, various computational resources, time and others are involved in the choice. For this reason, in addition to our original adiabatic mapping engine, we offer the user two engines based on LSQMAN (45,46) (one Cartesian-based and another based on internal phi, psi coordinates), which are faster but appear to be less realistic. Users wishing to add their own, non-trivial interpolation engines may contact the authors to make arrangements to do so. For example, a user wishing to analyze a very large number of trajectories (10 000 or more from, e.g., samplings from molecular dynamics simulations) might wish to supply a simplified interpolation engine and make other arrangements to allow the computations to be completed in a reasonable amount of time.

We chose our original technique, known in the literature as adiabatic mapping (47) for reasons of computational efficiency. It is a technique that produces chemically reasonable morphs with a modest amount of computational power and thus is most suitable for a Web-based server. This remains the default interpolation engine for the server. Using this engine, the server can produce a realistic interpolation of a protein and have the results rendered and returned to the user in < 3 min on a fast Pentium III machine. Using adiabatic mapping, we have also produced our own morph of the motion in GroEL which, although probably less accurate than the considerably more expensive technique of normal mode analysis (11–14), is probably good enough for most researchers seeking only a visual representation. Nevertheless, we believe that, for many proteins, most real motions will occur along the interpolated trajectories, and the morph server may be used to predict intermediate conformations should they exist. How close our predicted pathways come to reality is perhaps best answered through the emerging technique of time-resolved X-ray crystallography (48,49). Thus, an adiabatic mapping engine is much more suited to our goal of automatically interpolating a large percentage of the motions in our database.

Visual rendering

With the intermediate conformations morphed, the molecule is now visually rendered. We have written a Perl script that produces VRML 2.0 (Moving 3D Worlds) code (50,51) on-the-fly from the intermediate PDB files. The VRML 2.0 output is suitable for interactively viewing the moving 3D macromolecule in a VRML 2.0 Internet browser, such as SGI CosmoPlayer 2.0. The advantage of the 3D display format is that the remote Internet user may easily choose a preferred orientation and vantage point.

The molecule is also rendered as a 2D movie in the MultiGif, Quicktime, and MPEG formats, as well as an Adobe Portable Document Format (PDF) (52) page showing the individual frames. Remote adjustment of vantage point and orientation is not possible in the simpler 2D video format, so the molecule is rendered with the screw axis perpendicular to the plane of the display device, as was computed during the orientation process. The molecule is rendered in three display types (53,54): ribbons (with secondary structure indicated), lines (as a simple alpha chain), and ball and stick (showing all individual non-hydrogen atoms). The first two formats are also rendered into a small moving MultiGif icon to afford the database user with a quickly downloaded preview of the larger movies available.

Statistics

In the process, key standardized statistics are recorded. These include maximum Cα displacement, rotation angle in degrees around putative hinge regions, sequences of the putative hinge regions, average torsion angle change in the hinge region versus the overall average, distance of the putative hinge region from the screw axis, distance of the screw axis from the centroid, a structural comparison score between the two domains and a number of additional, useful statistics, such as the differences in torsion angles at every aligned position and the pseudo CHARMM/X-PLOR (42,55) energy at each point in the morph.

These statistics are detailed enough to perform an automatic preliminary classification of the motion and determine the location of the hinge relative to the transformed axes. (For example, a large rotation angle indicates a probable hinge motion.) A detailed description of our statistical results is given in Table 1 for five motions. Ranges and averages of some of these statistics after several hundred alignments are given in Table 2 along with similar but sparse statistics culled manually from the scientific literature for comparison.

For example, over approximately 175 motions submitted for analysis, the median motion has a maximum rotation of 9.5°

Table 1. Comprehensive statistics for alcohol dehydrogenase, reoverin, DNA polymerase beta, GroEL and diphtheria toxin as reported by the server

	Statistic	[Code]	Easy	Typical		Large	Impossible
			ADH	Reco-verin	DNA Pol-Beta	GroEL	Diphtheria Toxin
Input Structures	Motion ID	[ID]	adh	recvin	polbeta	groel	dt
	1st input frame	[inputframe0]	8ADH	1IKU	1BPD	1GRL	1DDT
	2nd input frame	[inputframe1]	6ADH	1JSA	2BPF	1AON	1MDT
	Size (Å) (in terms of window for rendering)	[max_x_or_y]	36	41	52	55	39
	Number of atoms	[natoms]	2887	1639	2697	3993	4110
	Number of residues	[nresidues]	374	201	335	548	535
Overall Motion	Overall RMS between first and last frames	[RMSoverall]	2.0	13	8.6	16	20
	Rotation (degrees)	[kappa]	4.9°	73°	9.9°	62°	62°
	Overall translation of centroid (Å)	[translation]	2.1	13	6.1	47	66
	X translation (Å)	[TransX]	1.1	-0.24	0.94	45	-45
	Y "" .	[TransY]	-0.95	-9.14	4.1	-2.1	-0.54
	Z "" .	[TransZ]	1.5	-9.78	-4.4	-10	48
1st Core	Number Cα's in 1st core	[AlignedCoreCAs]	187	95	160	259	262
	RMS of 1st core (Å)	[AlignedCoreRMS]	0.40	3.0	0.92	1.4	0.37
	Max Cα displacement in 1st Core (Å)	[MaxCore Deviation]	0.66	7.6	1.7	4.2	0.60
2nd Core	Num. Cα's in 2nd core	[2ndCoreCAs]	190	94	160	260	260
	RMS of 2nd core (Å)	[2ndCoreRMS]	2.9	18	12	23	29
	Max Cα displacement in 2nd core (Å)	[Max2ndCore Deviation]	7.1	38	28	49	60
	RMS of 2nd core (Å) after fitting on 1st core	[2ndCoreRMS postrefitting]	1.6	11	11	10	18
Hinge	Number of putative hinges detected	[NHinges]	0	0	0	1	1
	X position of 1st hinge (Å) rel. to centroid	[Hinge000X]	-	-	-	-4.7	-7.2
	Y position ""	[Hinge000Y]	-	-	-	11	-0.91
	Z position ""	[Hinge000Z]	-	-	-	3.3	-3.0
	1st Hinge Residue Selection	[Hinge000res]	-	-	-	380:403	352:375
	Sequence of 1st putative hinge	[Hinge000seq]	-	-	-	EVEMKE KKARVE DALHAT RAAVEE	NLFQVV HNSYNR PAYSPG HKTQP
Screw Axis	Distance betw. screw-axis (x0) & centroid (Å)	[x0ToCentroid Distance]	21	8.4	23	30	39
	X displacement centroid from screw axis (Å)	[x0X]	-0.16	-0.5	-2.5	17	-20
	Y ""	[x0Y]	-5.0	-6.2	-5.2	-16	-24
	Z ""	[x0Z]	-20	5.7	-22	19	-24
	Distance between screw axis and 1st hinge (Å)	[Hinge000x0dist]	-	-	-	26	45
Torsion Angles	Max phi change (Max of Abs. degrees, 0°-180°)	[MaxPhi]	180°	180°	180°	180°	180°
	Max psi change	[MaxPsi]	180°	180°	180°	180°	170°
	Max alpha change	[MaxAlpha]	150°	180°	180°	180°	170°

These statistics were automatically generated by the server in the course of morphing alcohol dehydrogenase, recoverin, DNA polymerase beta, the first chain of GroEL and diphtheria toxin. They are reported here to two significant figures except where exact. A brief explanation for each statistic may be found above. More comprehensive explanations may be found online.

Table 2. Comparison of statistics between automatically gathered (server gathered) and manually gathered statistics for maximum Cα displacement and maximum rotation

	Hand-gathered statistics			Automatically collected motion statistics				
Value	Min.	Max.	Mean	Min.	Max.	Mean	Median	SD
Maximum Cα displacement (Å)	1.5	60	12	0.90	81	23	17	19
Maximum hinge rotation (°)	5	148	24	0.0	150	35	9.5	46

Despite the sparseness of the manually culled data, the statistics are roughly comparable. Maximum Cα displacement was calculated by first sieve-fitting the protein conformations. The 81 Å motion in the database is due to Oxo-Acid-Lyase (5CTS to 1AJ8 in the PDB). The 12 references reporting maximum rotation in the literature reported a mean maximum rotation of 24°, whereas the server found a mean maximum rotation of 35° over the 176 entries present at the time the table was generated. The mean, however, is skewed by some of the larger motions; the median displacement is much smaller. The maximum value of 150° is due to Oxidoreductase (1FMC→1HDC in the PDB). To collect the manual data, we found 11 entries in the Database of Macromolecular Motions citing manually gathered Cα displacement statistics from the literature, and 12 entries giving manually gathered maximum hinge rotations. (Some researchers reported only Cα displacement while others reported only maximum hinge rotation, so these correspond to different sets of proteins.) Automatic collection used a sample of 184 motions for Cα displacement and 176 motions for maximum hinge rotation.

Table 3. Morph similarity score statistics

Statistic on 65 observations	Mean	Medium	Maximum
Number of residues aligned	250	5	780
Trimmed RMS	2.4	0.24	16
Trimmed RMS *P*-value	0.041	0.0	0.96
Sequence percent identity	55	7.9	100
Sequence identity *P*-value	0.23	0.0	1.00
Sequence Smith–Waterman score	1400	–7400	15 000
Structural similarity score	4400	97	15 000
Structural similarity score *P*-value	0.015	0.0	1.00

Morphs in the database were processed to eliminate redundancy (several PDB pairs have multiple morph movies of varying characteristics) and then fed into the Yale Structural Alignment Server (URL: http://bioinfo.mbb.yale.edu/align) based on structure alignment (29). Structure alignment was able to structurally align 65 of the 78 non-redundant protein chain pairs. The results for 65 observations are shown in the table above to two significant figures. On average, successful protein chain comparisons in the database have a sequence percent identity of 55%, although the server was able to successfully morph proteins with as little sequence identity as 7.9% identity and as high as 100% identity. Morphed proteins have a mean trimmed RMS (RMS after worst-fitting half of residues eliminated) of 2.4 Å, with a range between 0.245 and 16.46 Å. The server was able to successfully morph protein chains with *P*-values based on all three statistics (Trimmed RMS, Sequence percent identity and Structural Similarity Score) near one, suggesting that some protein chain pairs in the database are unlikely to be related wither evolutionarily or structurally.

Table 4. Morph torsion angle statistics

Name	Mean of max.	Min of max.	Max of max.
Maximum alpha change	140°	16°	180°
Maximum phi change	180°	140°	180°
Maximum psi change	150°	23°	180°

Maximum torsion angle changes is another example of the statistics collected by the server. For this table, maximum alpha, phi and psi torsion angle changes were computed for 134 protein chain pairs in the database and reported here to two significant figures. The mean, minimum and maximum of each statistic were computed for the table above. As expected, a motion can be found for each statistic with a torsion angle change of 180°, the maximum possible. Every motion involves at least one large phi angle change of at least 140°. However, few morphs have only small psi and alpha torsion angle changes. Alpha is the dihedral angle relating virtual bonds connecting Cα atoms between residues along the peptide chain; it is computed by pretending each residue is an atom with its center at its Cα atom.

over a range of 0–150° as computed by our algorithm, whereas the 12 motions culled from the scientific literature had an average rotation of 24° over a range of 5–148°. Similarly, our algorithms found a median maximum Cα displacement of 17 Å ranging from 0 to 81 Å for the submitted motions, whereas 11 motions reported in the scientific literature average 12 Å over a range of 1.5–60 Å. 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 (see Table 3). Most motions have at least one large torsion angle change (see Table 4).

The sparseness of manually culled data in Table 2 is due to the lack of a standardized nomenclature for these statistics in the scientific literature. It is worth noting that a different set of proteins had to be used for each of the manually culled tallies in Table 2. Because these statistics predate the server, they serve as a manual "gold standard" against which the results of the server may be compared. Table 1 presents a statistical description of motions in the database, a main scientific benefit of the server.

Integration with database

Privacy is a concern with some submissions, so users are afforded the option to either keep their submissions secret until the results have been published or to cause the submission to appear immediately in an index. For each successfully completed morph, the server produces a Web page allowing easy download of the coordinates (as an archive of PDB files or in NMR format) or movies (in a number of video formats), in addition to displaying the molecule in the moving VRML format. The page includes the standardized statistics discussed above generated for the conformations used in the morph. This page may be accessed through a URL containing a special code that is emailed back to the submitting user when the morph is complete; for users seeking to keep their morphs private (for

publication reasons), this URL serves as the user's password, allowing access to the morph page in the server. For public morphs, these pages are also accessible through an index, http://bioinfo.mbb.yale.edu/MolMovDB/movies

The ultimate flow of information is circular. For each motion we either link it via a motion ID to an existing entry in the Macromolecular Motions Database or we generate a new entry in the database. The results of analyzing particular ordered sets of structures ('strings' of structures) are entered under an appropriate identifier into the Database of Macromolecular Motions for further reference, and, in many cases, suggest further structures to study and analyze. Each comparison is assigned a unique ID entered into the 'comparison table' in the database that references the IDs of the PDB structures involved. These comparisons are, in turn, referenced by entries in the motions database (these references may be generated by comparing the IDs of the PDB structures referenced in each comparison table entry with the PDB structures referenced in each motion table entry.). Because many motions in the database are associated with more than two structures, more than one comparison is often possible and some database entries do reference multiple comparisons.

New movies, which lack a motion entry in the Database of Macromolecular Motions, have an entry automatically created with minimal or no annotation. This is indicated in the entry by setting the annotation level to zero. (Annotation levels range from 0 to 10. A level of '0' indicates the entry was automatically created with no human intervention. '10' indicates significant human intervention, typically in the form of a large amount of descriptive text present in the entry.) The user can annotate the new entry using an easy-to-use edit form displayed in his or her Web browser. Existing entries are also editable by the community through the same Web form with prior authorization from the database's maintainers. All changes are subsequently reviewed by the maintainers to assure quality control. In this way, the Database of Macromolecular Motions is used to classify and organize morphs submitted to the morph server.

Examples

To illustrate the technique of adiabatic mapping as implemented by the server, Figure 5 depicts the frames in five automatically generated morphs produced by the server's adiabatic mapping interpolation engine.

ADH. First is a 'trivial' morph, alcohol dehydrogenase (56,57). This morph is considered 'trivial' because a true motion is involved, and the endpoint conformations are sufficiently close together that a pathway between the two— not involving chain breaks or clearly distorted geometry— is easy to construct in the mind's eye. Therefore, one would intuitively expect software that claims to perform morphing to handle this case with similar ease. This is indeed the case, as can be seen in the figure, which depicts the frames generated by the morph server for the morph of alcohol dehydrogenase. The protein has very little movement; the figure shows the frames in the motion generated by our server with an arrow to indicate the region of movement. When the actual animation is played back, the arrow is not necessary, as the eye has evolved to be especially sensitive to motion and easily picks out the movement in the movie.

Recoverin. Recoverin (58–60) is an example of a 'typical' morph. The morph is considered 'typical' because a true motion is involved, as can be seen in the figure, the motion involves most of the molecule and is therefore qualitatively more extensive than that of a 'trivial' motion such as alcohol dehydrogenase. The motion is sufficiently complicated that a simple linear interpolation would produce at least some obvious distortion and physical impossibilities. Nevertheless, the adiabatic mapping interpolation engine produces a realistic morph without chain breaks or clearly distorted geometry.

Pol-beta. A morph of DNA polymerase beta (5,61), considered 'typical' for much the same reasons as recoverin.

GroEL. The user should be aware of a number of problems that can be encountered in the adiabatic mapping method. Problems arise for large deformations if the energy minimization methods cannot effectively remove the accumulated stresses (62). These problems are endemic to all adiabatic mapping systems, including our Web server. This problem is illustrated in Figure 5, which shows a morph of one subunit in GroEL (11–14), a 'medium difficulty' morph because of the considerable atomic displacements between the starting and ending conformations. However, this GroEL motion still represents a true protein motion, and the server still produces a fairly realistic interpolation. One means of improving this morph would be to have the user select additional interpolation frames (and, hence, additional energy minimizations). (This is in a sense a 'feature' that highlights which motions are sterically more difficult to achieve.)

DT. Our model will, of course, break when fed an 'impossible' morph, as shown in Figure 5. The endpoint conformations are that of diphtheria toxin (DT), not a true motion but rather an example of domain swapping (63,64) in which the domains in DT have been solved while bound in two different configurations. For one conformation to 'morph' into another, the easiest physically realistic route would be for one domain to unfold and refold. Indeed, the morph generated by the server does suggest a process of this sort.

While the morph server is unable to generate a physically realistic movie of this 'motion', it does suggest that the morph server may be used as a quick visual tool in evaluating the validity of a proposed motion. Comprehensive statistics for all five morphs may be found in Table 1.

DISCUSSION

Statistics

In a majority of cases the structures of a given macromolecule involved in motions have been solved in two or more conformations, so that endpoints for the motion are available. This, in turn, means that automatic conformation comparison tools are possible, which, when applied *en masse* to the motions database, allow the generation of a consistent, standardized set of statistics characterizing the motions in the

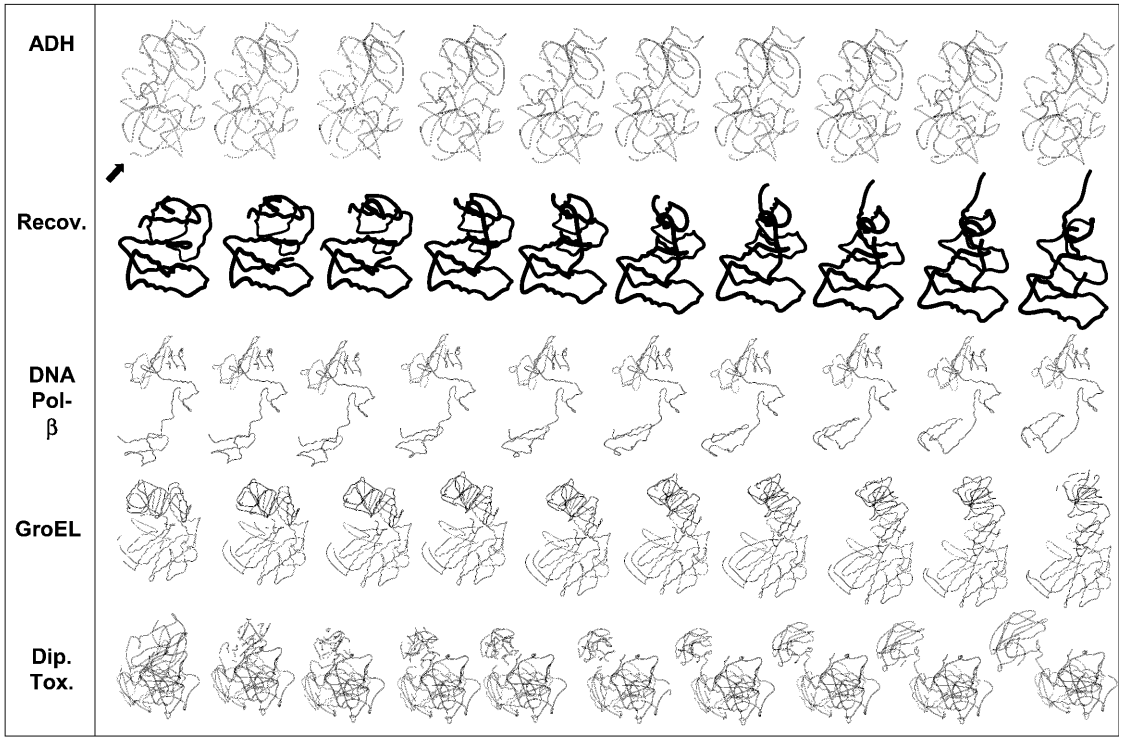

Figure 5. Sample morphs. An automatic morph of alcohol dehydrogenase (ADH) as produced by our server. Alcohol dehydrogenase is a 'trivial' case, as the motions involved are relatively small but nevertheless dramatic when viewed as a movie. It is shown in the top panel. The two panels below ADH show recoverin (1IKU→ 1JSA) and DNA polymerase beta, respectively, which are 'easy' cases. GroEL is shown as an intermediate case, as the motion are much larger than in alcohol dehydrogenase. The morph can still be reasonably handled by the server, and is especially dramatic on paper due to the large displacement of the motions involved. Diphtheria Toxin (Dip. Tox.) a hard or impossible case, because the rearrangement between the conformations does not involve a motion, but rather domain switching in the crystalline state. The poor quality of the morph provides the researcher with an immediate clue that the rearrangement pathway is unlikely to be a pure motion. The default MultiGif (or Moving Gif) using a combination of software, including Rasmol (54), Molscript (53), Ghostscript, and a gif to multigif utility, all driven through a Perl script. Additional software renders the molecule into Quicktime and MPEG formats to ensure display in a number of Internet browser environments. A simple HTML and Adobe PDF rendering of the sequence alignment of the residues between conformations is also available. In addition to visual output, the interpolated coordinates can also be downloaded as either an PDB NMR format archive or as an archive of PDB frames in the popular Unix Tape Archive ('.tar'file) format.

database. In the process of analyzing the structures, pathway interpolation is possible as well.

What constitutes an optimal morph?

Since the goal of the server was to output only a single interpolated pathway 'morph', it is necessary to define more precisely what is desired. Define the 'optimal morph' as the most likely (or most frequently taken) pathway between two conformations. In the large dimensional space of macromolecular atomic coordinate space, an infinite number of paths between conformations exist, so that establishing that a given local ensemble of pathways is the most statistically probable is, in general, computationally intractable. A more realistic approach would be simply to find a morph that is a reasonably good reaction coordinate that does not produce any large chemical distortions. This reduced the computational complexity of the problem, yet ensured that the resulting morph would be insightful, yet likely be very similar to the 'optimal' morph.

Unlike the adiabatic interpolation engine used in the server, a number of interpolation engines on proteins have taken approaches that do not meet these criteria. With the exception of the simplest motions, simple linear interpolations of atomic coordinates without consideration of physical reality yields intermediates with clearly distorted geometry. In some cases, atoms may be significantly closer than their van der Waals radii would permit, or further apart than a chemical bond would reasonably be expected to allow. A more sophisticated approach to morph movies not currently taken by the server due to its stringent computational requirements, but one which might be added in the future, involves the use of normal mode analysis, such as was done on GroEL by Xu *et al.* (11–14).

CONCLUSIONS

We have developed an integrated set of protein conformation comparison tools on the Web for use in conjunction with the Macromolecular Motions Database or as a stand-alone, publicly accessible server. When solved endpoint structures are available, the server can produce a useful comparison of the structures involved in protein motions. The server also implements a database of protein motions accessible on the Web or generated by Internet users through our server; this database is integrated into the Molecular Motions Database.

The server collects a number of statistics on the motion, including maximum Cα displacement and maximum rotation around the putative hinge, which are useful both in analyzing and classifying individual proteins and in generating a statistical picture of motions in the motions database as a whole. The software also homogenizes the incoming structures, attempting to solve for missing atoms using a molecular dynamics algorithm. The server then uses an adiabatic mapping technique to generate a visually rendered interpolated pathway, or 'morph', of the motion or evolution of the protein. The homogenized endpoint coordinates and the generated intermediate coordinates are made available for download.

The software presents the visual representation, statistics, orientation, alignment and interpolated coordinates to the user. At user option, these results may become public immediately or remain private until paper publication. Through an easy-to-use Web form, the user is afforded an opportunity to create a descriptive entry in the Database of Macromolecular Motions for the protein structures involved, referencing the morph results, as well. We have found the server useful in the analysis of protein motions and anticipate that use of the server will help standardize statistics and nomenclature for protein motions subsequently presented in the scientific literature.

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. Debrunner,P.G. and Frauenfelder,H. (1982) *Dynamics of proteins. Annu. Rev. Phys. Chem.*, **33**, 283.
2. Lipscomb,W.N. (1982) *Acceleration of reactions by enzymes. Accounts Chem. Res.*, **15**, 232.
3. Vonrhein,C., Schlauderer,G.J. and Schulz,G.E. (1995) *Movie of the structural changes during a catalytic cycle of nucleoside monophosphate kinases. Structure*, **3**, 483–490.
4. Vonrhein,C., Bonisch,H., Schafer,G. and Schulz,G.E. (1998) *The structure of a trimeric archaeal adenylate kinase. J. Mol. Biol.*, **282**, 167–179.
5. Sawaya,M.R., Prasad,R., Wilson,S.H., Kraut,J. and Pelletier,H. (1997) *Crystal structures of human DNA polymerase beta complexed with gapped and nicked DNA: evidence for an induced fit mechanism. Biochemistry*, **36**, 11205–11215.
6. Gilson,M.K., *et al.* (1994) *Open "Back Door" in a Molecular Dynamics Simulation of Acetylcholinesterase. Science*, **263**, 1276–1278.
7. Faerman,C., *et al.* (1996) *Site-directed mutants designed to test back-door hypotheses of acetylcholinesterase function. FEBS Lett.*, **386**, 65–71.
8. Ripoll,D.R., Faerman,C.H., Axelsen,P.H., Silman,I. and Sussman,J.L. (1993) *An electrostatic mechanism for substrate guidance down the aromatic gorge of acetylcholinesterase. Proc. Natl Acad. Sci. USA*, **90**, 5128–5132.
9. Pande,V.S. and Rokhsar,D.S. (1999) *Molecular dynamics simulations of unfolding and refolding of a beta-hairpin fragment of protein G [In Process Citation]. Proc. Natl Acad. Sci. USA*, **96**, 9062–9067.
10. Pande,V.S. and Rokhsar,D.S. (1999) *Folding pathway of a lattice model for proteins. Proc. Natl Acad. Sci. USA*, **96**, 1273–1278.
11. Xu,Z., Horwich,A.L. and Sigler,P.B. (1997) *The crystal structure of the asymmetric GroEL-GroES-(ADP)7 chaperonin complex. Nature*, **388**, 741–750.
12. Rye,H., *et al.* (1997) *Distinct actions of cis and trans ATP within the double ring of the chaperonin GroEL. Nature*, **388**, 792–798.
13. Xu,Z. and Sigler,P.B. (1998) *GroEL/GroES: structure and function of a two-stroke folding machine. J. Struct. Biol.*, **124**, 129–141.
14. Sigler,P.B., *et al.* (1998) *Structure and function in GroEL-mediated protein folding. Annu. Rev. Biochem.*, **67**, 581–608.
15. Verschelde,J.L., *et al.* (1998) *Analysis of three human interleukin 5 structures suggests a possible receptor binding mechanism. FEBS Lett.*, **424**, 121–126.
16. Crofts,A.R., *et al.* (1999) *Pathways for proton release during ubihydroquinone oxidation by the bc(1) complex. Proc. Natl Acad. Sci. USA*, **96**, 10021–10026.
17. Crofts,A.R. and Berry,E.A. (1998) *Structure and function of the cytochrome bc1 complex of mitochondria and photosynthetic bacteria. Curr. Opin. Struct. Biol.*, **8**, 501–509.
18. Bystrom,C.E., Pettigrew,D.W., Branchaud,B.P., O'Brien,P. and Remington,S.J. (1999) *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**, 3508–3518.
19. Feese,M.D., Faber,H.R., Bystrom,C.E., Pettigrew,D.W. and Remington,S.J. (1998) *Glycerol kinase from Escherichia coli and an Ala65→Thr mutant: the crystal structures reveal conformational changes with implications for allosteric regulation. Structure*, **6**, 1407–1418.
20. Thompson,A.B., *et al.* (1992) *Aerosolized beclomethasone in chronic bronchitis. Improved pulmonary function and diminished airway inflammation. Am. Rev. Respir. Dis.*, **146**, 389–395.
21. Sykes,J.A., Thomas,M.J., Goldie,D.J. and Turner,G.M. (1982) *Plasma lactoferrin levels in pregnancy and cystic fibrosis. Clin. Chim. Acta*, **122**, 385–393.
22. Martz,E. (1999) *Protein Explorer (software package)*. URL: http://www.umass.edu/microbio/chime/explorer/index.htm
23. Schnecke,V., Swanson,C.A., Getzoff,E.D., Tainer,J.A. and Kuhn,L.A. (1998) *Screening a peptidyl database for potential ligands to proteins with side-chain flexibility. Proteins*, **33**, 74–87.
24. Gerstein,M. and Krebs,W. (1998) *A Database of Macromolecular Movements. Nucleic Acids Res.*, **26**, 4280.
25. Gerstein,M.B., Jansen,R., Johnson,T., Park,B. and Krebs,W. (1999) *Studying Macromolecular Motions in a Database Framework: From Structure to Sequence*. In Thorpe,M.F. and Duxbury,P.M. (eds), *Rigidity Theory and Applications.* Kluwer Academic/Plenum Press, Norwell, MA, USA. pp. 401–442.
26. Gerstein,M. (1997) *A Structural Census of Genomes: Comparing Bacterial, Eukaryotic, and Archaeal Genomes in terms of Protein Structure. J. Mol. Biol.*, **274**, 562–576.
27. Wilson,C.A., Kreychman,J. and Gerstein,M. (2000) *Assessing Annotation Transfer for Genomics: Quantifying the relations between protein sequence, structure, and function through traditional and probabilistic scores. J. Mol. Biol.*, **297**, 233–249.
28. Gerstein,M. and Levitt,M. (1998) *Comprehensive Assessment of Automatic Structural Alignment against a Manual Standard, the Scop Classification of Proteins. Protein Sci.*, **7**, 445–456.
29. Levitt,M. and Gerstein,M. (1998) *A Unified Statistical Framework for Sequence Comparison and Structure Comparison.* (1998) *Proc. Natl Acad. Sci. USA*, **95**, 5913–5920.
30. Murzin,A., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) *SCOP: A Structural Classification of Proteins for the Investigation of Sequences and Structures. J. Mol. Biol.*, **247**, 536–540.
31. Hubbard,T.J.P., Murzin,A.G., Brenner,S.E. and Chothia,C. (1997) *SCOP: a structural classification of proteins database. Nucleic Acids Res.*, **25**, 236–239.
32. Barton,G.J. and Sternberg,M.J.E. (1987) *A strategy for the rapid multiple alignment of protein sequences: Confidence levels from tertiary structure comparisons. J. Mol. Biol.*, **198**, 327–338.
33. Barton,G.J. and Sternberg,M.J.E. (1990) *Flexible protein sequence patterns: A sensitive method to detect weak structural similarities. J. Mol. Biol.*, **212**, 389–402.
34. Barton,G.J. (1990) *Methods Enzymol.*, **183**, 403–428.
35. Lesk,A.M. (1991) *Protein Architecture: A Practical Approach*. IRL Press, Oxford.
36. Gerstein,M. and Chothia,C.H. (1991) *Analysis of Protein Loop Closure: Two Types of Hinges Produce One Motion in Lactate Dehydrogenase. J. Mol. Biol.*, **220**, 133–149.

37. Gerstein,M. and Altman,R. (1995) *Average core structures and variability measures for protein families: Application to the immunoglobulins. J. Mol. Biol.*, **251**, 161–175.
38. Lesk,A.M. and Chothia,C. (1984) *Mechanisms of Domain Closure in Proteins. J. Mol. Biol.*, **174**, 175–191.
39. Gerstein,M., Lesk,A. and Chothia,C. (1994) *Structural Mechanisms for Protein Motions. Biochemistry*, **33**, 6739–6749
40. Wriggers,W. and Schulten,K. (1997) *Protein domain movements: detection of rigid domains and visualization of hinges in comparisons of atomic coordinates. Proteins*, **29**, 1–14.
41. Maiorov,V. and Abagyan,R. (1997) *A new method for modeling large-scale rearrangements of protein domains. Proteins*, **27**, 410–424.
42. Brünger,A.T. (1993) *X-PLOR 3.1, A System for X-ray Crystallography and NMR*. Yale University Press, New Haven.
43. Hogue,C.W., Ohkawa,H. and Bryant,S.H. (1996) *A dynamic look at structures: WWW-Entrez and the Molecular Modeling Database. Trends Biochem. Sci.*, **21**, 226–229.
44. Ohkawa,H., Ostell,J. and Bryant,S. (1995) *MMDB: an ASN.1 specification for macromolecular structure. ISMB*, **3**, 259–267.
45. Kleywegt,G.J. and Jones,T.A. (1995) *Where freedom is given, liberties are taken. Structure*, **3**, 535–540.
46. Kleywegt,G.J. and Jones,T.A. (1996) *Phi/psi-chology: Ramachandran revisited. Structure*, **4**, 1395–1400.
47. McCammon,J.A. and Harvey,S.C. (1987) *Dynamics of Proteins and Nucleic Acids*. Cambridge University Press.
48. Moffat,K. (1989) *Time-resolved macromolecular crystallography. Annu. Rev. Biophys. Biophys. Chem.*, **18**, 309–332.
49. Genick,U.K., *et al.* (1997) *Structure of a protein photocycle intermediate by millisecond time-resolved crystallography. Science*, **275**, 1471–1475.
50. Silicon-Graphics (1996) *VRML 2 Specification*. URL: http://webspace.sgi.com/moving-worlds/Design.html
51. Pesce,M. (1995) *VRML*. New Riders Indianapolis, IN.
52. Introduction to Acrobat Reader, in *Description of Portable Document Format*, Adobe Corporation: URL: http://www.adobe.com/supportservice/custsupport/SOLUTIONS/ac76.htm
53. Kraulis,P.J. (1991) *MOLSCRIPT – A program to produce both detailed and schematic plots of protein structures. J. Appl. Crystallogr.*, **24**, 946–950.
54. Sayle,R. and Milner-White,E.J. (1995) *RASMOL: biomolecular graphics for all. Trends Biochem. Sci.*, **20**, 374.
55. Brooks,B.R., *et al.* (1983) *CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations. J. Comp. Chem.*, **4**, 187–217.
56. Colonna-Cesari,F., *et al.* (1986) *Interdomain motion in liver alcohol dehydrogenase: structural and energetic analysis of the hinge bending mode. J. Biol. Chem.*, **261**, 15273–15280.
57. Eklund,H., *et al.* (1981) *Structure of a triclinic ternary complex of horse liver alcohol dehydrogenase at 2.9 Å resolution. J. Mol. Biol.*, **146**, 561–587.
58. Ames,J.B., *et al.* (1997) *Molecular mechanics of calcium-myristoyl switches. Nature*, **389**, 198–202.
59. Tanaka,T., Ames,J.B., Harvey,T.S., Stryer,L. and Ikura,M. (1995) *Sequestration of the membrane-targeting myristoyl group of recoverin in the calcium-free state. Nature*, **376**, 444–447.
60. Ames,J.B., Tanaka,T., Ikura,M. and Stryer,L. (1995) *Nuclear magnetic resonance evidence for $Ca^{(2+)}$-induced extrusion of the myristoyl group of recoverin. J. Biol. Chem.*, **270**, 30909–30913.
61. Sawaya,M.R., Pelletier,H., Kumar,A., Wilson,S.H. and Kraut,J. (1994) *Crystal structure of rat DNA polymerase beta: evidence for a common polymerase mechanism. Science*, **264**, 1930–1935.
62. Tung,C.S., Harvey,S.C. and McCammon,J.A. (1984) *Large-amplitude bending motions in phenylalaine transfer RNA. Biopolymers*, **23**, 2173.
63. Bennett,M.J., Schlunegger,M.P. and Eisenberg,D. (1995) *3D domain swapping: a mechanism for oligomer assembly. Protein Sci.*, **4**, 2455–2468.
64. Bennett,M.J., Choe,S. and Eisenberg,D. (1994) *Domain swapping: entangling alliances between proteins. Proc. Natl Acad. Sci. USA*, **91**, 3127–3131.