STUDYING MACROMOLECULAR MOTIONS IN A DATABASE FRAMEWORK: FROM STRUCTURE TO SEQUENCE

Mark Gerstein,[1] Ronald Jansen,[1]
Ted Johnson,[1] Jerry Tsai,[2] and Werner Krebs[1]

[1]Department of Molecular Biophysics and Biochemistry
Yale University, 266 Whitney Avenue
New Haven, CT 06511

[2]Department of Structural Biology, Stanford University
Stanford, CA 94305

ABSTRACT

We describe database approaches taken in our lab to the study of protein and nucleic acid motions. We have developed a database of macromolecular motions, which is accessible on the World Wide Web with an entry point at http://bioinfo.mbb.yale.edu/MolMovDB. This attempts to systematize all instances of macromolecular movement for which there is at least some structural information. At present it contains detailed descriptions of more than 100 motions, most of which are of proteins. Protein motions are further classified hierarchically into a limited number of categories, first on the basis of size (distinguishing between fragment, domain, and subunit motions) and then on the basis of packing. Our packing classification divides motions into various categories (shear, hinge, other) depending on whether or not they involve sliding over a continuously maintained and tightly packed interface. We quantitatively systematize the description of packing through the use of Voronoi polyhedra and Delaunay triangulation. In addition to the packing classification, the database provides some indication about the evidence behind each motion (i.e. the type of experimental information or whether the motion is inferred based on structural similarity) and attempts to describe many aspects of a motion in terms of a standardized nomenclature (e.g. the maximum rotation, the residue selection of a fixed core, etc). Currently, we use a standard relational design to implement the database. However, the complexity and heterogeneity of the information kept in the database makes it an ideal application for an object-relational approach, and we are moving it in this direction. The database, moreover, incorporates innovative Internet cooperatively features that allow authorized remote experts to serve as database editors. The database also contains plausible representations for

motion pathways, derived from restrained 3D interpolation between known endpoint conformations. These pathways can be viewed in a variety of movie formats, and the database is associated with a server that can automatically generate these movies from submitted coordinates. Based on the structures in the database we have developed sequence patterns for linkers and flexible hinges and are currently using these for the annotation of genome sequence data.

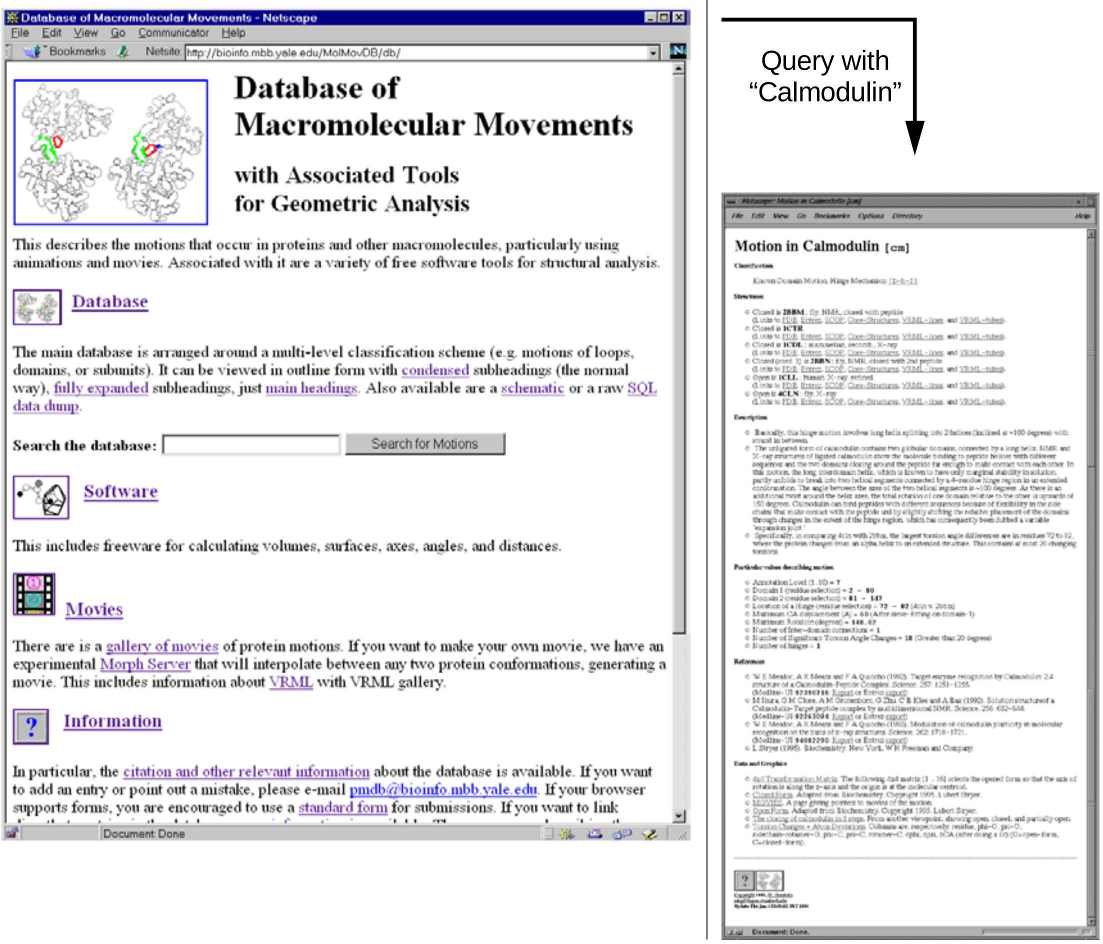
INTRODUCTION

Motion is frequently the way macromolecules (proteins and nucleic acid) carry out particular functions; thus motion often serves as an essential link between structure and function. In particular, protein motions are involved in numerous basic functions such as catalysis, regulation of activity, transport of metabolites, formation of large assemblies and cellular locomotion. In fact, highly mobile proteins have been implicated in a number of diseases—e.g., the motion of gp41 in AIDS and that of the prion protein in scrapie[1-5]. Another reason for the study of macromolecular motions results from their fundamental relationship to the principles of protein and nucleic acid structure and stability.

Macromolecular motions are amongst the most complicated biological phenomena that can be studied in great quantitative detail, involving concerted changes in thousands of precisely specified atomic coordinates. Fortunately, it is now possible to study these motions in a database framework, by analyzing and systematizing many of the instances of protein structures solved in multiple conformations. We summarize here some recent work in our laboratory relating to the construction of a database of protein motions[6]) and the use of Voronoi polyhedra to study packing[7]. We also present some preliminary results relating to creating sequence patterns for hinges and flexible linkers.

Table 1. Statistics for the Mechanism of the Motions. This table cross-tabulates the two main classifying attributes of motions: their size (row heads) and their packing characteristics (column heads). We define a known motion to be a motion with two or more solved conformations, and a suspected motion is defined to have only one or fewer solved conformations. (Adapted from Gerstein and Krebs (1998).[6])

Size	Domain		Fragment		Subunit		Total	
Mechanism								
Hinge	38	51%	16	59%			54	45%
Shear	14	19%	3	11%			17	14%
Partial Refolding	5	7%					5	4%
Allosteric					8	57%	8	7%
Other/Non-Allosteric	2	3%	1	4%	6	43%	9	7%
Unclassifiable	15	20%	7	26%			22	18%
Notably Motionless							1	1%
Complex							2	2%
Nucleic Acid							3	2%
Known / % category	53	72%	25	93%	11	79%	94	78%
Suspected / % category	21	28%	2	7%	3	21%	27	22%
Totals / % DB	74	62%	27	23%	14	12%	121	100%



Database of Macromolecular Movements - Netscape
File Edit View Go Communicator Help
Bookmarks
Netsite: http://bioinfo.mbb.yale.edu/MolMovDB/db/

Database of Macromolecular Movements

with Associated Tools for Geometric Analysis

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

Database

The main database is arranged around a multi-level classification scheme (e.g. motions of loops, domains, or subunits). It can be viewed in outline form with condensed subheadings (the normal way), fully expanded subheadings, just main headings. Also available are a schematic or a raw SQL data dump.

Search the database: [ ] Search for Motions

Software

This includes freeware for calculating volumes, surfaces, axes, angles, and distances.

Movies

There are is a gallery of movies of protein motions. If you want to make your own movie, we have an experimental Morph Server that will interpolate between any two protein conformations, generating a movie. This includes information about VRML with VRML gallery.

Information

In particular, the citation and other relevant information about the database is available. If you want to add an entry or point out a mistake, please e-mail pmdb@bioinfo.mbb.yale.edu. If your browser supports forms, you are encouraged to use a standard form for submissions. If you want to link

Query with "Calmodulin"

Netscape: Motion in Calmodulin [cm]
File Edit View Go Bookmarks Options Directory Help

Motion in Calmodulin [cm]

Classification

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

Structures

- Closed is **2BBM** : fly, NMR, closed with peptide (Links to PDB, Entrez, SCOP, Core-Structures, VRML-lines, and VRML-tubes).
- Closed is **1CTR** (Links to PDB, Entrez, SCOP, Core-Structures, VRML-lines, and VRML-tubes).
- Closed is **1CDL** : mammelian, recomb., X-ray (Links to PDB, Entrez, SCOP, Core-Structures, VRML-lines, and VRML-tubes).
- Closed (conf. 3) is **2BBN** : fly, NMR, closed with 2nd peptide (Links to PDB, Entrez, SCOP, Core-Structures, VRML-lines, and VRML-tubes).
- Open is **1CLL** : human, X-ray, refined (Links to PDB, Entrez, SCOP, Core-Structures, VRML-lines, and VRML-tubes).
- Open is **4CLN** : fly, X-ray (Links to PDB, Entrez, SCOP, Core-Structures, VRML-lines, and VRML-tubes).

Description

- Basically, this hinge motion involves long helix splitting into 2 helices (inclined at ~100 degrees) with strand in between.
- The unligated form of calmodulin contains two globular domains, connected by a long helix. NMR and X-ray structures of ligated calmodulin show the molecule binding to peptide helices with different sequences and the two domains closing around the peptide far enough to make contact with each other. In this motion, the long interdomain helix, which is known to have only marginal stability in solution, partly unfolds to break into two helical segments connected by a 4-residue hinge region in an extended conformation. The angle between the axes of the two helical segments is ~100 degrees. As there is an additional twist around the helix axes, the total rotation of one domain relative to the other is upwards of 150 degrees. Calmodulin can bind peptides with different sequences because of flexibility in the side chains that make contact with the peptide and by slightly shifting the relative placement of the domains through changes in the extent of the hinge region, which has consequently been dubbed a variable 'expansion joint.'
- Specifically, in comparing 4cln with 2bbm, the largest torsion angle differences are in residues 72 to 82, where the protein changes from an alpha helix to an extended structure. This contains at most 20 changing torsions.

Particular values describing motion

- Annotation Level (1..10) = **7**
- Domain 1 (residue selection) = **2 - 80**
- Domain 2 (residue selection) = **81 - 147**
- Location of a Hinge (residue selection) = **72 - 82** (4cln v. 2bbm)
- Maximum CA displacement (A) = **60** (After sieve-fitting on domain-1)
- Maximum Rotation (degrees) = **148.02**
- Number of Inter-domain connections = **1**
- Number of Significant Torsion Angle Changes = **18** (Greater than 20 degrees)
- Number of hinges = **1**

References

- W E Meador, A R Means and F A Quiocho (1992). Target enzyme recognition by Calmodulin: 2.4 structure of a Calmodulin-Peptide Complex. Science. 257: 1251-1255. (Medline-UI **92390718**: Export or Entrez export)
- M Ikura, G M Clore, A M Gronenborn, G Zhu, C B Klee and A Bax (1992). Solution structureof a Calmodulin-Target peptide complex by multidimensional NMR. Science. 256: 632-638. (Medline-UI **92263094**: Export or Entrez export)
- W E Meador, A R Means and F A Quiocho (1993). Modulation of calmodulin plasticity in molecular recognition on the basis of x-ray structures. Science. 262: 1718-1721. (Medline-UI **94082290**: Export or Entrez export)
- L Stryer (1995). Biochemistry. New York, W H Freeman and Company.

Data and Graphics

- 4x4 Transformation Matrix. The following 4x4 matrix [1 .. 16] orients the opened form so that the axis of rotation is along the z-axis and the origin is at the molecular centroid.
- Closed Form. Adapted from Biochemistry. Copyright 1995, Lubert Stryer.
- MOVIES. A page giving pointers to movies of the motion.
- Open Form. Adapted from Biochemistry. Copyright 1995, Lubert Stryer.
- The closing of calmodulin in 5 steps. From another viewpoint, showing open, closed, and partially open.
- Torsion Changes + Atom Deviations. Columns are, respectively: residue, phi-O, psi-O, sidechain-rotamer-O, phi-C, psi-C, rotamer-C, dphi, dpsi, dCA (after doing a fit) (O=open-form, C=closed-form).

Figure 1. 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. RIGHT shows an entry retrieved by such a keyword search (the entry for calmodulin). Graphics and movies are accessed by clicking on an entry page. (These have been deliberately segregated from the textual parts of the database since the interface was designed to make it easy to use on a low-bandwidth, text-only browser, e.g. lynx or the original www_3.0.) The main URL for the database is http://bioinfo.mbb.yale.edu/MolMovDB**.** 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 [8-15].

THE DATABASE

The primary public interface to the database consists of coupled hypertext documents available over the World Wide Web at http://bioinfo.mbb.yale.edu/MolMovDB. As shown in Figure 1, use of the web interface is straightforward and simple. The database may be browsed either by typing various search keywords into the main page or by navigating through an outline. Either way brings one to the entries. Thus far, the database has ~120 entries, which reference over 240 structures in the Protein Databank (PDB) (Table 1).

Table 2. Standard Statistics for the Magnitude of the Motions. The motions in the database range greatly in size, with maximum mainchain displacements between 1.5 and 60 Å. All the statistics are for version 1.7 of the database, based on the relatively small set of values culled from the literature. The averages are only approximate given the sparse nature of the data. We are developing software tools to extract these values automatically from structural data. (Adapted from Gerstein and Krebs (1998).[86])

Value	Num. Entries	min	max	average
Maximum Cα displacement	11	1.5	60	12
Maximum Atomic Displacement	3	8.8	10	9.3
Maximum Rotation	12	5	148	24
Maximum Translation	2	0.7	2.7	1.7

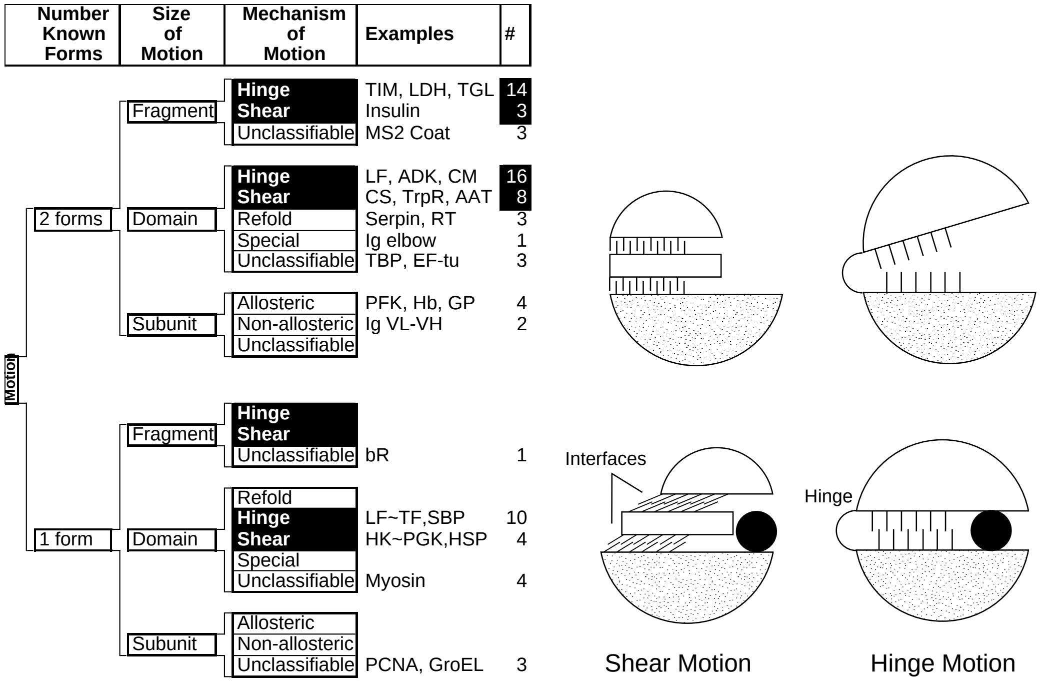


Figure 2. Schematic Showing the Overall Classification Scheme for Motions. TOP-LEFT, the database is organized around a hierarchical classification scheme, based on size (fragment, domain, subunit) and then packing (hinge or shear). Currently, the hierarchy also contains a third level for whether or not the motion is inferred. TOP-RIGHT is a schematic showing the difference between shear (sliding) and hinge motions. Figure adapted from[20,45]. 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. The essential characteristics of the various motions are summarized below. (Adapted from Gerstein and Krebs (1998).[86])

Unique Motion Identifier

Each entry is indexed by a *unique motion identifier*, rather than around individual proteins and nucleic acids. This is necessary because a single macromolecule can not only have a number of motions, but the essential motion can be shared amongst a number of different macromolecules.

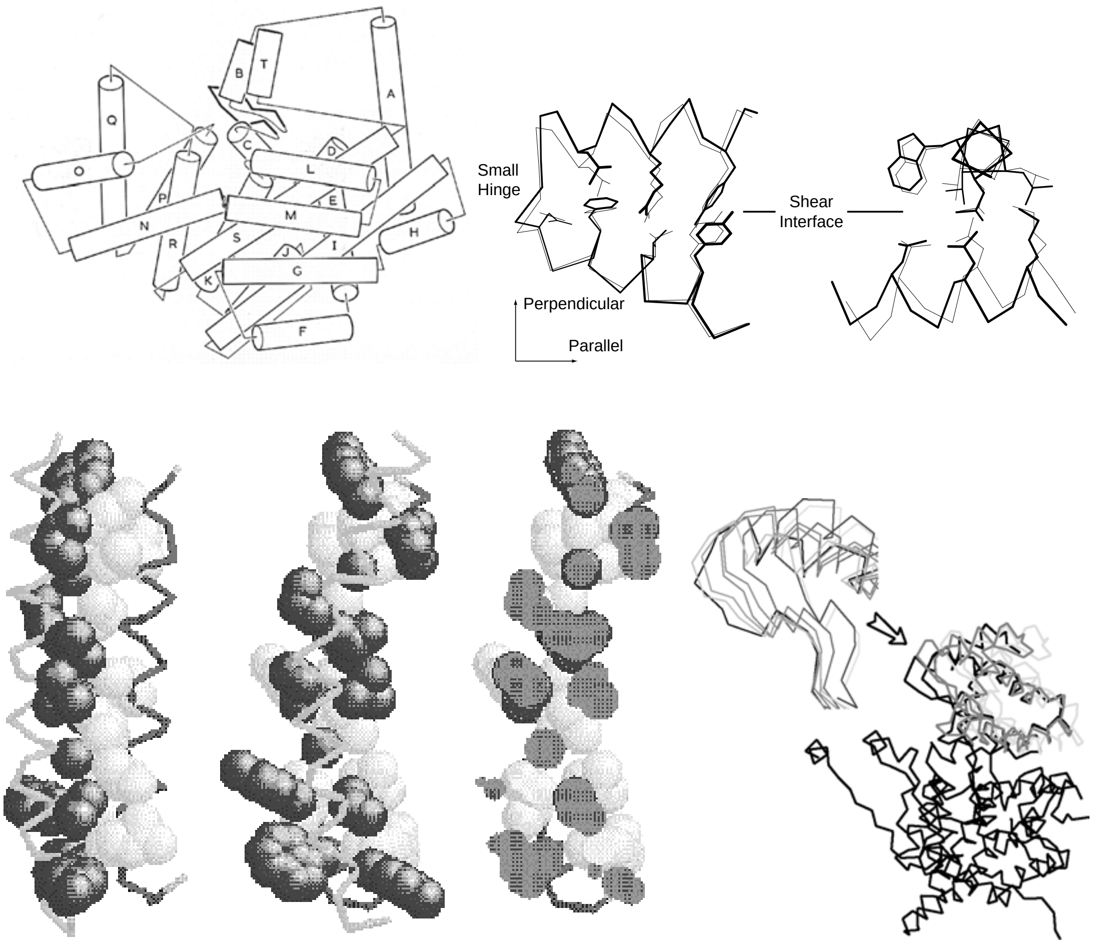


Figure 3. Closeup on the Shear Mechanism. The figure gives a close up illustrating shear motion in one protein, citrate synthase[20,93]. TOP-LEFT, Cartoon of one subunit of citrate synthase (1CTS) gives an overall view of the protein showing that it is composed of many helices. The adjacent one is related by two-fold axis shown. The small two-stranded sheet is omitted to improve clarity. a-helices are represented by cylinders. The small domain contains helices N, O, P, Q, and R. TOP-MIDDLE and TOP-RIGHT show representative shear motions between close-packed helices. Note how the mainchain only shifts by a small amount and the sidechains stay in the same rotamer configuration. BOTTOM-LEFT highlights the “knobs into holes” interdigitation of two close-packed helices. BOTTOM-RIGHT shows how these small motions can be added together to produce a large overall motion. Specifically, many small motions add up to shift helix O by 10.1 Å and rotate it by 28°. The incremental motion in shear domain closure is shown by Ca traces of the whole protein and of a closeup of the OP loop. BLACK is the apo form; WHITE, holo form; GRAY, cumulative effect of motion over the K, P, and then Q helix-helix interfaces. (The apo form was fit to the holo form, first on the core, and then on the K, P, and Q helices.) (Parts adapted from Gerstein and Krebs (1998).[86])

[*] At the time of writing, the PDB contained in excess of 6600 protein structures, but less than 600 nucleic acids structures.

[†] There is, of course, also the motion (i.e. rotation) of individual sidechains, often on the protein surface. However, this is on a much smaller scale than the motion of fragments or domains. It also occurs in all proteins. Consequently, sidechain motions are not considered to constitute individual motions in the database, being considered here a kind of background, intrinsic flexibility, common to all proteins.

insulin[31-33]. Often domain and fragment motions involve portions of the protein closing around a binding site, with a bound substrate stabilizing a closed conformation. They, consequently, provide a specific mechanism for induced-fit in protein recognition[34,35]. In enzymes this closure around a binding site has been analyzed in particular detail[36-40]. It serves to position important chemical groups around the substrate, shielding it from water and preventing the escape of reaction intermediates.

Subunit motion is distinctly different from fragment or domain motion. It affects two large sections of polypeptide that are *not* covalently connected. It is frequently part of an allosteric transition and tied to regulation[41,42]. The relative motions of the subunits in the transport protein hemoglobin and the enzyme glycogen phosphorylase change the affinity with which these proteins bind to their primary substrates[43,44] and are good examples.

Packing Classification: Hinge and Shear

For protein motions of domains and smaller units, we have systematized the motions on the basis of packing, using a scheme developed previously[6,28]. This is because the tight packing of atoms inside of proteins provides a most fundamental constraint on protein structure[45-50]. Unless there is a cavity or packing defect, it is usually impossible for an atom inside a protein to move much without colliding with a neighboring atom[51,52].

Internal interfaces between different parts of a protein are packed very tightly[7,28,53]. Furthermore, they are not smooth, but are formed from interdigitating sidechains. Common sense consideration of these aspects of interfaces places strong constraints on how a protein can move and still maintain its close packing. Specifically, maintaining packing throughout a motion implies that the sidechains at the interface must maintain their same relative orientation and pattern of inter-sidechain contacts in both conformations (e.g. open and closed).

These straightforward constraints on the types of motions that are possible at interfaces allow an individual movement within a protein to be described in terms of two basic mechanisms, shear and hinge, depending on whether or not it involves sliding over a continuously maintained interface[28] (Figure 2). A complete protein motion (which can contain many of these smaller “movements”) can be built up from these basic mechanisms. For the database, a motion is classified as *shear* if it predominately contains shear movements and as *hinge* if it is predominately composed of hinge movements. More detail on the characteristics of the two types of motion follows.

Shear. As shown in Figure 3, the shear mechanism basically describes the special kind of sliding motion a protein must undergo if it wants to maintain a well-packed interface. Because of the constraints on interface structure described above, individual shear motions have to be very small. Sidechain torsion angles maintain the same rotamer configuration[54] (with <15° rotation of sidechain torsions); there is no appreciable mainchain deformation; and the whole motion is parallel to the plane of the interface, limited to total translations of ~2 Å and rotations of 15°. Since an individual shear motion is so small, a single one is not sufficient to produce a large overall motion, and a number of shear motions have to be concatenated to give a large effect — in a similar fashion to each plate in a stack of plates sliding slightly to make the whole stack lean considerably. Examples include the Trp repressor and aspartate amino transferase[55,56].

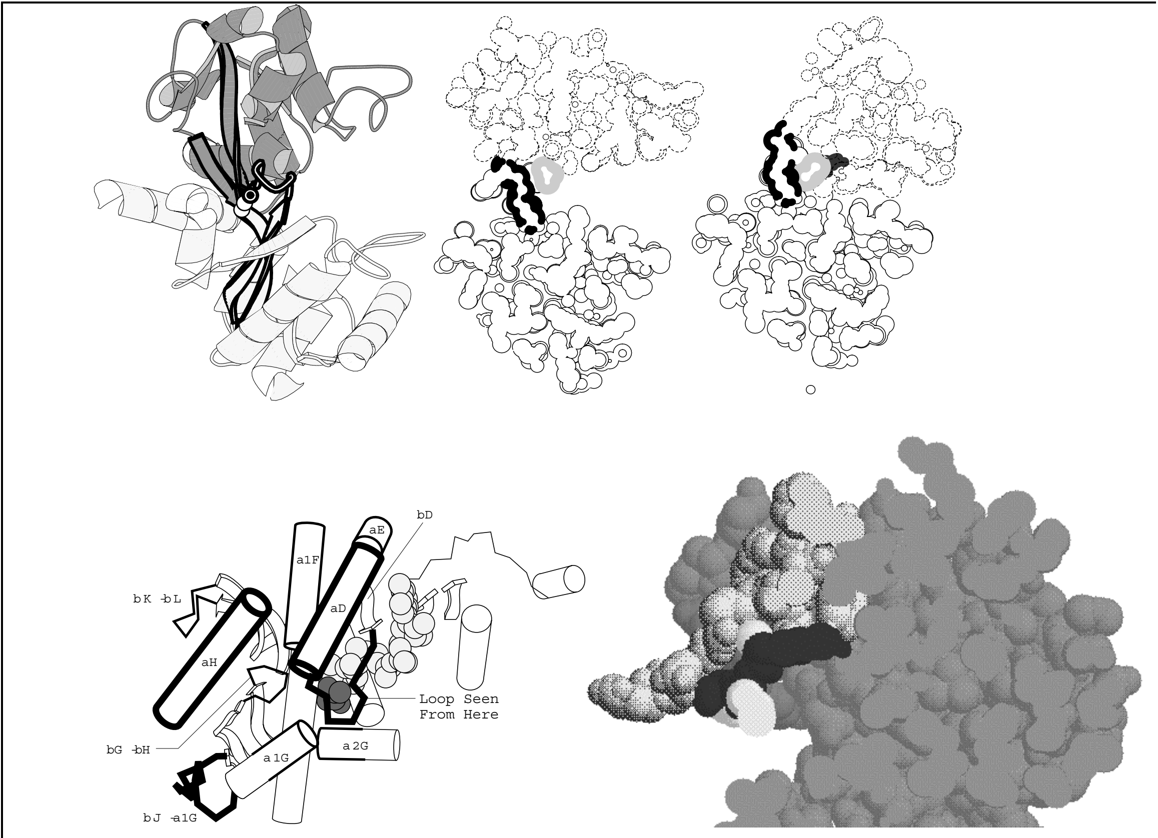
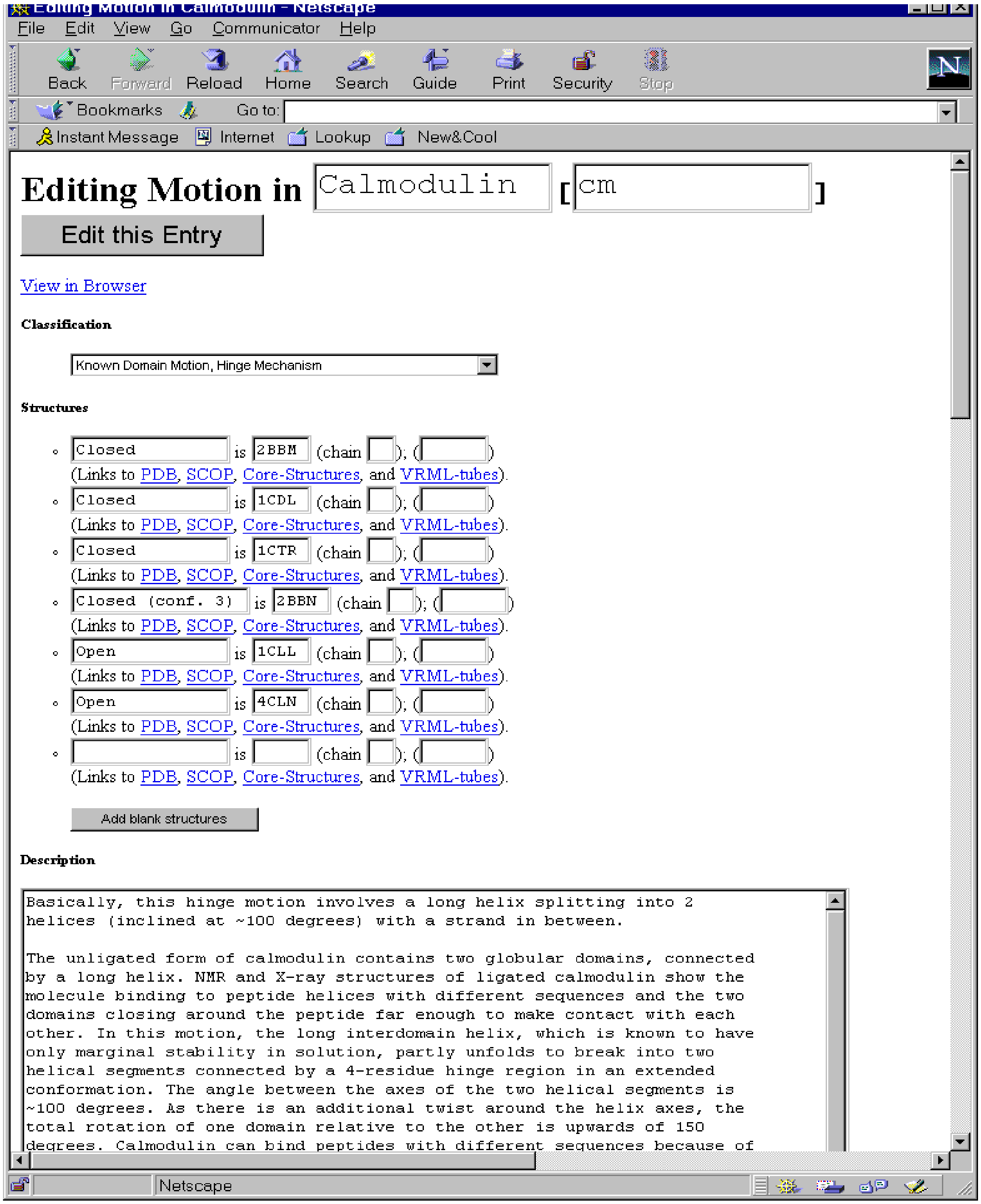


Figure 4. Close-up on the Hinge Mechanism. The figure shows the hinge motion in lactoferrin[20,45]. FAR-LEFT shows a ribbon drawing of the protein in the open conformation. The view is down the screw-axis, which is indicated in the figure by the circle with the dot in it. The screw-axis passes very close to the hinge region, which occurs in the middle of two beta strands (highlighted in bold). MIDDLE-LEFT and MIDDLE-RIGHT show the open and closed conformations in terms of space filling slices. The hinge region is highlighted by a thick black line. Note how few packing constraints there are on the hinge in contrast to the other atoms in the protein. (Figure adapted from Gerstein (1993).[45]) BOTTOM-LEFT shows the placement of a mobile loop in another protein, lactate dehydrogenase.
BOTTOM-RIGHT shows a close-up of this loop that highlights the absence of close-packing at the base of the hinge. Hinge mainchain is shown in black (first hinge) and almost white (second hinge). Rest of protein is shown in shades of gray.

Hinge. As shown in Figure 4, hinge motions occur when there is *no* continuously maintained interface constraining the motion. These motions usually occur in proteins that have two domains (or fragments) connected by linkers (i.e. hinges) that are relatively unconstrained by packing. A few large torsion angle changes in the hinges are sufficient to produce almost the whole motion. The rest of the protein rotates essentially as a rigid body, with the axis of the overall rotation passing through the hinges. The overall motion is always perpendicular to the plane of the interface (so the interface exists in one conformation but not in the other, as in the closing and opening of a book) and is identical to the local motion at the hinge. Examples include lactoferrin and tomato bushy stunt virus (TBSV)[57,58].


Editing Motion in Calmodulin - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Guide Print Security Stop
Bookmarks Go to:
Instant Message Internet Lookup New&Cool
Edit this Entry
Add blank structures
Netscape

Editing Motion in Calmodulin [cm]

View in Browser

Classification

Known Domain Motion, Hinge Mechanism

Structures

- Closed is 2BBM (chain); ( )
  (Links to PDB, SCOP, Core-Structures, and VRML-tubes).
- Closed is 1CDL (chain); ( )
  (Links to PDB, SCOP, Core-Structures, and VRML-tubes).
- Closed is 1CTR (chain); ( )
  (Links to PDB, SCOP, Core-Structures, and VRML-tubes).
- Closed (conf. 3) is 2BBN (chain); ( )
  (Links to PDB, SCOP, Core-Structures, and VRML-tubes).
- Open is 1CLL (chain); ( )
  (Links to PDB, SCOP, Core-Structures, and VRML-tubes).
- Open is 4CLN (chain); ( )
  (Links to PDB, SCOP, Core-Structures, and VRML-tubes).
- is (chain); ( )
  (Links to PDB, SCOP, Core-Structures, and VRML-tubes).

Description

```
Basically, this hinge motion involves a long helix splitting into 2
helices (inclined at ~100 degrees) with a strand in between.

The unligated form of calmodulin contains two globular domains, connected
by a long helix. NMR and X-ray structures of ligated calmodulin show the
molecule binding to peptide helices with different sequences and the two
domains closing around the peptide far enough to make contact with each
other. In this motion, the long interdomain helix, which is known to have
only marginal stability in solution, partly unfolds to break into two
helical segments connected by a 4-residue hinge region in an extended
conformation. The angle between the axes of the two helical segments is
~100 degrees. As there is an additional twist around the helix axes, the
total rotation of one domain relative to the other is upwards of 150
degrees. Calmodulin can bind peptides with different sequences because of
```

Figure 5. Editing a motion remotely over the Internet. The Database of Macromolecular Movements features an innovative Web form (shown here) that allows authorized remote users to collaborate and edit motions from remote sites around the world. Saved changes to motions may be previewed to see how they would appear to an end user and then applied to the database. If desired, saved changes can be made to appear immediately in the public Web interface to the database.

Gerstein et al.[53,59] analyzed the hinged domain and loop motion in specific proteins (lactate dehydrogenase, adenylate kinase, lactoferrin). These studies emphasized how critical the packing at the base of a protein hinge is (in the same sense that the "packing" at the base of an everyday door hinge determines whether or not the door can close). Protein hinges are special regions of the mainchain in the sense that they are exposed and have few packing constraints on them and are thus free to sharply kink (Figure 4). Most mainchain atoms, in contrast, are usually buried beneath layers of other atoms (usually sidechain atoms), precluding large torsion angle changes and hinge motions.

It is important to note that because most shear motions do, in fact, contain hinges, (joining the various sliding parts) the existence of a hinge is not the salient difference between the two basic mechanisms. Instead, it is the existence of a continuously maintained interface.

Other Classification

Most of the fragment and domain motions in the database fall within the hinge-shear classification. However, we have created additional categories to deal with the small number of exceptions.

Data Entry

One innovative feature of the database is that it allows authorized remote researchers to enter motions in their area of expertise directly into the database via a Web form. Authorization to edit a given motion entry, if necessary, works in conjunction with the standard password feature built into modern Web browser systems. The layout of the Web form is analogous to that of a normal HTML page describing a motion in the database, except that the various fields have been replaced by textboxes and pull-down selectors to make the Web page editable. The user retrieves either a blank form or a form corresponding to a pre-existing motion entry, makes appropriate changes remotely over the Internet via his or her Web browser, and then simply clicks the ‘Submit’ button to save changes into the database. Depending on whether or not the user has editing privileges over a particular motion entry, the changes may be published immediately or upon further approval by the database maintainers. The remote user may immediately preview the edited motion entry to see what it will look like once it becomes public.

The Web form system (Figure 5) takes advantage of advanced features of the Informix Dynamic Server with Universal Option to enable user previews. The Web Datablade module allows database content to be dynamically and rapidly translated into Web content with little additional overhead compared to static pages. Because updates to the database can be translated instantaneously into updated Web content, remote editors are able to preview their changes as it will appear to the end database user instantaneously before submitting or publishing them. Previously, we stored the database using the MSQL database software package, which is freely available to academic users. Unlike the commercial Informix system, the MSQL package does not support Application Program Interfaces (APIs) that allow for an efficient, rapid translation of database content into Web content. Consequently, it was necessary to store the Web interfaces as static HTML files on the server. For Web content to remain current, these pages would need to be rebuilt each time the database changed, a time-consuming process that would have prevented accurate previews. In addition, the Informix database system also features state-of-the-art transaction concurrency and logging, important features when multiple users are simultaneously updating the database.

In this way, the database takes full advantage of the cooperatively features of the Internet and modern database software, allowing experts in distant parts of the world to collaborate simultaneously on macromolecular motions. In addition to accelerating the rate at which the database may be populated, this feature improves the accuracy and timeliness of existing database entries by allowing them to be edited, revised, and updated, if necessary, by experts in the field.

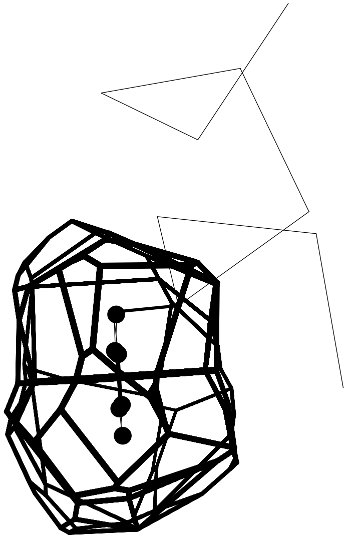


Figure 6. Voronoi Polyhedra. Two representative Voronoi polyhedra from 1CSE (subtilisin). On the left is shown the polyhedron around the sidechain hydroxyl oxygen (OG) of a serine. On right is shown the six polyhedra around the atoms in a Phe ring.

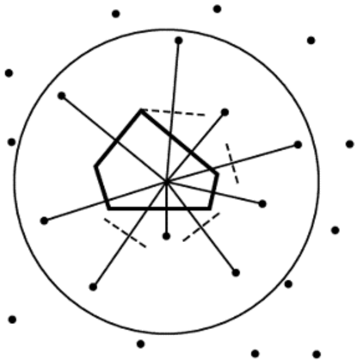


Figure 7. The Voronoi Polyhedra Construction. A schematic showing the construction of a Voronoi polyhedron in 2-dimensions. The asymmetry parameter is defined as the ratio of the distances between the central atom and the farthest and nearest vertex.

Internet Hits

The database is currently receiving over 65,000 hits from over 45,000 sites each month. Internet traffic on the database's main web server grew approximately exponentially between November, 1997, and February 1998, with database usage doubling approximately every other month during this period. In recent months, database usage has continued to grow, albeit at a somewhat reduced rate. We expect this trend to continue as the database becomes established in the structural biology community.

STANDARDIZED TOOLS FOR PROTEIN MOTIONS

Quantification of packing using Voronoi polyhedra

Packing clearly is an essential component of the motions 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." We have attempted to systematize and quantify the discussion of packing in the context of the motions database through the use of particular geometric constructions called Voronoi polyhedra and Delaunay triangulation.[53]

Voronoi polyhedra are a useful way of partitioning space amongst a collection of atoms. Each atom is surrounded by a single convex polyhedron and allocated the space within it (Figure 6). The faces of Voronoi polyhedra are formed by constructing dividing planes perpendicular to vectors connecting atoms, and the edges of the polyhedra result from the intersection of these planes.

Voronoi polyhedra were originally developed (obviously enough) by Voronoi[60] nearly a century ago. Bernal and Finney[61] used them to study the structure of liquids in the 1960s. 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[62] found a solution to this problem and first applied Voronoi polyhedra to proteins in 1974. He has, subsequently, reviewed their use in this application[48,49].

Voronoi polyhedra are particularly useful in studying the packing of the protein interior. This is because the construction of Voronoi polyhedra allocates all space amongst a 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[63]. 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 residues in proteins[46,64] a worthwhile proposition.

Voronoi polyhedra calculations have been applied to other aspects of packing in protein structure. In particular, they have been used to study protein-protein recognition[65], protein motions[53], and the protein surface[7,66-68]. As the Voronoi volume of an atom is a weighted average of the distances to all its neighbors (where the contact area with a neighbor is the weight), Voronoi polyhedra are very useful in assessing interatomic contacts[68-70]. Furthermore, the faces of Voronoi polyhedra have been used to characterize protein accessibility and to assess the fit of docked substrates in enzymes[71,72].

Voronoi polyhedra have many uses beyond the analysis of protein structures. For instance, they have also been used in the analysis of liquid simulations[73] and in weighting sequences to correct for over- or under-representation in an alignment[74]. In non-biological applications, they are used in “nearest-neighbor” problems (trying to find the neighbor of a query point) and in finding the largest empty circle in a collection of points[75]. The dual of a Voronoi diagram is a Delaunay triangulation. Since this triangulation has the “fattest” possible triangles, it is convenient for such procedures as finite element analysis. Furthermore, the border of Delaunay triangulation is the convex hull of an object, which is useful in graphics[75].

The simplest method for calculating volumes with Voronoi polyhedra 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. It is, furthermore, a useful approach for high-dimensional integration[74] and for the curved dividing surface approach discussed later.

More realistic 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 basic Voronoi construction (Figure 7), 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

[*] That is, one uses four sets of coordinates (x,y,z) to solve for the center (a,b,c) of the sphere: $(x-a)^2+(y-b)^2+(z-c)^2=r^2$. (This method can fail for certain pathological arrangements of atoms that would

four atoms than any other atom; if so, it is a vertex.

In the procedure outlined above, all the atoms are considered equal, and the dividing planes are positioned midway between atoms (Figure 6). 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 an excess to the smaller atom.

Two principal methods of re-positioning the dividing plane have been proposed to make the partition more physically reasonable: method B[62] and the radical-plane method[77]. Both methods depend on the radii of the atoms in contact (R_1 and R_2) and the distance between the atoms (D).

Representing Motion Pathways as "Morph Movies"

One of the most interesting of the complex data types kept in the database are "morph movies" giving a plausible representation for the pathway of the motion. These movies can immediately give the viewer an idea of whether the motion is a rigid-body displacement or involves significant internal deformations (as in tomato bushy stunt virus versus citrate synthase). Pathway movies were pioneered by Vorhein et al.[78], who used them to connect the many solved conformations of adenylate kinase.

Normal molecular-dynamics simulations (without special techniques, such as high temperature simulation or Brownian dynamics[79-81]) cannot approach the timescales of the large-scale motions in the database. Consequently a pathway movie cannot be generated directly via molecular simulation. Rather, it is constructed as an interpolation between known endpoints (usually two crystal structures). The interpolation can be done in a number of ways.

Straight Cartesian interpolation. The difference in each atomic coordinate (between the known endpoint structures) is simply divided into a number of evenly spaced steps, and intermediate structures are generated for each step. This was the method used by Vorhein et al. It is easy to do, only requiring that the beginning and ending structures be intelligently positioned by fitting on a motionless core. However, it produces intermediates with clearly distorted geometry.

Interpolation with restraints. This is the above method where each intermediate structure is restrained to have correct stereochemistry and/or valid packing. One simple approach is to minimize the energy of each intermediate (with only selected energy terms) using a molecular mechanics program, such as X-PLOR[82]. This technique will be described more fully in a forthcoming paper (Krebs & Gerstein, manuscript in preparation). The database, furthermore, is currently home to an experimental server that applies this interpolation technique to two arbitrary structures, generating a movie.

ANALYSIS OF AMINO ACID COMPOSITION OF LINKER SEQUENCES

Now that we have developed a database of protein motions, an essentially structure-orientated database, we want to use this to help interpret the mass of sequence data coming out of genome sequencing projects. In this way we are extrapolating ideas developed on the (relatively) smaller structure database to the much larger sequence database. We propose to do this through the calculation of two propensity scales for amino acids to be in linkers or flexible hinges.

not normally be encountered in a real protein structure; see Proacci and Scateni[76]).

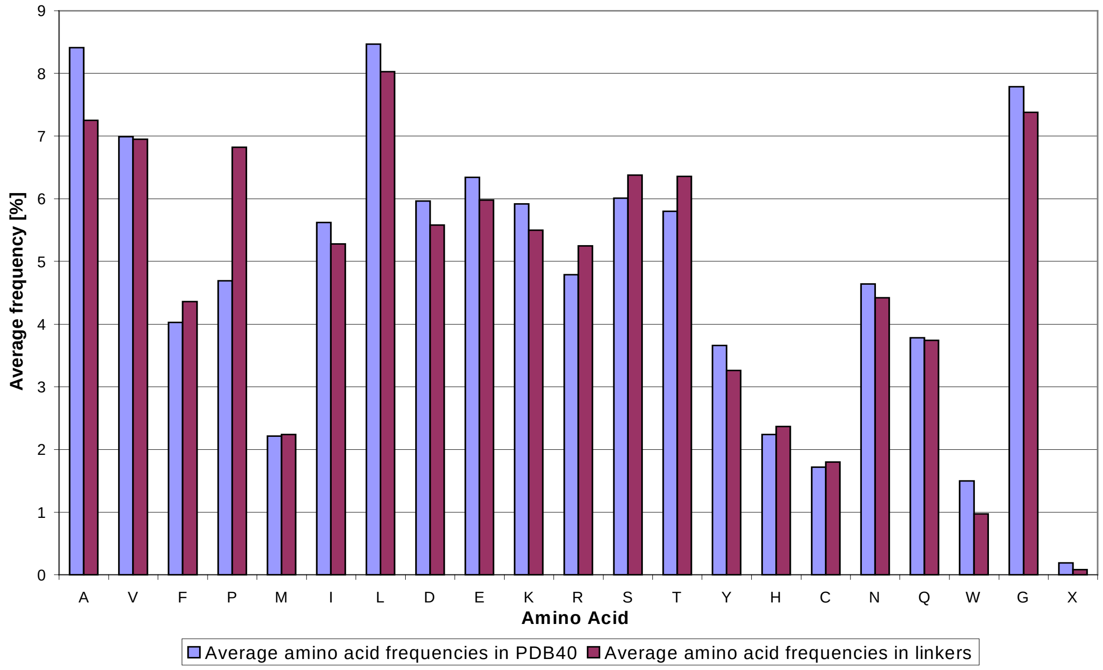


Figure 8. Comparison of the average amino acid composition in linker sequences and proteins in general (as represented by the PDB40 database).

Solved protein structures typically reveal different domains of proteins and linker regions between these domains. Linker regions are typically flexible, and, as such, form the basis for the hinge regions that allow two protein domains or fragments to move relative to each other as a part of a hinge mechanism.

Information about the amino acid composition of linker sequences can potentially be used to predict protein domains in protein sequences of unknown structure. In particular, a profile of flexible linker regions might be used to predict the location of domain hinges, for structural annotation of genome sequences.[93] Here we present some preliminary results involving two methods for statistical analysis of linker sequences.

Propensities for Linkers in General

Our first method of analysis of linker sequences includes both flexible as well as inflexible linkers. In this method we have arbitrarily defined a linker sequence as the 16 residue region centered around the peptide bond linking two domains.

The analysis of the amino acid composition of linker sequences is an example of deriving sequence information from structural information. The structural information (i.e., the location of protein domains) can be found in the Structural Classification of Proteins (SCOP)[19,20]. SCOP contains several databases of amino acid sequences of protein domains. In our study, the PDB40 database provided by SCOP has been used to create a database of linker sequences. The PDB40 database comprises a subset of proteins in the Protein Data Bank (PDB) with known structure 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. We were able to extract 234 linker sequences from the PDB40 database, although the PDB40 database itself contains about 1,500 protein sequences. This mainly reflects the fact that many proteins consist of only a single domain and therefore contain no linker region.

Figure 8 compares the average amino acid composition of the linker sequences with the average amino acid composition of the PDB40 database, while Table 3 shows in more detail the profile of the amino acid composition at each of the sixteen positions in the linker sequence. For an interpretation of these results it is important to compute two-sided P-values to determine which amino acids show statistically different frequencies in linkers than in the database as a whole. (A two-sided P-value represents the probability that, in a data set of equal size drawn at random from the PDB40 database, a given amino acid would have a frequency of occurrence as different as or more different from its occurrence in the entire PDB40 database than what was actually observed in the linker subset.) Figure 9 shows the P-values for the average amino acid composition in the linkers. We are able to conclude, 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. Table 4 shows the P-values of the amino acids at each of the sixteen linker positions.

In Table 4 and Figure 9 the amino acids have been roughly grouped according to the attributes hydrophobic, charged, and polar (following the classification of Branden and Tooze[83]). As shown in Table 4 and Figure 9, the frequencies of the remaining amino acids in linkers are not statistically different from the database as a whole at the 5% significance level.

The statistical significance of the results of the computed amino acid averages can be assessed by comparing the composition of the linker sequences with random data sets of

Table 3. 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.

																	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 4. 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[76]) does not provide a satisfactory explanation for the observed levels of amino acids (see also Figure 9).

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**	

sequences of the same length and the same amount taken from the PDB40 database. The number of times a single amino acid occurs in multiple random data sets follows the binomial distribution according to the familiar equation:

$$P^N(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Here, p is the probability that the amino acid occurs in the PDB40 database, and $P^n(k)$ is the probability that the amino acid occurs k times in a data set of n samples (n = 234 for the distribution of every *single* of the sixteen linker positions and n = 234 x 16 for the distribution of the linker *average*). The ratio k/n represents the fraction of the amino acid in the data set. Knowledge of the distribution functions of the amino acids then allows the calculation of P-values from the cumulative distribution function:

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

The value of $CDF^n(k)$ is the probability that the number of counts of an amino acid in a random data set would be less than k. Consequently, if o and e represent the observed and expected counts, then the two-sided P-value is given by $1\text{-}CDF^n(e+|o\text{-}e|) + CDF^n(e\text{-}|o\text{-}e|)$. This is simply the probability that the number of counts observed in a random subset of PDB40 would take on a value more different from what was expected than what was observed. In order to assign a P-value to an amino acid frequency in the linkers data set, the discrete values of the cumulative distribution function have been linearly interpolated. In

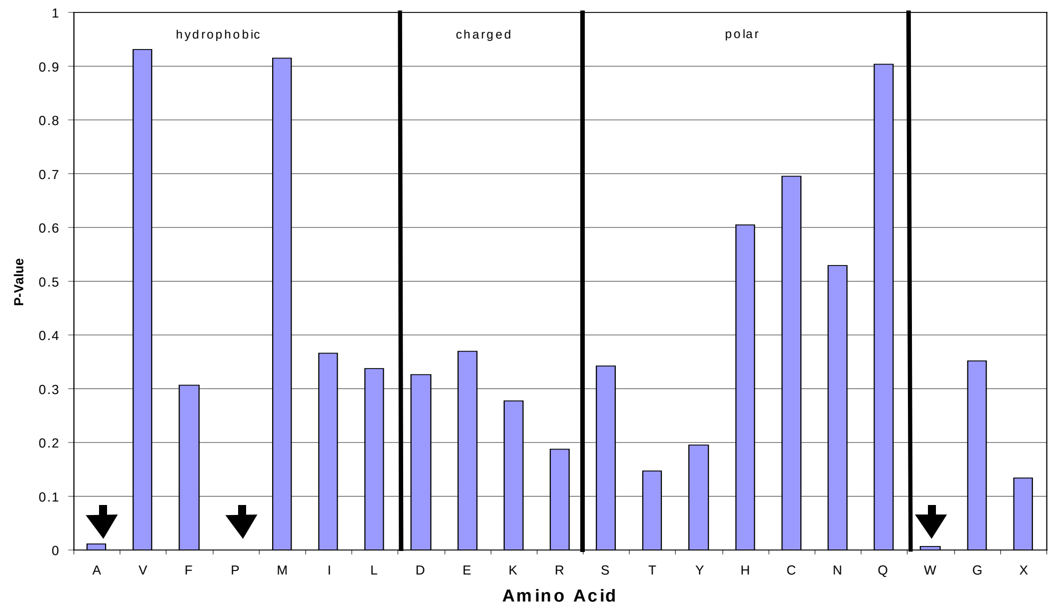


Figure 9. P-values for the average amino acid compositons in linker sequences. The P-values of alanine, proline, and tryptophan are close to zero. The difference between the content of these amino acids in linkers and protein sequences in general (as represented by the PDB40 database) is statistically significant at better than 98% confidence.

most cases, it is also possible to obtain a satisfactory approximation to the P-values by applying the two-sided significance test to the Normal approximation of the Binomial distribution.

Towards Propensities for Flexible Linkers

A variant on this procedure involves focusing just on linkers that are known to be flexible. Our Database of Macromolecular Motions contains residue selections for known protein hinge regions (i.e., flexible linkers) that have been culled from the scientific literature. These sequences have been verified manually to be true flexible linker regions, and thus this database constitutes 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 may be obtained. A FASTA search with a suitable cutoff (e.g., e-value 0.001) may then be performed on known linker sequence to obtain a series of near homologues (Table 6).These homologues can then be arranged into a multiple alignment (via the CLUSTALW) program[84,85] and the multiple alignment can be fused into a variety of consensus pattern representations, such as Hidden Markov Models or simply consensus sequences[86-90]. A sample multiple alignment for the hinge in calmodulin is shown in Table 6 and a number of consensus sequences are shown in Table 5. The amino acid composition may be averaged over all the different hinges and different positions within a hinge to give a single composition vector for flexible hinges. Finally, this can be compared to the overall amino acid composition or that of linkers to obtain a preliminary scale of amino acid propensity in mobile linkers, as shown in Table 7. This can be compared with the scale of amino acid propensities in linkers as obtained by the procedure previously described and shown in Table 3.

Table 5. 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

CONCLUSION AND FUTURE DIRECTIONS

We have developed a number of database-based techniques for the study of macromolecular motions. We have constructed a database of macromolecular motions, which currently documents ~120 motions, and have developed a classification scheme for the database based on size then packing (whether or not there is motion across a well-packed interface). The database incorporates innovative cooperatively features, allowing authorized remote experts to act as database editors via the Internet. We also developed a standardized nomenclature, such as maximum atomic displacement or degrees of rotation. We are developing automated tools to analyze protein and nucleic acid structures and sequences with possible motions, to extract standardized statistics on macromolecular motions from structural data, and allow the database to be more readily populated.

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[91,92]. 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 an (over-simplified!) 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.

Table 6. 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	Sequence
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 7. 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

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the financial support of the National Science Foundation (Grant DBI-9723182) and the numerous people who have either contributed entries or information to the database 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.

All correspondence to Mark.Gerstein@yale.edu.

REFERENCES

1. N. Wade, *Scientists Find A Key Weapon Used by H.I.V.*, in *New York Times*. 1997: New York. p. A1.
2. D.G. Donne, *et al.*, *Proc Natl Acad Sci USA*. **94**:13452–13457 (1997).
3. D.C. Chan, *et al.*, *Cell*. **89**:(2):263-73 (1997).
4. D. Peretz, *et al.*, *J Mol Biol*. **273**:(3):614-22 (1997).
5. P.M. Harrison, *et al.*, *Curr Opin Struct Biol*. **7**:(1):53-9 (1997).
6. M. Gerstein and W. Krebs, *Nucl Acids Res* (In press) (1998).
7. M. Gerstein and C. Chothia, *Proc Natl Acad Sci USA*. **93**:10167-10172 (1996).
8. A. Bairoch and B. Boeckmann, *Nucl Acids Res*. **20**:2019-2022 (1992).
9. L. Holm and C. Sander, *Nuc Acid Res*. **22**:3600-3609 (1994).
10. G.D. Schuler, *et al.*, *Meth Enz*. **266**:141-162 (1996).
11. E. Abola, *et al.*, *Meth Enz*. **277**:556-571 (1997).
12. C.A. Orengo, D.T. Jones, and J.M. Thornton, *Nature*. **372**:631-634 (1994).
13. R.B. Altman, N.F. Abernethy, and R.O. Chen, *Ismb*. **5**:15-24 (1997).
14. R.O. Chen, R. Felciano, and R.B. Altman, *Ismb*. **5**:84-7 (1997).
15. A. Bairoch, P. Bucher, and K. Hofmann, *Nucleic Acids Research*. **24**:(1):189-196 (1996).
16. H.M. Berman, *et al.*, *Biophys J*. **63**:(3):751-759 (1992).
17. J.A. Epstein, J.A. Kans, and G.D. Schuler, *2nd Ann Int WWW Conf*. :(in press) (1994).
18. C.W. Hogue, H. Ohkawa, and S.H. Bryant, *Trends Biochem Sci*. **21**:(6):226-9 (1996).
19. A. Murzin, *et al.*, *J Mol Biol*. **247**:536-540 (1995).
20. T.J.P. Hubbard, *et al.*, *Nucleic Acids Res*. **25**:(1):236-9 (1997).
21. W.G. Scott, J.T. Finch, and A. Klug, *Cell*. **81**:(7):991-1002 (1995).
22. H.W. Pley, K.M. Flaherty, and D.B. McKay, *Nature*. **372**:(6501):68-74 (1994).
23. J.H. Cate, *et al.*, *Science*. **273**:(5282):1678-85 (1996).
24. B. Rees, J. Cavarelli, and D. Moras, *Biochimie*. **78**:(7):624-31 (1996).
25. M. Ruff, *et al.*, *Science*. **252**:(5013):1682-9 (1991).
26. S. Remington, G. Wiegand, and R. Huber, *J Mol Biol*. **158**:111-152 (1982).
27. W.S. Bennett, Jr and T.A. Steitz, *Proc Natl Acad Sci USA*. **75**:4848-4852 (1978).
28. M. Gerstein, A.M. Lesk, and C. Chothia, *Biochemistry*. **33**:6739-6749 (1994).
29. W.S. Bennett and R. Huber, *Crit Rev Biochem*. **15**:291-384 (1984).
30. J. Janin and S. Wodak, *Prog Biophys Mol Biol*. **42**:21-78 (1983).
31. C. Abad-Zapatero, *et al.*, *J Mol Biol*. **198**:445-67 (1987).
32. R.K. Wierenga, *et al.*, *Proteins*. **10**:93 (1991).
33. C. Chothia, *et al.*, *Nature*. **302**:500-505 (1983).
34. D.E. Koshland, *Sci Am*. **229**:52-64 (1973).
35. D.E. Koshland, Jr, *Proc Natl Acad Sci USA*. **44**:98-104 (1958).
36. C.M. Anderson, F.H. Zucker, and T. Steitz, *Science*. **204**:375-380 (1979).
37. J.R. Knowles, *Nature*. **350**:121-4 (1991).
38. L. Stryer. *Biochemistry*. 4th ed, W H Freeman and Company, New York (1995).
39. N.S. Sampson and J.R. Knowles, *Biochemistry*. **31**:8482-8487 (1992a).
40. J.R. Knowles, *Phil Trans R Soc Lond B*. **332**:115-121 (1991).
41. M. Perutz, *Quart Rev Biophys*. **22**:139-236 (1989).
42. P.R. Evans, *Curr Opin Struc Biol*. **1**:773-779 (1991).
43. G. Fermi and M.F. Perutz. *Haemoglobin and Myoglobin*, Claredon Press, Oxford (1981).
44. L.N. Johnson and D. Barford, *J Biol Chem*. **265**:2409-2412 (1990).
45. F.M. Richards and W.A. Lim, *Quart Rev Biophys*. **26**:423-498 (1994).
46. Y. Harpaz, M. Gerstein, and C. Chothia, *Structure*. **2**:641-649 (1994).

47. M. Levitt, *et al.*, *Ann Rev Biochem*. **66**:549-579 (1997).
48. F.M. Richards, *Methods in Enzymology*. **115**:440-464 (1985).
49. F.M. Richards, *Ann Rev Biophys Bioeng*. **6**:151-76 (1977).
50. L.M. Gregoret and F.E. Cohen, *J Mol Biol*. **211**:(4):959-974 (1990).
51. S.J. Hubbard and P. Argos, *Protein Science*. **3**:(12):2194-2206 (1994).
52. S.J. Hubbard and P. Argos, *J Mol Biol*. **261**:289-300 (1996).
53. M. Gerstein, *et al.*, *J Mol Biol*. **234**:357-372 (1993).
54. J.W. Ponder and F.M. Richards, *J Mol Biol*. **193**:775-791 (1987).
55. C.L. Lawson, *et al.*, *Proteins*. **3**:18-31 (1988).
56. C.A. McPhalen, *et al.*, *J Mol Biol*. **227**:197-213 (1992).
57. A.J. Olson, G. Bricogne, and S.C. Harrison, *J Mol Biol*. **171**:61 (1983).
58. B.F. Anderson, *et al.*, *Nature*. **344**:784-787 (1990).
59. M. Gerstein and C.H. Chothia, *J Mol Biol*. **220**:133-149 (1991).
60. G.F. Voronoi, *J Reine Angew Math*. **134**:198-287 (1908).
61. J.D. Bernal and J.L. Finney, *Disc Faraday Soc*. **43**:62-69 (1967).
62. F.M. Richards, *J Mol Biol*. **82**:1-14 (1974).
63. G.J. Kleywegt and T.A. Jones, *Acta Cryst*. **D50**:178-185 (1994).
64. C. Chothia, *Nature*. **254**:304-308 (1975).
65. J. Janin and C. Chothia, *J Biol Chem*. **265**:16027-16030 (1990).
66. J.L. Finney, *J Mol Biol*. **96**:721-732 (1975).
67. J.L. Finney, *et al.*, *Biophys J*. **32**:(1):17-33 (1980).
68. M. Gerstein, J. Tsai, and M. Levitt, *J Mol Biol*. **249**:955-966 (1995).
69. J. Tsai, M. Gerstein, and M. Levitt, *J Chem Phys*. **104**:9417-9430 (1996).
70. J. Tsai, M. Gerstein, and M. Levitt, *Protein Science*. :(in press) (1997).
71. J.L. Finney, *J Mol Biol*. **119**:415-441 (1978).
72. C.W. David, *Biopolymers*. **27**:339-344 (1988).
73. J.P. Shih, S.Y. Sheu, and C.Y. Mou, *J Chem Phys*. **100**:(3):2202-2212 (1994).
74. P.R. Sibbald and P. Argos, *J Mol Biol*. **216**:813-818 (1990).
75. J. O'Rourke. *Computational Geometry in C*, Cambridge UP, Cambridge (1994).
76. P. Procacci and R. Scateni, *Int J Quant Chem*. **42**:151-1528 (1992).
77. B.J. Gellatly and J.L. Finney, *J Mol Biol*. **161**:305-322 (1982).
78. C. Vonrhein, G.J. Schlauderer, and G.E. Schulz, *Structure*. **3**:483-490 (1995).
79. D. Joseph, G.A. Petsko, and M. Karplus, *Science*. **249**:1425-1428 (1990).
80. R.C. Wade, *et al.*, *Biophys J*. **64**:9-15 (1993).
81. J.A. McCammon and S.C. Harvey. *Dynamics of Proteins and Nucleic Acids*, Cambridge UP, (1987).
82. A.T. Brünger. *X-PLOR 3.1, A System for X-ray Crystallography and NMR*, Yale University Press, New Haven (1993).
83. C. Branden and J. Tooze. *Introduction to Protein Structure*, Garland Publishing Incorporated, New York (1991).
84. J.D. Thompson, D.G. Higgins, and T.J. Gibson, *Nuc Acid Res*. **22**:4673-4680 (1994).
85. D.G. Higgins, J.D. Thompson, and T.J. Gibson, *Methods Enzymol*. **266**:383-402 (1996).
86. E.L. Sonnhammer, *et al.*, *Nucleic Acids Res*. **26**:(1):320-2 (1998).
87. A. Krogh, *et al.*, *J Mol Biol*. **235**:1501-1531 (1994).
88. S.R. Eddy, *Curr Opin Struc Biol*. **6**:361-365 (1996).
89. S.R. Eddy, G. Mitchison, and R. Durbin, *J Comp Bio*. **9**:9-23 (1994).
90. P. Baldi, Y. Chauvin, and T. Hunkapiller, *Proc Natl Acad Sci*. **91**:(1059-1063) (1994).
91. S.E. Brenner, C. Chothia, and T.J. Hubbard, *Curr Opin Struct Biol*. **7**:(3):369-76 (1997).
92. C. Chothia, *Nature*. **357**:543-544 (1992).
93. M. Gerstein, *J Mol Biol*. **274**: 562-576 (1997).