D296–D301 *Nucleic Acids Research, 2006, Vol. 34, Database issue*
doi:10.1093/nar/gkj046

The Database of Macromolecular Motions: new features added at the decade mark

Samuel Flores[1,3], Nathaniel Echols[2], Duncan Milburn[3], Brandon Hespenheide[6], Kevin Keating[4], Jason Lu[3], Stephen Wells[6], Eric Z. Yu[3], Michael Thorpe[6,7] and Mark Gerstein[3,4,5,*]

[1]Department of Physics, Yale University, P.O. Box 208120, New Haven, CT 06520-8120, USA, [2]Department of Molecular and Cell Biology, University of California, Berkeley, CA 94720, USA, [3]Molecular Biophysics and Biochemistry Department and [4]Computational Biology and Bioinformatics Program and [5]Department of Computer Science, Bass 432A, 266 Whitney Avenue, Yale University, New Haven, CT 06520, USA, [6]Center for Biological Physics, Department of Physics and Astronomy and [7]Department of Chemistry and Biochemistry, Arizona State University, Tempe, AZ 85287, USA

Received August 15, 2005; Revised and Accepted October 4, 2005

ABSTRACT

The database of molecular motions, MolMovDB (http://molmovdb.org), has been in existence for the past decade. It classifies macromolecular motions and provides tools to interpolate between two conformations (the Morph Server) and predict possible motions in a single structure. In 2005, we expanded the services offered on MolMovDB. In particular, we further developed the Morph Server to produce improved interpolations between two submitted structures. We added support for multiple chains to the original adiabatic mapping interpolation, allowing the analysis of subunit motions. We also added the option of using FRODA interpolation, which allows for more complex pathways, potentially overcoming steric barriers. We added an interface to a hinge prediction service, which acts on single structures and predicts likely residue points for flexibility. We developed tools to relate such points of flexibility in a structure to particular key residue positions, i.e. active sites or highly conserved positions. Lastly, we began relating our motion classification scheme to function using descriptions from the Gene Ontology Consortium.

INTRODUCTION

The study of macromolecular motions is important for the understanding of function. Motion is crucial for the mechanism of catalysis, signaling and for the formation of complexes. Also, knowledge of the accessible conformations can be used to improve the performance of docking codes. For these reasons a server which receives pairs of structures and generates putative motion trajectories plays a unique role in structural biology. The Database of Molecular Motions (1–5) is not only a repository of such motions but also aims to characterize them systematically and provide tools for their analysis.

MolMovDB is a resource for studying conformational changes in protein and other macromolecules, primarily through analysis of crystal structures. It has been used to design and test a wide variety of structural analysis algorithms. The Morph Server in particular has been used by many scientists to analyze pairs of conformations and produce realistic animations.

MolMovDB sits within a constellation of databases focusing on protein structure. These include the Structural Classification of Proteins (6), the Protein Data Bank (PDB) (7), CATH (8) and many others. Most of these databases are designed as repositories of information or systems of classification for single protein structures. MolMovDB differs from most in that it focuses on motions.

Early studies of domain movements based on comparison of two structures (9,10) led to the idea of creating a database of pairs of structures. Initially a simple collection of web pages (10), MolMovDB soon developed into a proper database with a classification scheme (1,2,11). An automatic pipeline for finding and morphing related proteins in the PDB followed (3). Updates in recent years have included a normal mode analysis server to try to predict probable motions from a single structure (3), and automated graphs showing distribution of flexibility statistics (4).

*To whom correspondence should be addressed. Tel: +1 203 432 6105; Fax: +1 360 838 7861; Email: Mark.Gerstein@yale.edu.
Correspondence may also be addressed to Samuel Flores. Tel: +1 203 747 2682; Email: Samuel.Flores@yale.edu

In the present work we describe recent improvements to MolMovDB. We have begun relating our motion classification scheme to function classification using definitions provided by GO (12). New tools have been added to relate motions to particular sites, namely active sites and highly conserved residues. We have further developed the morph server to produce more realistic interpolations between two structures and handle larger motions. Specifically, an option has been added to use FRODA (13) to find a sterically allowed trajectory, and a multiple chain option has been made available to obtain the trajectory of a complex using adiabatic mapping. We have also added an interface to our flexibility prediction program, FlexOracle (S. Flores *et al*., submitted).

IMPROVEMENTS TO THE MORPH SERVER

The original morph server uses an adiabatic mapping approach to generate morphs for single chains. Under this scheme, the distance between each atom on one structure and the corresponding atom on a second structure is evenly reduced. After each reduction, the thus interpolated structure is subjected to an energy minimization step. The usefulness of the trajectories generated has been limited by two factors. First, previous versions of the server could not handle complexes. This limitation stood in the way of large scale studies of interior voids in proteins, helix–helix packing, flexibility prediction and many other large-scale structural analyses. Second, we found that when the trajectory of conformational change strayed far from a linear interpolation, the morph server often gave unphysical results.

To address these limitations, we added two new morph methods. First, a new multiple chain option enables the morphing of complexes using adiabatic mapping (11). Second, a new FRODA option gives our morph server the ability to circumvent the steric clashes that sometimes occur in adiabatic mapping.

Multiple chain morphing using adiabatic mapping approach

The new multiple chain option of the Morph Server was developed to aid in studies of conformational changes of large complexes, including mixed protein–nucleic acid structures.

Although interpolations of complexes were possible with previous versions of the server, they were limited by the requirement for precisely matching sequences and limited gaps. Proteins with very distant homology could be morphed, but only single chains at a time. Our new server can determine consensus sequences and coordinates for an arbitrarily large number of chains, and has successfully been used with structures related by 55% sequence identity. Compared with either of the previous versions, the output is more faithful to the original crystallographic data, preserving most atomic positions, residue numbering and gaps. Currently the processing of homologous structures results in alanine mutations for residue mismatches, but the design is sufficiently flexible to allow addition of other methods for obtaining a consensus sequence.

New FRODA option for the morph server

To address the potential clashes inherent in non-linear trajectories, we used the newly developed FRODA (13) module that is part of the FIRST5 software suite. Our submissions page now offers a 'FRODA lite' (14) option, which invokes a 'directed dynamics' FRODA run with a set of default parameters. In this 'lite' mode, only covalent bond lengths and angles are maintained, along with appropriate van der Waals radii on all atoms to avoid collisions (13). Since no hydrogen-bond constraints are considered in the 'lite' mode, there is no need for the input structure to have hydrogens added. A full-feature version of FIRST5 with FRODA is available for download and online usage at http://flexweb.asu.edu, which also includes added hydrogen atoms, as well as hydrogen bond constraints and hydrophobic tethers (13).

The central concept behind FRODA is the use of geometric simulation to explore conformational space. The simulation begins by mapping a set of ghost templates onto the protein such that every atom belongs to at least one template. These templates overlap each other only at rotatable dihedral angles. Figure 1a and b, respectively, shows the atoms and the two ghost templates, colored yellow and blue, FRODA assigns to an ethane molecule. The two carbon atoms belong to both ghost templates because the carbon–carbon bond is rotatable (flexible). Initially, these templates map onto the structure perfectly, with each edge of a template mapping to a covalent bond in protein. Noncovalent interactions are not included in the morphing procedure of FRODA lite.

Once the ghost templates have been mapped to the protein atoms, the simulation proceeds through a series of displacement and matching steps. The result of each step is a new conformation. One of these simulation steps is depicted for an ethane molecule, shown with green carbon spheres and white hydrogen spheres in Figure 1a. Figure 1b shows two ghost templates, yellow and blue, mapped onto the ethane molecule such that each hydrogen atom is associated with a single vertex while each carbon atom is associated with two vertices, one in each ghost template. Each step begins with random displacement of every atomic position (Figure 1c), essentially breaking all of the bonds. Now begins an iterative procedure to realign the atoms and their associated template(s). First, the ghost templates are fit as best as possible to the new positions of the atoms (Figure 1d). The position and orientation of each ghost template is computed by a least-squares fit to the new positions of the atoms. The displaced atoms are then fit onto the new position of the ghost templates (Figure 1e). The hydrogen atoms fit exactly onto their respective ghost template positions because they each belong to only one template. The carbon atoms, however, each belong to two templates, and are thus positioned equidistant from each of their associated ghost template points. This concludes one iteration of atom-ghost template fitting. A predefined fitting tolerance determines whether the templates have been adequately realigned, thus concluding one step in FRODA. The templates in Figure 1e are outside of the fitting tolerance, and so iteration of the ghost template-atom fitting continues. Figure 1f and g, respectively, shows refitting of the templates to the atoms, and the atoms back onto the templates. It can be seen that the second iteration, Figure 1g, results in templates that are much better aligned compared with the result of the first iteration in Figure 1e. The procedure continues until the atoms and the ghost templates are aligned within tolerance. Alignment is measured by the distance between atoms and vertices. The tolerance in FRODA lite is $\leqslant$0.125 Å. One complete step

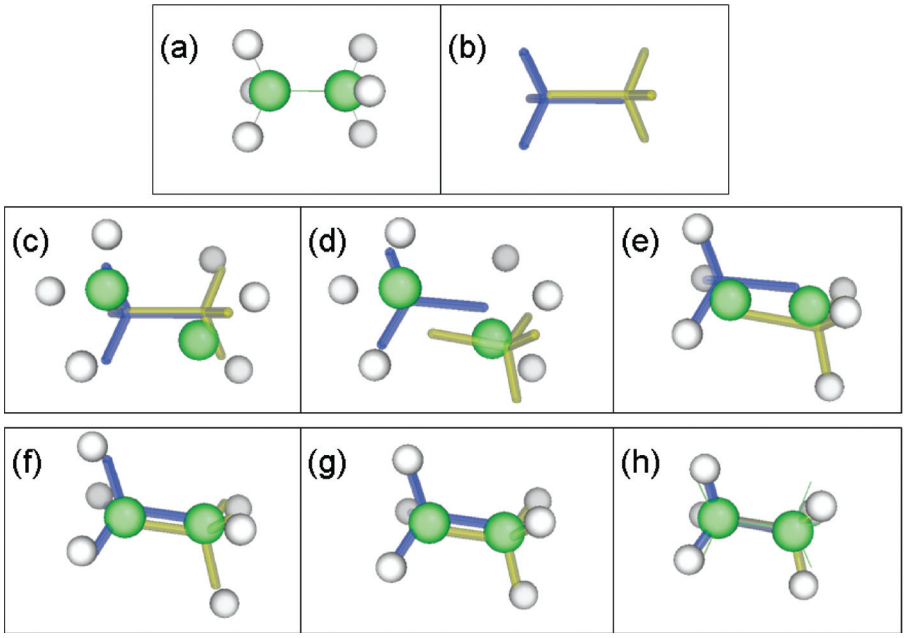

Figure 1. The motion of an ethane molecule as determined by geometric simulation in FRODA. (**a**) Initial atomic positions; (**b**) ghost templates; (**c**) random atomic displacement; (**d**) fitting of ghost templates to atoms; (**e**) refitting of atoms to ghost templates; (**f** and **g**) further iterations of (d and e); (**h**) until a new conformer is found[13].

of FRODA produces a new conformation of the ethane molecule (Figure 1h).

The morphing procedure used by FRODA lite is directed from an initial structure to a target and so differs slightly from the example given in Figure 1. During morphing, the initial random displacement of the atoms at the beginning of each step is now biased to move the atoms toward their respective position in the target structure. The result is a gradual transition from the initial structure to the target structure. It is important to note that steric overlaps are computed during each iteration so that the atoms move during fitting both to match the ghost templates and to obey excluded-volume constraints.

If the structure finds itself in a jammed position, such that the tolerances cannot be satisfied, it will revert to a previous conformer and continue the morph. The random element in the atomic displacements provides for a degree of simulated annealing so that the structure can find its way around small obstacles. The paths produced by FRODA therefore avoid sterically impossible trajectories.

INTERFACE TO FLEXIBILITY ANALYSIS TOOL

Ultimately, most studies of flexibility are oriented towards the goal of predicting specific conformational changes. To this end we built an interface to our FlexOracle flexibility analysis tool, accessible from the front page of MolMovDB. The goal of this server is to provide hinge predictions for structures submitted by the public. FlexOracle and the Normal Mode Analysis server (3) differ from the morph server in that the first two operate on a single structure submitted by the user, and the third operates on a pair of structures.

The hinge-prediction server submission form is linked to from the front page of MolMovDB. Users are invited to submit a single PDB (7) file containing a single chain. Upon submission, FlexOracle is run on the structure. The user is sent an email with the URL at which the results may be viewed.

In the FlexOracle hinge prediction algorithm (S. Flores *et al.*, submitted for publication), a cut is introduced into the structure after a residue i. The resulting N-terminal fragment with residues 1 to $i - 1$ is separated from the C-terminal fragment with residues i to N. The intra-molecular potential energy of each fragment is calculated using CHARMm (15). The implicit solvent model is used to account for the protein–solvent interactions. The energies corresponding to the two fragments are summed. The process is repeated for $i = 2$ to N. The procedure is similar to that used by Janin and Wodak (9) in their solvent exposed area calculations. Continuing that comparison, values of i that result in lower energy correspond to residues more likely to be in hinges. As implemented on our server, the predictor only works when the submitted chain represents the biological molecule (i.e. does not occur in complex), and is soluble. For these cases the predictions compare well with known hinges.

For proteins that have had FlexOracle run on a submitted structure, we link to a graph of energy versus i (Figure 2f). In order to compare the predicted with actual hinges, we have prepared a small set of morphs that had FlexOracle run on the first of the two submitted structures. These can be viewed at molmovdb.org/sets/curatedFlexOracle.

IMPROVEMENTS TO THE UNDERLYING CLASSIFICATION

In addition to improving the two-structure and single-structure servers, we also made improvements to the underlying classification in MolMovDB. These improvements have been

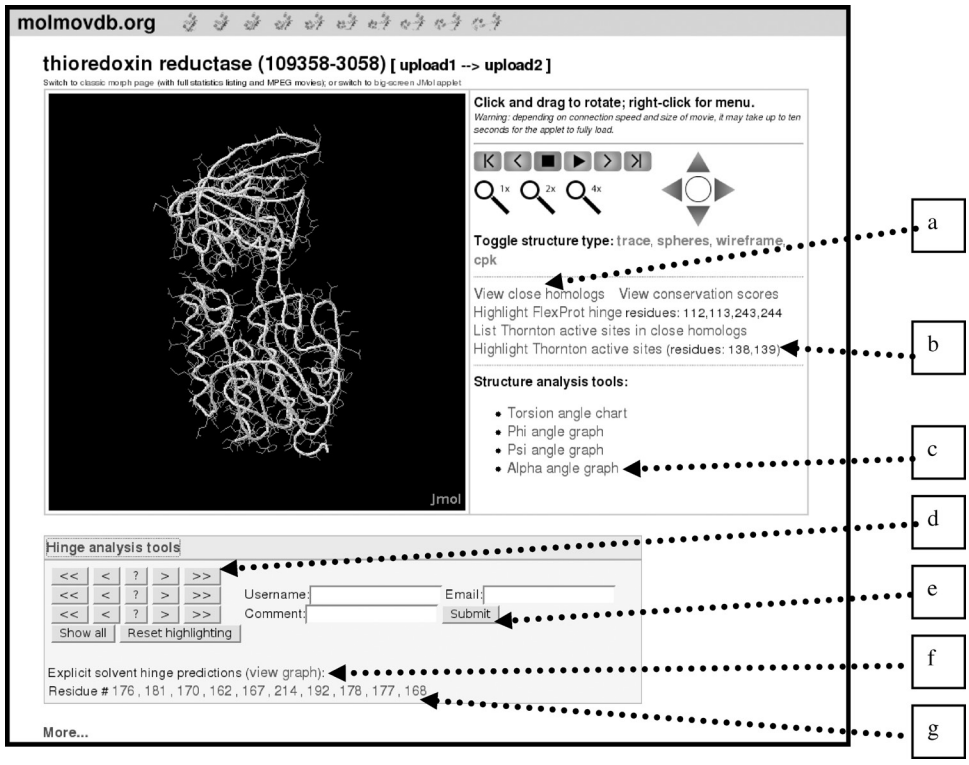

Figure 2. The new morph page. The previous page, now called morph-classic.cgi, can still be accessed by a link. Features: (**a**) a page with links to PDB entries with >99% sequence homology; (**b**) highlight active sites from the CSA database, if entries exist in any close homologs; (**c**) Torsion angle plots can be useful in guiding your hinge selection efforts; (**d**) if you wish to contribute to our hinge research, use the arrow buttons to manually select up to three hinges by visual inspection; (**e**) Submit your hinge selection, plus any comments. Comments appear in our public bulletin board; (**f**) if our FlexOracle hinge prediction program has been run on the first frame of the morph, the energy versus residue number plot can be viewed here; (**g**) the 10 best (lowest energy, in ascending order of energy) hinges can be highlighted in the viewer.

oriented towards relating structure to function and allowing us to group related morphs together with their homologs.

GO annotation

We have integrated a subset of the dataset from the EBI Gene Ontology Annotation (GOA) project (12), into the server. In particular, we have implemented the lookup and display of GO terms for PDB identifiers that feature in a given motion or morph. Terms from each of the three GO organizing principles—molecular function, cellular component and biological process—are displayed when available (at the time of writing, there were 191 040 references to 24 703 PDB structures), and links are provided to reveal the definition of individual terms. For example, for the motion in DNA polymerase I from *Thermus aquaticus* (database motion 'taq-pol', PDB codes 2ktq and 3ktq), the motion report can be annotated with six GO terms: DNA binding, nuclease activity, 5′–3′ exonuclease activity, DNA-directed DNA polymerase activity (all molecular function), intracellular (cellular component) and DNA replication (biological process)—all appropriate terms for this enzyme.

The addition of the GOA dataset for the annotation of the motion and morph reports is not only useful in itself, as the addition of the GO terms also facilitates searching the database in a broader fashion than was possible previously. Furthermore, we have added some interesting new subsets for use with the automatic plotter[3] derived from searches with GO terms. Figure 3 shows the distribution of one particular statistic, maximum Cα displacement of the second (moving) core of the structure for two subsets derived in this manner, and a third included for comparative purposes, showing the distribution for the same statistic across all canonical morphs in the database. These plots are a first step in addressing the question of whether particular types of motion are associated with a particular function or role in a biological system.

PDB ID VERSUS MORPH ID BLAST FEATURES

The front page of MolMovDB has long provided a feature to search by PDB ID. This feature was limited by the fact that many users uploaded structure files directly rather than providing PDB identifier. Also, the submitted structures were often unpublished and therefore absent from the PDB altogether. For these morphs, no PDB ID was assigned, therefore it was impossible to find them with this search method.

In order to overcome these shortcomings, it was necessary to assign a PDB ID to such morphs. To do so, we searched the PDB for structures with >99% sequence identity to the morphs in our database. This information was then used to provide an additional search option on our front page. It is also possible to use this feature to connect to our database by providing a PDB ID, as has been done on LinkHub (A. Smith *et al*., manuscript in preparation). Conversely, it is possible to search for PDB ID's with high sequence identity to a given morph

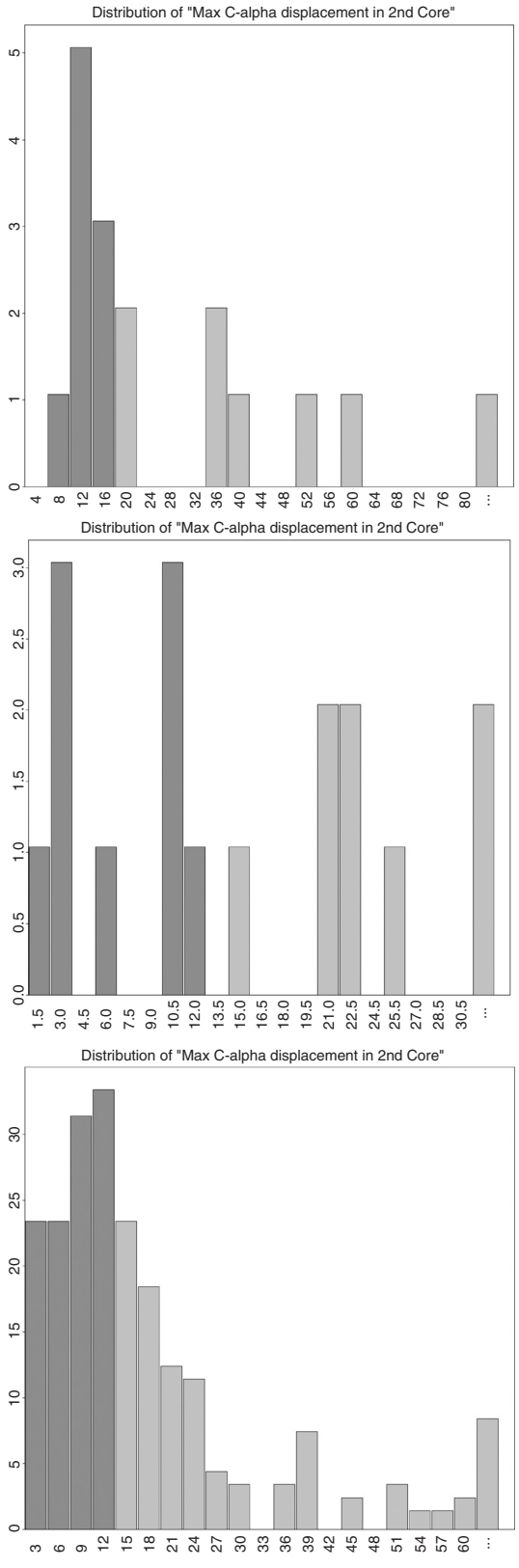

Figure 3. Example plots of the distribution of maximum Cα displacement of the mobile component (second core) in structures annotated with GO terms ('DNA binding', 17 morphs—top; 'metabolism',17 morphs—centre), compared with the reference morph dataset (200 canonical morphs—bottom). Dark bars indicate morphs whose second core max. Cα displacement falls below the median, while light bars indicate those above.

(Figure 2, a). As an additional benefit, it is possible to transfer annotation from closely related PDB structures to the corresponding morph. We took advantage of this capability to assign active site annotation to our morphs from the Catalytic Site Atlas (CSA) (19).

TOOLS RELATING MOTION TO SITES ON STRUCTURE

Users often want to relate the motion of particular residues in a structure to particular structural sites and features, e.g. the position of active sites or highly conserved residues. We built tools that now allow us to do both of these things.

Catalytic site highlighting

We related our morphs to active sites identified in the CSA (19). The CSA is manually curated, thus avoiding the various pitfalls of using the PDB's SITE records.

We obtained the active site residue numbers and corresponding PDB ID's from a table provided by the Thornton group. We annotated all the morphs that were linked by our homology table to an entry in the CSA. When annotation is available, a button appears on the morph page to highlight the active site residues. This can serve as a visual aid to understanding the link between catalysis and motion.

For easier browsing of this feature we've added a new gallery (2) named 'Catalytic Site Atlas'. Every morph in this gallery has active site information available for viewing.

CONSERVATION SCORE VIEWER

Since highly conserved residues are more probable to play an important role in the function of a protein, we implemented a tool to calculate the conservation score for each residue of a submitted sequence. The server highlights the top 5% of the most conserved residues (i.e. the residues with the highest conservation scores).

To calculate these conservation scores we first performed a BLAST search of the input sequence against nrdb90, a non-redundant sequence database in which protein sequences have no >90% sequence identity with each other (20). Next we extracted up to 50 top-aligned sequences to a given morph to generate an multiple sequence alignment using Clustal W (21). For each position in the multiple sequence alignment, we used information content to evaluate the consensus of each of the 20 types of amino acids at this position (22). Then we ranked each position according to the magnitude of the information content.

ACKNOWLEDGEMENTS

The work at Arizona State University was supported by the NSF, NIH and the Arizona State University Foundation.

Conflict of interest statement. None declared.

REFERENCES

1. Gerstein,M. and Krebs,W. (1998) A database of macromolecular motions. *Nucleic Acids Res.*, **26**, 4280–4290.

2. Krebs,W.G. and Gerstein,M. (2000) The morph server: a standardized system for analyzing and visualizing macromolecular motions in a database framework. *Nucleic Acids Res*., **28**, 1665–1675.
3. Krebs,W.G., Alexandrov,V., Wilson,C.A., Echols,N., Yu,H. and Gerstein,M. (2002) Normal mode analysis of macromolecular motions in a database framework: developing mode concentration as a useful classifying statistic. *Proteins*, **48**, 682–695.
4. Krebs,W.G., Tsai,J., Alexandrov,V., Junker,J., Jansen,R. and Gerstein,M. (2003) Tools and databases to analyze protein flexibility; approaches to mapping implied features onto sequences. *Methods Enzymol*., **374**, 544–584.
5. Qian,J., Stenger,B., Wilson,C.A., Lin,J., Jansen,R., Teichmann,S.A., Park,J., Krebs,W.G., Yu,H., Alexandrov,V. *et al.* (2001) PartsList: a web-based system for dynamically ranking protein folds based on disparate attributes, including whole-genome expression and interaction information. *Nucleic Acids Res*., **29**, 1750–1764.
6. Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol*., **247**, 536–540.
7. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res*., **28**, 235–242.
8. Pearl,F., Todd,A., Sillitoe,I., Dibley,M., Redfern,O., Lewis,T., Bennett,C., Marsden,R., Grant,A., Lee,D. *et al.* (2005) The CATH Domain Structure Database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis. *Nucleic Acids Res*., **33**, D247–D251.
9. Janin,J. and Wodak,S.J. (1983) Structural domains in proteins and their role in the dynamics of protein function. *Prog. Biophys. Mol. Biol*., **42**, 21–78.
10. Gerstein,M., Lesk,A.M. and Chothia,C. (1994) Structural mechanisms for domain movements in proteins. *Biochemistry*, **33**, 6739–6749.
11. M Gerstein,R.J., Johnson,T., Tsai,J. and Krebs,W. (1999) Studying macromolecular motions in a database framework: from structure to sequence. *Rigidity Theory Appl*., 401–442.
12. Camon,E., Magrane,M., Barrell,D., Lee,V., Dimmer,E., Maslen,J., Binns,D., Harte,N., Lopez,R. and Apweiler,R. (2004) The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Res*., **32**, D262–D266.
13. Wells,S., Menor,S., Hespenheide,B.M. and Thorpe,M.F. (2005) *Phys. Biol*. (in press).
14. Jacobs,D.J., Rader,A.J., Kuhn,L.A. and Thorpe,M.F. (2001) Protein flexibility predictions using graph theory. *Proteins*, **44**, 150–165.
15. Brooks,B.R., Bruccoleri,R.E., Olafson,B.D., States,D.J., Swaminathan,S. and Karplus,M. (1983) CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem*., **4**, 187–217.
16. Lindahl,E., Hess,B. and van der Spoel,D. (2001) GROMACS: a package for molecular simulation and trajectory analysis. *J. Mol. Mod*., **7**, 306–317.
17. Rader,A.J., Hespenheide,B.M., Kuhn,L.A. and Thorpe,M.F. (2002) Protein unfolding: rigidity lost. *Proc. Natl Acad. Sci. USA*, **99**, 3540–3545.
18. Thorpe,M.F., Lei,M., Rader,A.J., Jacobs,D.J. and Kuhn,L.A. (2001) Protein flexibility and dynamics using constraint theory. *J. Mol. Graph. Model*., **19**, 60–69.
19. Porter,C.T., Bartlett,G.J. and Thornton,J.M. (2004) The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res*., **32**, D129–D133.
20. Holm,L. and Sander,C. (1998) Removing near-neighbour redundancy from large protein sequence collections. *Bioinformatics*, **14**, 423–429.
21. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*., **22**, 4673–4680.
22. Schneider,T.D., Stormo,G.D., Gold,L. and Ehrenfeucht,A. (1986) Information content of binding sites on nucleotide sequences. *J. Mol. Biol*., **188**, 415–431.