Conformational dynamics data bank: a database for conformational dynamics of proteins and supramolecular protein assemblies

Do-Nyun Kim[1], Josiah Altschuler[2], Campbell Strong[2], Gaël McGill[2,3] and Mark Bathe[1,*]

[1]Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, [2]Digizyme, Inc., Brookline, MA 02446 and [3]Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, MA 02115, USA

ABSTRACT

The conformational dynamics data bank (CDDB, http://www.cdyn.org) is a database that aims to provide comprehensive results on the conformational dynamics of high molecular weight proteins and protein assemblies. Analysis is performed using a recently introduced coarse-grained computational approach that is applied to the majority of structures present in the electron microscopy data bank (EMDB). Results include equilibrium thermal fluctuations and elastic strain energy distributions that identify rigid versus flexible protein domains generally, as well as those associated with specific functional transitions, and correlations in molecular motions that identify molecular regions that are highly coupled dynamically, with implications for allosteric mechanisms. A practical web-based search interface enables users to easily collect conformational dynamics data in various formats. The data bank is maintained and updated automatically to include conformational dynamics results for new structural entries as they become available in the EMDB. The CDDB complements static structural information to facilitate the investigation and interpretation of the biological function of proteins and protein assemblies essential to cell function.

INTRODUCTION

High molecular weight protein assemblies are actively involved in diverse cellular functions including transcription, translation, protein folding and degradation, nuclear–cytoplasmic translocation of biomolecules and cell division. While the static structure of proteins provides invaluable insight into their functional mechanism, their conformational dynamics provide additional important insight that often cannot be inferred from static structure alone. Examples include the collective motion of molecular domains, allosteric coupling and the identification of molecular regions that are important to complex stability (1–5).

The conformational dynamics data bank (CDDB) distributes these results to the broader scientific community, focusing on high molecular weight protein assemblies whose structure is based on single-particle cryo-electron microscopy (cryo-EM) reconstruction. More than 900 such structures are available at present in the electron microscopy data bank (EMDB) (6,7). Since its inception in 2002, the EMDB has been growing rapidly in recent years, with a submission rate that has reached ~150 structures per year, paralleling early growth of the Protein Data Bank (PDB) (8,9). While the conformational dynamics of proteins stored in the PDB are publically accessible through several data banks and servers (10–15), similar data banks do not exist at present for entries in the EMDB, despite their importance to molecular cell biology and the fact that many of these structures are not available in the PDB. The CDDB provides detailed conformational dynamics information for the majority of structures deposited in the EMDB, including conformational flexibility (16–18), dynamical coupling between distant domains that may be involved in allosteric mechanisms (2–4), and elastic strain energies associated with specific functional motions that identify molecular regions important for structural stability in these motions (19–21). These data both support further computational analyses to gain insight into the biological function of high molecular weight protein assemblies, as well as serve as a basis set for classification in single particle reconstructions (22).

Conformational dynamics are computed using a well-established procedure called Normal Mode Analysis

*To whom correspondence should be addressed. Tel: +1 617 324 5685; Fax: +1 617 324 7554; Email: mark.bathe@mit.edu

(NMA) (16,23,24), which achieves relative computational efficiency over molecular or Brownian dynamics by approximating molecular motions as a linear superposition of harmonic oscillations, called normal modes (NMs) (16). In practice, only a subset (20–100) of the lowest non-rigid body NMs is required to accurately describe the dynamical motions of macromolecules (25,26). NMA is performed using a recently introduced coarse-grained computational framework based on the finite element (FE) method (25) that is computationally efficient, capable of reproducing atomic-level NMA results quantitatively (25), and particularly well-suited to the analysis of EM-based structures that lack atomic coordinates because it is based on the molecular volume of proteins (26). The data bank is updated actively in an unsupervised manner as new depositions become available in the EMDB. The data bank may be extended to incorporate viral assemblies in the VIPERdb (27,28), as well as structures from the PDB in the future.

DATABASE ARCHITECTURE

The CDDB is driven by the MySQL open source relational database management system (http://dev.mysql.com/doc/refman/5.1/en/index.html), and all search queries via the client browser are implemented on the server in PHP (http://www.php.net/manual/en/). Table columns contain in-house conformational dynamics data corresponding to five major categories: Molecular Surface, Root Mean Square Fluctuations, Correlations in Molecular Domains, Elastic Strain Energy Distributions, and FE Model and Results (26). In addition, there is an overview category with columns corresponding to visualization and sample information present in the EMDB. These data are automatically downloaded from the EMDB FTP (Rutgers) site and integrated into the CDDB on a weekly basis. Figure 1 displays the information flow from the database to the visiting user at cdyn.org. Individual database tables have been normalized to provide users with a simple search interface where they may search by any key word. Further, advanced search filters are presented to the user in a drop down box. These advanced search fields essentially constrain the user to create queries more closely aligned to the actual tables of the database.

DATA GENERATION

Conformational dynamics results are computed and maintained using an unsupervised procedure (26) that consists of several distinct computational steps ranging from EM density map retrieval from the EMDB to NMA results processing (Supplementary Data). The EMDB is monitored regularly for suitable new entry depositions, for which NMA and results processing are automatically performed. Entries that are excluded from analysis include electron tomograms, entries for which neither contour level nor molecular weight are provided thereby precluding molecular surface calculation, entries that consist of multiple disconnected bodies and entries with defective surface meshes (26).
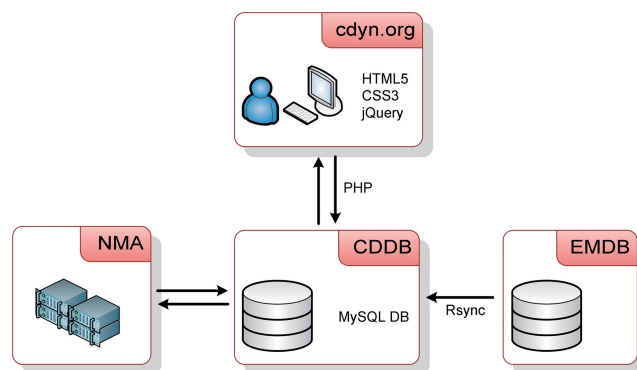

Figure 1. Schematic illustration of information flow from the database to the user. The user interface of cdyn.org is designed using HTML5 (HyperText Markup Language), CSS3 (Cascading Style Sheets) and jQuery, and is compliant across all major web browsers (Firefox, Safari, Internet Explorer, Chrome and Opera). HTML provides the base structure of each page, CSS provides the styling of fonts, colors and layouts and jQuery is used for animations such as the slide down boxes under 'Additional search parameters'. When the user clicks on the Search button, another language, PHP (PHP: Hypertext Preprocessor), is used to query the database for information. This database houses all of the data needed to drive cdyn.org, and is implemented using the MySQL relational database management system. PHP processes the returned results from each of its queries to MySQL and outputs them as dynamically generated web pages to the user. The database itself contains all of the NMA data, as well as data from the EMDB. Data in the EMDB is available via the FTP site at ftp://emdb.rutgers.edu/. The integration of EMDB data into the local database allows the user to specify both NMA and EMDB search parameters in cdyn.org.

NMs and resultant dynamical properties are computed using the FE framework for proteins, which treats molecules as homogeneous isotropic elastic bodies defined by their closed molecular surface (25,26). Computation and discretization of the molecular surface from the EM density map is performed using the marching cubes algorithm (29) implemented in UCSF Chimera (30) based on the suggested contour level provided in the EMDB. The triangulated molecular surface is then processed by several surface mesh filters available in Meshlab (31) to remove defective triangles that hinder generation of the 3D volumetric FE mesh. The commercially available FE analysis program ADINA (ADINA R&D, Inc., Watertown, MA, USA) is used to generate the 3D volumetric FE mesh and calculate the lowest 100 NMs using NMA.

Root mean square fluctuations (RMSFs) are calculated in results processing by weighting NMs according to the equipartition theorem of statistical thermodynamics (32), which states that the mean equilibrium elastic energy corresponding to each NM is $\frac{1}{2}k_BT$, where k_B is the Boltzmann constant and T is temperature, assumed to be 298 K (Supplementary Data). Weighted NMs can additionally be used to calculate dynamical coupling between distinct molecular regions using a mutual information metric (26,33,34), and corresponding elastic strain energy distributions associated with each NM are computed and stored, indicating rigid versus flexible regions corresponding to distinct functional motions (26).
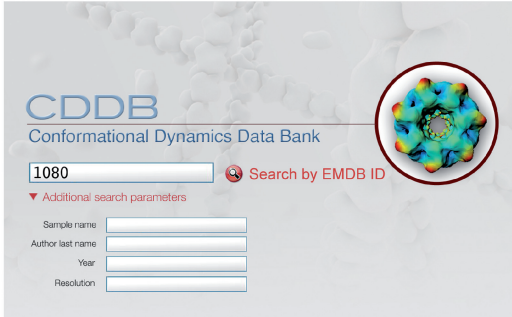
DATA ACCESS AND CONTENT

The CDDB offers a practical web-based search interface for conformational dynamics data. Users already familiar with the EMDB can search for any entry using the corresponding EMDB ID. A keyword search is additionally provided for an advanced search using the sample name, author last name, aggregation state and resolution. The search leads to the search results page that lists all matching entries with basic identifying information including the EMDB ID, the sample source, the resolution, the title and authors of the primary citation and the EM density map release date. Clicking on an entry header on the search results page shows the corresponding entry page where conformational dynamics results are available for download in various formats (Figure 2).

Each entry page begins with an overview of the molecule with images including a surface rendering of the molecule from the EMDB, EM density map surface renderings at the suggested contour level prior to FE mesh generation, molecular surfaces after preparation for FE mesh generat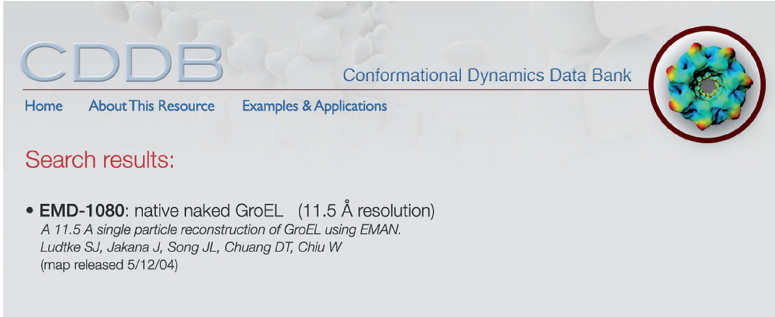ion and molecular surfaces colored by RMSFs. Several sections follow as drop down boxes where users can access various results. The molecular surface section provides deformed molecular surfaces corresponding to each NM, as well as the initial molecular surface used for NM calculation in STL file format, which is native to the stereolithography CAD software created by 3D Systems, Inc., Rock Hill, SC, USA. These files are useful for animation of thermal fluctuations of the molecule using programs such as Molecular Maya (http://www.molecularmovies.org/toolkit/) and Maya (Autodesk, Inc., San Rafael, CA, USA). Several such movies are available in the 'Examples & Applications' page in the CDDB. The entry sections for RMSF and elastic strain energy distributions also provide the initial molecular surfaces, whose vertices are colored by relative values of RMSFs and elastic strain energy densities in PLY file format, also known as the polygon file format or the Stanford triangle format, supported by many computer graphics software packages.

The last section of the entry page is devoted to all other data formats including the FE model used for

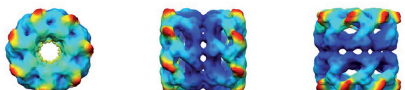**Figure 2.** CDDB web interface. (**A**) A practical search interface is provided in the main page of the CDDB where users can search by EMDB ID or keyword. (**B**) Search results when EMDB ID 1080 is entered, corresponding to the chaperonin molecule GroEL solved at 11.5 Å resolution (37). (**C**) Entry page corresponding to EMDB ID 1080, beginning with an overview of the molecule and followed by surface renderings that illustrate the EM density map rendered at the contour level suggested in the EMDB entry, the filtered molecular surface mesh used for generation of the FE model, and the RMSF-colored molecular surface obtained from NMA. Several sections follow as drop down boxes where users can access specific results in various formats for download.

conformational dynamics analysis, which may be further used to simulate mechanical response properties of the molecule. Eigenvectors and their magnitudes, corresponding elastic strain energy densities for each NM and total RMSF amplitudes at FE nodal points are stored in ASCII format. These results are additionally interpolated to each voxel of the original electron density map and stored in MRC electron density map format (http://www2.mrc-lmb.cam.ac.uk/image2000.html). Users may view both the original density map and result maps simultaneously using, for example, Chimera. Molecular animations that illustrate the dynamical motions associated with each NM are also available for download in both high (640 × 480 pixels) and low (320 × 240 pixels) resolutions.

CONCLUSIONS

The CDDB distributes for the first time to the molecular cell biology community conformational dynamics results for supramolecular protein assemblies. Many of these results are uniquely available from the EM-based structures employed because high molecular weight complexes, as well as some other lower molecular weight proteins and complexes, are not amenable to crystallography. EM has the additional advantage over crystallography that samples are prepared in their native, physiological state, including effects of pH, salt concentration and ligand concentration on structural states. It is well-established that the presented NMA-based conformational dynamics results are useful in identifying and understanding allosteric mechanisms that mediate signal transduction and catalysis in proteins (2), molecular regions and residues that are key to functional transitions in proteins (26), as well in the analysis of protein stability and mechanics (34–36). While the present rendition of the data bank is focused on structures deposited in the EMDB, the approach employed is generally applicable to atom-based structures present in other structural databases including the VIPERdb and PDB. Conformational dynamics results from these databases may be included in the CDDB in the future, providing a single database for dynamical information related to proteins and protein assemblies.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

REFERENCES

1. Ma,J.P. and Karplus,M. (1998) The allosteric mechanism of the chaperonin GroEL: a dynamic analysis. *Proc. Natl Acad. Sci. USA*, **95**, 8502–8507.
2. Cui,Q. and Karplus,M. (2008) Allostery and cooperativity revisited. *Protein Sci.*, **17**, 1295–1307.
3. Kern,D. and Zuiderweg,E.R.P. (2003) The role of dynamics in allosteric regulation. *Curr. Opin. Struct. Biol.*, **13**, 748.
4. Csermely,P., Palotai,R. and Nussinov,R. (2010) Induced fit, conformational selection and independent dynamic segments: an extended view of binding events. *Trends Biochem. Sci.*, **35**, 539–546.
5. Zheng,W.J., Brooks,B.R. and Thirumalai,D. (2006) Low-frequency normal modes that describe allosteric transitions in biological nanomachines are robust to sequence variations. *Proc. Natl Acad. Sci. USA*, **103**, 7664–7669.
6. Tagari,M., Newman,R., Chagoyen,M., Carazo,J.M. and Henrick,K. (2002) New electron microscopy database and deposition system. *Trends Biochem. Sci.*, **27**, 589.
7. Henrick,K., Newman,R., Tagari,M. and Chagoyen,M. (2003) EMDep: a web-based system for the deposition and validation of high-resolution electron microscopy macromolecular structural information. *J. Struct. Biol.*, **144**, 228–237.
8. Bernstein,F.C., Koetzle,T.F., Williams,G.J.B., Meyer,E.F., Brice,M.D., Rodgers,J.R., Kennard,O., Shimanouchi,T. and Tasumi,M. (1977) The protein data bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.*, **112**, 535–542.
9. 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.
10. Alexandrov,V., Lehnert,U., Echols,N., Milburn,D., Engelman,D. and Gerstein,M. (2005) Normal modes for predicting protein motions: a comprehensive database assessment and associated Web tool. *Protein Sci.*, **14**, 633–643.
11. Suhre,K. and Sanejouand,Y.-H. (2004) ElNemo: a normal mode web server for protein movement analysis and the generation of templates for molecular replacement. *Nucleic Acids Res.*, **32**, W610–W614.
12. Wako,H., Kato,M. and Endo,S. (2004) ProMode: a database of normal mode analyses on protein molecules with a full-atom model. *Bioinformatics*, **20**, 2035–2043.
13. Yang,L.W., Liu,X., Jursa,C.J., Holliman,M., Rader,A., Karimi,H.A. and Bahar,I. (2005) iGNM: a database of protein functional motions based on Gaussian network model. *Bioinformatics*, **21**, 2978–2987.
14. Hollup,S.M., Salensminde,G. and Reuter,N. (2005) WEBnm@: a web application for normal mode analyses of proteins. *BMC Bioinformatics*, **6**, 52.
15. Lindahl,E., Azuara,C., Koehl,P. and Delarue,M. (2006) NOMAD-Ref: visualization, deformation and refinement of macromolecular structures based on all-atom normal mode analysis. *Nucleic Acids Res.*, **34**, W52–W56.
16. Cui,Q. and Bahar,I. (eds), (2006) *Normal Mode Analysis: Theory and Applications to Biological and Chemical Systems*. Chapman & Hall/CRC, Boca Raton.
17. Brooks,B. and Karplus,M. (1985) Normal modes for specific motions of macromolecules: application to the hinge-bending mode of lysozyme. *Proc. Natl Acad. Sci. USA*, **82**, 4995–4999.
18. Levitt,M., Sander,C. and Stern,P.S. (1985) Protein normal-mode dynamics: trypsin inhibitor, crambin, ribonuclease and lysozyme. *J. Mol. Biol.*, **181**, 423–447.
19. Zheng,W.J., Brooks,B.R., Doniach,S. and Thirumalai,D. (2005) Network of dynamically important residues in the open/closed transition in polymerases is strongly conserved. *Structure*, **13**, 565–577.
20. Maragakis,P. and Karplus,M. (2005) Large amplitude conformational change in proteins explored with a plastic network model: adenylate kinase. *J. Mol. Biol.*, **352**, 807–822.
21. Miyashita,O., Onuchic,J.N. and Wolynes,P.G. (2003) Nonlinear elasticity, proteinquakes, and the energy landscapes of functional transitions in proteins. *Proc. Natl Acad. Sci. USA*, **100**, 12570–12575.
22. Brink,J., Ludtke,S.J., Kong,Y.F., Wakil,S.J., Ma,J.P. and Chiu,W. (2004) Experimental verification of conformational variation of human fatty acid synthase as predicted by normal mode analysis. *Structure*, **12**, 185–191.

23. 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.

24. Go,N., Noguti,T. and Nishikawa,T. (1983) Dynamics of a small globular protein in terms of low frequency vibrational modes. *Proc. Natl Acad. Sci. USA.*, **80**, 3696–3700.

25. Bathe,M. (2008) A finite element framework for computation of protein normal modes and mechanical response. *Proteins: Struct. Funct. Bioinform.*, **70**, 1595–1609.

26. Kim,D.-N., Nguyen,C.-T. and Bathe,M. (2010) Conformational dynamics of supramolecular protein assemblies. *J. Struct. Biol.* (in press).

27. Shepherd,C.M., Borelli,I.A., Lander,G., Natarajan,P., Siddavanahalli,V., Bajaj,C., Johnson,J.E., Brooks,C.L. and Reddy,V.S. (2006) VIPERdb: a relational database for structural virology. *Nucleic Acids Res.*, **34**, D386–D389.

28. Carrillo-Tripp,M., Shepherd,C.M., Borelli,I.A., Venkataraman,S., Lander,G., Natarajan,P., Johnson,J.E., Brooks,C.L. III and Reddy,V.S. (2009) VIPERdb2: an enhanced and web API enabled relational database for structural virology. *Nucleic Acids Res.*, **37**, D436–D442.

29. Lorensen,W.E. and Cline,H.E. (1987) Marching cubes: a high resolution 3D surface construction algorithm. *Computer Graphics (SIGGRAPH '87 Proceedings)*, **21**, 163–169.

30. Pettersen,E.F., Goddard,T.D., Huang,C.C., Couch,G.S., Greenblatt,D.M., Meng,E.C. and Ferrin,T.E. (2004) UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.*, **25**, 1605–1612.

31. Cignoni,P., Callieri,M., Corsini,M., Dellepiane,M., Ganovelli,F. and Ranzuglia,G. (2008) *Sixth Eurographics Italian Chapter Conference*, pp. 129–136.

32. McQuarrie,D.A. (1975) *Statistical Mechanics*. Harper & Row, New York.

33. Lange,O.F. and Grubmüller,H. (2006) Generalized correlation for biomolecular dynamics. *Proteins: Struct. Funct. Bioinform.*, **62**, 1053–1061.

34. Sedeh,R.S., Fedorov,A.A., Fedorov,E.V., Ono,S., Matsumura,F., Almo,S.C. and Bathe,M. (2010) Structure, evolutionary conservation, and conformational dynamics of Homo sapiens fascin-1, an F-actin crosslinking protein. *J. Mol. Biol.*, **400**, 589–604.

35. Gibbons,M.M. and Klug,W.S. (2007) Nonlinear finite-element analysis of nanoindentation of viral capsids. *Phys. Rev. E.*, **75**, 031901.

36. Gibbons,M.M. and Klug,W.S. (2008) Influence of nonuniform geometry on nanoindentation of viral capsids. *Biophys. J.*, **95**, 3640–3649.

37. Ludtke,S.J., Jakana,J., Song,J.-L., Chuang,D.T. and Chiu,W. (2001) A 11.5 Å single particle reconstruction of GroEL using EMAN. *J. Mol. Biol.*, **314**, 253–262.