Proteins: Structure, Function, and Genetics
Article No. 7e1021

Copy of e-mail Notification

Your article (01247R) from "Proteins: Structure, Function, and Genetics" is available for download
=====
RE: Your article (01247R) from "Proteins: Structure, Function, and Genetics" is available for download

Proteins: Structure, Function, and Genetics Published by John Wiley & Sons, Inc.

Dear Sir or Madam,

PDF page proofs for your article are ready for review.

Please refer to this URL address

http://mothra.cadmus.com/cgi-bin/s-proof/login?628691

Login: your e-mail address
Password: ----

The site contains 1 file. You will need to have Adobe Acrobat Reader software to read these files. This is free software and is available for user downloading at
http://www.adobe.com/products/acrobat/readstep.html.

This file contains:

Author Instructions Checklist
Adobe Acrobat Users - NOTES tool sheet
Reprint Order form
Copyright Transfer Agreement
Return fax form
A copy of your page proofs for your article

After printing the PDF file, please read the page proofs carefully and:

1) indicate changes or corrections in the margin of the page proofs;
2) answer all queries (footnotes A,B,C, etc.) on the last page of the PDF proof;
3) proofread any tables and equations carefully;
4) check that any Greek, especially "mu", has translated correctly.

Within 48 hours, please return the following to the address given below:

1) original PDF set of page proofs,
2) Reprint Order form,
3) Return fax form

Return to:

Mike Evans
Journals Editorial/Production
John Wiley & Sons, Inc.
605 Third Avenue

Proteins: Structure, Function, and Genetics
Article No. 7e1021

Copy of e-mail Notification

New York, N.Y. 10158
U.S.A.

Your article will be published online via our EarlyView service within a few days of correction receipt. Your prompt attention to and return of page proofs is crucial to faster publication of your work.
If you experience technical problems, please contact Doug Frank (e-mail: FrankD@cadmus.com, phone: 800-238-3814 (X615).

If you have any questions regarding your article, please contact me. PLEASE ALWAYS INCLUDE YOUR ARTICLE NO. (01247R) WITH ALL CORRESPONDENCE.

Sincerely,

Mike Evans
Senior Production Editor
John Wiley & Sons, Inc.
E-mail: mjevans@wiley.com
Tel: 212-850-6952
Fax: 212-850-6052

WILEY
Publishers Since 1807

JOHN WILEY & SONS
605 THIRD AVENUE, NEW YORK, NY 10158

*****IMMEDIATE RESPONSE REQUIRED*****

Your article will be published online via Wiley's EarlyView® service (www.interscience.wiley.com) shortly after receipt of corrections. EarlyView® is Wiley's online publication of individual articles in full-text HTML and/or pdf format before release of the compiled print issue of the journal. Articles posted online in EarlyView® are peer-reviewed, copyedited, author corrected, and fully citable. EarlyView® means you benefit from the best of two worlds--fast online availability as well as traditional, issue-based archiving.

☐ **READ PROOFS CAREFULLY**

- This will be your only chance to review these proofs.
- Please note that the volume and page numbers shown on the proofs are for position only.

☐ **ANSWER ALL QUERIES ON PROOFS** (Queries for you to answer are noted on the manuscript.)

- Mark all corrections directly on the proofs, not on the manuscript. Note that excessive author alterations may ultimately result in delay of publication and extra costs may be charged to you.

☐ **CHECK FIGURES AND TABLES CAREFULLY** (Color figures will be sent under separate cover.)

- Check size, numbering, and orientation of figures. Check quality of figures directly from the galley proofs. The reproduction is 1200dpi, and although it is not indicative of final printed quality, it is adequate for checking purposes.
- Review figure legends to ensure that they are complete.
- Check all tables. Review layout, title, and footnotes.

☐ **COMPLETE REPRINT ORDER FORM**

- Fill out the attached reprint order form. It is important to return the form even if you are not ordering reprints. You may, if you wish, pay for the reprints with a credit card. Reprints will be mailed only after your article appears in print. The time you return proofs is the most opportune time to order reprints. If you wait until after your article comes off press, the reprints will be considerably more expensive.

RETURN

- ☐ **PROOFS**
- ☐ **ORIGINAL MANUSCRIPT**
- ☐ **REPRINT ORDER FORM**
- ☐ **ORIGINAL FIGURES**
- ☐ **Copyright Transfer Agreement (If you have not already signed one)**

Send complete package to:

John Wiley & Sons, Inc.
STM Journal Production (3400)
605 Third Avenue (9th Fl.)
New York, NY 10158-0012
attn: Mike Evans

You may fax your corrected proofs to 212-850-6052 to save time, but please also forward all original materials via Express Mail to the above address.

RETURN IMMEDIATELY AS YOUR ARTICLE WILL BE POSTED IN ORDER OF RECEIPT. YOU CAN EXPECT TO SEE YOUR ARTICLE ONLINE SHORTLY AFTER RECEIPT OF CORRECTIONS.
QUESTIONS?

Contact: Mike Evans, Senior Production Editor

Refer to article #_____

E-mail: mjevans@wiley.com

Telephone: 212-850-6952

Softproofing for advanced Adobe Acrobat Users - NOTES tool

NOTE: ACROBAT READER FROM THE INTERNET DOES NOT CONTAIN THE NOTES TOOL USED IN THIS PROCEDURE.

Acrobat annotation tools can be very useful for indicating changes to the PDF proof of your article. By using Acrobat annotation tools, a full digital pathway can be maintained for your page proofs.

The NOTES annotation tool can be used with either Adobe Acrobat 3.0x or Adobe Acrobat 4.0. Other annotation tools are also available in Acrobat 4.0, but this instruction sheet will concentrate on how to use the NOTES tool. Acrobat Reader, the free Internet download software from Adobe, DOES NOT contain the NOTES tool. In order to softproof using the NOTES tool you must have the full software suite Adobe Acrobat Exchange 3.0x or Adobe Acrobat 4.0 installed on your computer.

Steps for Softproofing using Adobe Acrobat NOTES tool:

1. Open the PDF page proof of your article using either Adobe Acrobat Exchange 3.0x or Adobe Acrobat 4.0. Proof your article on-screen or print a copy for markup of changes.

2. Go to File/Preferences/Annotations (in Acrobat 4.0) or File/Preferences/Notes (in Acrobat 3.0) and enter your name into the "default user" or "author" field. Also, set the font size at 9 or 10 point.

3. When you have decided on the corrections to your article, select the NOTES tool from the Acrobat toolbox and click in the margin next to the text to be changed.

4. Enter your corrections into the NOTES text box window. Be sure to clearly indicate where the correction is to be placed and what text it will effect. If necessary to avoid confusion, you can use your TEXT SELECTION tool to copy the text to be corrected and paste it into the NOTES text box window. At this point, you can type the corrections directly into the NOTES text box window. **DO NOT correct the text by typing directly on the PDF page.**

5. Go through your entire article using the NOTES tool as described in Step 4.

6. When you have completed the corrections to your article, go to File/Export/Annotations (in Acrobat 4.0) or File/Export/Notes (in Acrobat 3.0). Save your NOTES file to a place on your harddrive where you can easily locate it. **Name your NOTES file with the article number assigned to your article in the original softproofing e-mail message.**

7. **When closing your article PDF be sure NOT to save changes to original file.**

8. To make changes to a NOTES file you have exported, simply re-open the original PDF proof file, go to File/Import/Notes and import the NOTES file you saved. Make changes and re-export NOTES file keeping the same file name.

9. When complete, attach your NOTES file to a reply e-mail message. Be sure to include your name, the date, and the title of the journal your article will be printed in.

John Wiley & Sons, Inc.

Publishers Since 1807

REPRINT BILLING DEPARTMENT • 605 THIRD AVENUE • NEW YORK, NY 10158-0012
PHONE: (212) 850-8789; FAX: (212) 850-6326
E-MAIL: reprints @ wiley.com

PREPUBLICATION REPRINT ORDER FORM

Please complete this form even if you are not ordering reprints. This form **MUST** be returned with your corrected proofs and original manuscript. Your reprints will be shipped approximately 4 weeks after publication. Reprints ordered after printing are substantially more expensive.

JOURNAL: ***PROTEINS: Structure, Function, and Genetics*** VOLUME_____ ISSUE_____

TITLE OF MANUSCRIPT_____

MS. NO._____ NO. OF PAGES_____ AUTHOR(S)_____

REPRINTS 8 1/4 X 11

No. of Pages	100 Reprints	200 Reprints	300 Reprints	400 Reprints	500 Reprints
	$	$	$	$	$
1-4	336	501	694	890	1,052
5-8	469	703	987	1,251	1,477
9-12	594	923	1,234	1,565	1,850
13-16	714	1,156	1,527	1,901	2,273
17-20	794	1,340	1,775	2,212	2,648
21-24	911	1,529	2,031	2,536	3,037
25-28	1,004	1,707	2,267	2,828	3,388
29-32	1,108	1,894	2,515	3,135	3,755
33-36	1,219	2,092	2,773	3,456	4,143
37-40	1,329	2,290	3,033	3,776	4,528

****** REPRINTS ARE ONLY AVAILABLE IN LOTS OF 100. IF YOU WISH TO ORDER MORE THAN 500 REPRINTS, PLEASE CONTACT OUR REPRINTS DEPARTMENT AT (212)850-8789 FOR A PRICE QUOTE.

☐ Please send me _____ reprints of the above article at..... $_____

Please add appropriate State and Local Tax {Tax Exempt No._____} $_____
Please add 5% Postage and Handling..... $_____
TOTAL AMOUNT OF ORDER** ..... $_____

***International orders must be paid in U.S. currency and drawn on a U.S. bank*

Please check one: ☐ Check enclosed ☐ Bill me ☐ Credit Card
If credit card order, charge to: ☐ American Express ☐ Visa ☐ MasterCard ☐ Discover
Credit Card No._____ Signature_____ Exp. Date_____

Bill To:

Name_____
Address/Institution_____
_____
_____
_____
Purchase Order No. _____ Phone _____ Fax _____
E-mail: _____

Ship To:

Name_____
Address/Institution_____
_____
_____
_____

WILEY
Publishers Since 1807

605 Third Avenue
New York, NY 10158-0012
212-850-6000
FAX: 212-850-8852

COPYRIGHT TRANSFER AGREEMENT

Date:

To:

Production/Contribution ID#_____ Publisher/Editorial office use only

Re: Manuscript entitled_____
_____ (the "Contribution") for publication in _____ (the "Journal") published by Wiley-Liss, Inc., a subsidiary of John Wiley & Sons, Inc. ("Wiley").

Dear Contributor(s):

Thank you for submitting your Contribution for publication. In order to expedite the editing and publishing process and enable Wiley to disseminate your work to the fullest extent, we need to have this Copyright Transfer Agreement signed and returned to us as soon as possible. If the Contribution is not accepted for publication this Agreement shall be null and void.

A. COPYRIGHT

1. The Contributor assigns to Wiley, during the full term of copyright and any extensions or renewals of that term, all copyright in and to the Contribution, including but not limited to the right to publish, republish, transmit, sell, distribute and otherwise use the Contribution and the material contained therein in electronic and print editions of the Journal and in derivative works throughout the world, in all languages and in all media of expression now known or later developed, and to license or permit others to do so.

2. Reproduction, posting, transmission or other distribution or use of the Contribution or any material contained therein, in any medium as permitted hereunder, requires a citation to the Journal and an appropriate credit to Wiley as Publisher, suitable in form and content as follows: (Title of Article, Author, Journal Title and Volume/Issue Copyright © [year] Wiley-Liss, Inc. or copyright owner as specified in the Journal.)

B. RETAINED RIGHTS

Notwithstanding the above, the Contributor or, if applicable, the Contributor's Employer, retains all proprietary rights other than copyright, such as patent rights, in any process, procedure or article of manufacture described in the Contribution, and the right to make oral presentations of material from the Contribution.

C. OTHER RIGHTS OF CONTRIBUTOR

Wiley grants back to the Contributor the following:

1. The right to share with colleagues print or electronic "preprints" of the unpublished Contribution, in form and content as accepted by Wiley for publication in the Journal. Such preprints may be posted as electronic files on the Contributor's own website for personal or professional use, or on the Contributor's internal university or corporate networks/intranet, or secure external website at the Contributor's institution, but not for commercial sale or for any systematic external distribution by a third party (e.g., a listserve or database connected to a public access server). Prior to publication, the Contributor must include the following notice on the preprint: "This is a preprint of an article accepted for publication in [Journal title] © copyright (year) (copyright owner as specified in the Journal)". After publication of the Contribution by Wiley, the preprint notice should be amended to read as follows: "This is a preprint of an article published in [include the complete citation information for the final version of the Contribution as published in the print edition of the Journal]", and should provide an electronic link to the Journal's WWW site, located at the following Wiley URL: http://www.interscience.Wiley.com/. The Contributor agrees not to update the preprint or replace it with the published version of the Contribution.

2. The right, without charge, to photocopy or to transmit online or to download, print out and distribute to a colleague a copy of the published Contribution in whole or in part, for the Colleague's personal or professional use, for the advancement of scholarly or scientific research or study, or for corporate informational purposes in accordance with Paragraph D.2 below.

3. The right to republish, without charge, in print format, all or part of the material from the published Contribution in a book written or edited by the Contributor.

4. The right to use selected figures and tables, and selected text (up to 250 words, exclusive of the abstract) from the Contribution, for the Contributor's own teaching purposes, or for incorporation within another work by the Contributor that is made part of an edited work published (in print or electronic format) by a third party, or for presentation in electronic format on an internal computer network or external website of the Contributor or the Contributor's employer.

5. The right to include the Contribution in a compilation for classroom use (course packs) to be distributed to students at the Contributor's institution free of charge or to be stored in electronic format in datarooms for access by students at the Contributor's institution as part of their course work (sometimes called "electronic reserve rooms") and for in-house training programs at the Contributor's employer.

D. **CONTRIBUTIONS OWNED BY EMPLOYER**

1. If the Contribution was written by the Contributor in the course of the Contributor's employment (as a "work-made-for-hire" in the course of employment), the Contribution is owned by the company/employer which must sign this Agreement (in addition to the Contributor's signature), in the space provided below. In such case, the company/employer hereby assigns to Wiley, during the full term of copyright, all copyright in and to the Contribution for the full term of copyright throughout the world as specified in paragraph A above.

2. In addition to the rights specified as retained in paragraph B above and the rights granted back to the Contributor pursuant to paragraph C above, Wiley hereby grants back, without charge, to such company/employer, its subsidiaries and divisions, the right to make copies of and distribute the published Contribution internally in print format or electronically on the Company's internal network. Upon payment of Wiley's reprint fee, the institution may distribute (but not resell) print copies of the published Contribution externally. Although copies so made shall not be available for individual re-sale, they may be included by the company/employer as part of an information package included with software or other products offered for sale or license. Posting of the published Contribution by the institution on a public access website may only be done with Wiley's written permission, and payment of any applicable fee(s).

E. **GOVERNMENT CONTRACTS**

In the case of a Contribution prepared under U.S. Government contract or grant, the U.S. Government may reproduce, without charge, all or portions of the Contribution and may authorize others to do so, for official U.S. Government purposes only, if the U.S. Government contract or grant so requires. (U.S. Government Employees: see note at end.)

F. **COPYRIGHT NOTICE**

The Contributor and the company/employer agree that any and all copies of the Contribution or any part thereof distributed or posted by them in print or electronic format as permitted herein will include the notice of copyright as stipulated in the Journal and a full citation to the Journal as published by Wiley.

G. **CONTRIBUTOR'S REPRESENTATIONS**

The Contributor represents that the Contribution is the Contributor's original work. If the Contribution was prepared jointly, the Contributor agrees to inform the co-Contributors of the terms of this Agreement and to obtain their signature to this Agreement or their written permission to sign on their behalf. The Contribution is submitted only to this Journal and has not been published before, except for "preprints" as permitted above. (If excerpts from copyrighted works owned by third parties are included, the Contributor will obtain written permission from the copyright owners for all uses as set forth in Wiley's permissions form or in the Journal's Instructions for Contributors, and show credit to the sources in the Contribution.) The Contributor also warrants that the Contribution contains no libelous or unlawful statements, does not infringe upon the rights (including without limitation the copyright, patent, or trademark rights) or privacy of others, or contain material or instructions that might cause harm or injury.

CHECK ONE:

[____]Contributor-owned work

_____ _____
Contributor's signature Date

_____
Type or print name and title

_____ _____
Co-contributor's signature Date

_____
Type or print name and title

ATTACH ADDITIONAL SIGNATURE PAGE AS NECESSARY

[____]Company/Institution-owned work (made-for-hire in the course of employment)

_____ _____
Company or Institution (Employer-for-Hire) Date

_____ _____
Authorized signature of Employer Date

[____]**U.S. Government work**

Note to U.S. Government Employees

A Contribution prepared by a U.S. federal government employee as part of the employee's official duties, or which is an official U.S. Government publication is called a "U.S. Government work," and is in the public domain in the United States. In such case, the employee may cross out Paragraph A.1 but must sign and return this Agreement. If the Contribution was not prepared as part of the employee's duties or is not an official U.S. Government publication, it is not a U.S. Government work.

[____]**U.K. Government work (Crown Copyright)**

Note to U.K. Government Employees

The rights in a Contribution prepared by an employee of a U.K. government department, agency or other Crown body as part of his/her official duties, or which is an official government publication, belong to the Crown. In such case, Wiley will forward the relevant form to the Employee for signature.

605 THIRD AVENUE, NEW YORK, NY 10158

Telephone Number: 212.850.6952 Facsimile Number: 212.850.6052

To: Mike Evans, Senior Production Editor

Company: STM Journals Production, John Wiley & Sons, Inc.

Phone:

Fax:

From:

Date:

Pages including this cover page:

re:

PROTEINS: Structure, Function, and Genetics 00:000–000 (2002)

AQ: 1

Automatic Classification of a Database of Macromolecular Motions Based on Normal Mode Statistics

W. G. Krebs,[1] Vadim Alexandrov,[1] Cyrus A. Wilson,[2] Nathaniel Echols,[1] Haiyuan Yu,[1] and Mark Gerstein[1]*
[1]*Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut*
[2]*Department of Biochemistry, Stanford University, Stanford, California*

ABSTRACT **We investigated protein motions using normal modes within a database framework, determining on a large sample the degree to which normal modes anticipate the direction of the observed motion and were useful for motions classification. As a starting point for our analysis, we identified a large number of examples of protein flexibility from a comprehensive set of structural alignments of the proteins in the PDB. Each example consisted of a pair of proteins that were considerably different in structure given their sequence similarity. On each pair, we performed geometric comparisons and adiabatic-mapping interpolations in a high-throughput pipeline, arriving at a final list of 3,814 putative motions and standardized statistics for each. We then computed the normal modes of each motion in this list, determining the linear combination of modes that best approximated the direction of the observed motion. We integrated our new motions and normal mode calculations in the Macromolecular Motions Database, through a new ranking interface at http://molmovdb.org. Based on the normal mode calculations and the interpolations, we identified a new statistic, mode concentration, related to the mathematical concept of information content, which describes the degree to which the direction of the observed motion can be summarized by a few modes. Using this statistic, we were able to determine the fraction of the 3,814 motions where one could anticipate the direction of the actual motion from only a few modes. We also investigated mode concentration in comparison to related statistics on combinations of normal modes and correlated it with quantities characterizing protein flexibility (e.g., maximum backbone displacement or number of mobile atoms). Finally, we evaluated the ability of mode concentration to automatically classify motions into a variety of simple categories (e.g., whether or not they are "fragment-like"), in comparison to motion statistics. This involved the application of decision trees and feature selection (particular machine-learning techniques) to training and testing sets derived from merging the "list" of motions with manually classified ones. Proteins 2002;00:000–000.** © 2002 Wiley-Liss, Inc.

INTRODUCTION

Protein motions play a key role in a wide range of biological phenomena, including chemical concentration regulation, signal transduction, transport of metabolites, and cellular locomotion.[1–3] Motion is typically the way a structure actually carries out a specific function; for this reason, motions are an essential link between function and structure.

We previously developed a database of macromolecular motions,[1,4,5] which consisted of crystallographically documented protein motions. We also developed a morph server coupled to a collection of protein "morph" movies and related statistics.[6] Here:

1. we identify ~4,000 putative new motions from automatic structural comparison on the PDB[7];
2. we add these to our database and present the results in a new ranking interface;
3. we analyze the dynamics of these many motions, perform normal mode analysis on them, and calculate statistics to encapsulate the results of the normal mode analysis;
4. from the normal mode analysis and the interpolations, we assemble a corpus of statistics and perform datamining and feature extraction on this corpus; and
5. we identify a number of statistics, in particular, mode concentration, that we find useful.

Our work builds upon a rich literature in macromolecular motions.[8–11] Motion related to proteins' mechanical function has mainly been studied experimentally by X-ray crystallography. Traditional X-ray crystallography has provided key insights into the relationships between conformational change and macromolecular function; GroEL[12] and beta-actin[13] are just two of many examples. Progress in the field of time-resolved X-ray crystallography[14–16] has also enhanced the study of biologically significant protein conformational change. Recently, it has become possible to study larger protein conformational changes via NMR.[17] Other approaches have focused on the use of computational methods.[18–25] A systematic comparison of PDB-

W.G. Krebs's current address is Department of Integrative Biosciences, San Diego Supercomputer Center MC 0505, The University of California, 9500 Gilman Drive, La Jolla, CA 92093-0505.

Grant sponsor: Keck Foundation; Grant sponsor: National Science Foundation; Grant number: DBI - 9723182.

*Correspondence to: Mark Gerstein, Dept of Molecular Biophysics and Biochemistry, Yale University, PO Box 208114, New Haven, CT 06520. E-mail: Mark.Gerstein@yale.edu

Received 10 September 2001; Accepted 18 March 2002

Published online 00 Month 2002 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.10168

derived difference vectors has been published elsewhere on a much smaller scale.[26]

Normal mode analysis is a computational approach that can be applied to protein conformational change. Widely used by spectroscopists for many years to associate IR and Raman experimental peaks with small molecule vibrational modes,[27] advances in computer technology over the last few decades has made normal mode analysis of proteins and other large molecules practical. This was first applied to proteins in the mid 1980s and has subsequently been scaled up.[28–34] The concept of normal mode analysis is to find a set of basis vectors (normal modes) describing the molecule's concerted atomic motion and spanning the set of all $3N - 6$ degrees of freedom. For very large molecules, it is often of more interest to try to find a small subset of these normal modes that seem in some way especially important. By modeling the interatomic bonds as springs and analyzing the protein as a large set of coupled harmonic oscillators, one can calculate a frequency of periodic motion associated with each normal mode, and then attempt to find normal modes with low frequencies. The low-frequency normal modes of proteins are thought to correspond to the large-scale real-world vibrations of the protein, and can be used to deduce significant biological properties. There is evidence to suggest[35–40] that proper, symmetric normal mode vibration of binding pockets is crucial to correct biological activity in some proteins.

The principal of normal mode analysis is to solve an eigenvalue equation of the form

$$\ddot{\mathbf{q}} + \mathbf{F} \cdot \mathbf{q} = \mathbf{0} \tag{1}$$

where the vector $\mathbf{q}$ is a vector representing the displacements in three dimensions of the various atoms of the molecule, and $\mathbf{F}$ is a matrix that can be computed from the system's mass and potential energy functions. Solutions to the above system are vectors of periodic functions (the normal modes) vibrating in unison at the characteristic frequency of the mode.

Normal modes have proved to be highly useful in both modeling protein motions and in interpretation of the experimental results.[29,32,41–53] Macromolecular motions can be often characterized by a long (nanosecond or beyond) time-scale, and it has been suggested[54,55] that it may be possible to identify one or a few low-frequency normal modes, which would connect conformational endpoints. However, in certain cases (e.g., calmodulin motion) the amplitudes for the actual (observed) motion and the normal modes displacement vectors may differ by several orders of magnitude. For these cases, our theory may only be valid in interpretation of the motion initiation stage and in analysis of facilitating factors causing the actual motion.

In this paper we apply normal mode analysis to the study of protein motions. Fundamentally, we chose normal modes over MD and other related computational techniques because normal mode analysis gives a concise description of a motion (in terms of a small number of modes) that is ideal for subsequent statistical tabulation. Also, the application of normal mode analysis techniques to ~4,000 conformational changes is much less expensive than most of the competing techniques.

In this analysis, the question we are trying to answer is to what degree the direction of the observed motion (a set of vectors connecting the structure pair) occurs along with the displacement vectors of the lowest normal modes for the initial conformation. This may indirectly provide an insight about how much protein dynamics is dominated by anharmonic contributions, even though it was not a goal of this work to develop any such quantitative anharmonicity measure. Since the structure pairs may not always be available, one of the main motivations behind this work was to see if it were possible to develop an inexpensive motion analysis technique capable of assessing the direction of the actual protein motion.

Our normal mode analyses are related to the "Essential Dynamics" (ED) methods of Berdensen[56,57] on normal modes, involving a singular value decomposition analysis of normal mode atomic displacements and how they relate to experimentally solved conformations. (Essential Dynamics can also be applied to other dynamical approaches that generate displacements including techniques that do not make a harmonic assumption such as MD simulations or experimentally determined ensembles of structures.[56] However, our analysis is in many ways formally different, and we apply it within a database framework. Many of the problems customarily found in ED analyses also apply: e.g., the superfluous rotational and translational differences must be eliminated by superimposing the experimental structures to fix at least one domain; in the process, the motion's screw-axis may be characterized.[58] Previously, we developed web software tools to solve these problems in a different way using purely experimental information.[6] Here, we analyze a comprehensive database of thousands of putative protein motions, whereas existing publications limit their scope to single proteins or databases specific to certain types of proteins. AQ: 2

MATERIALS AND METHODS

Data Sources

Full outlier set

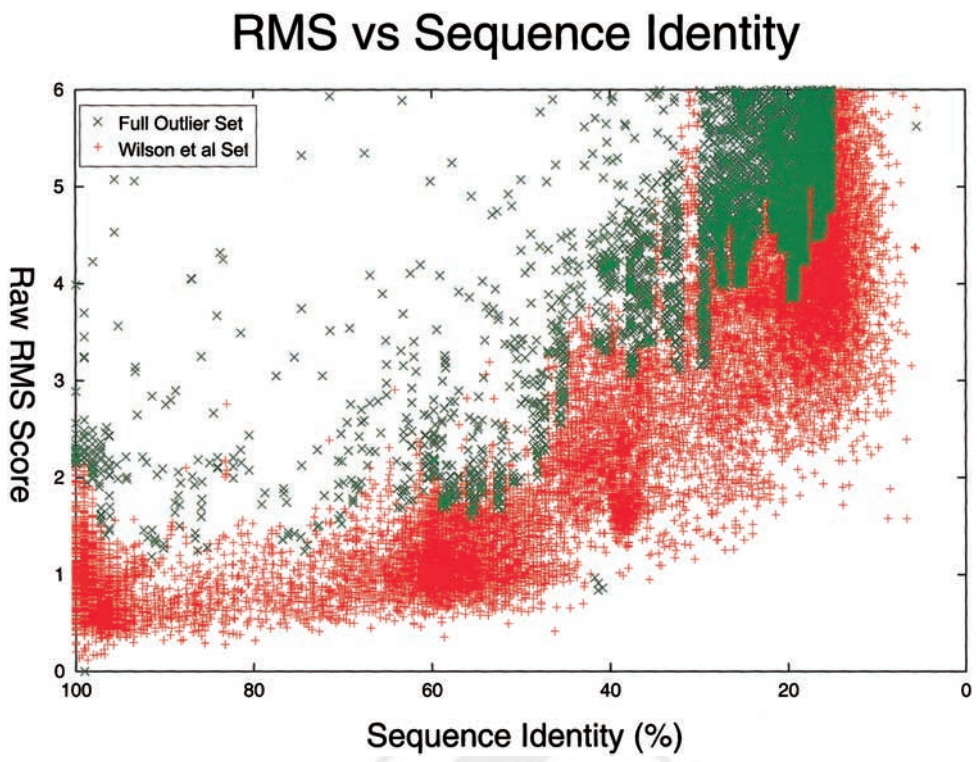
To identify a large dataset of proteins with conformational changes, Wilson et al.[59] performed automatic pairwise sequence, structure, and function comparisons on about 30,000 pairs of protein domains constructed according to scop fold classification.[60–65] Using this set of alignments, we were able to identify ~4,400 pairs of likely protein motions. We call this set the "full outlier set" (the definitions of these terms are shown in Table Ia). Its construction is described in detail in Figure 1. Basically, we plotted RMS structure alignment scores against sequence percent identity for the ~30,000 scop domain pairs aligned in Wilson et al.[59] We binned the plot into one-percent-wide bins. For each bin, we computed a mean RMS and standard deviation. Points lying more than two standard deviations above the mean were removed from the dataset and used to generate a new dataset, the full outlier dataset, which ultimately consisted of 4,400 such pairs. T1 F1

TABLE I. Definitions Table[†]

A. Term	Definition or URL lcation
Macromolecular motions database	http://bioinfo.mbb.yale.edu/MolMovDB Used for classification and annotation of motions in outlier database
SCOP database	http://scop.mrc-lmb.cam.ac.uk/scop/ Used for classification and annotation of motions via SCOP extension technique.
Wilson et al.[59] set	As shown in Figure 1, a set of 30,000 of SCOP identifier pairs was constructed for Wilson et al.[59] This was then separated into two sets: the 30,000 pair "Wilson et al." set used in that paper, and the "Full Outlier Set" (described immediately below), which we use in this text. See the caption to Figure 1 for more information.
Full outlier set	Text file http://bioinfo.mbb.yale.edu/molmovdb/datasets/outliers.txt Pairs of proteins (SCOP domains) whose structural similarity score more than two standard deviations above the mean structural similarity for their sequence similarity. See the caption to Figure 1 for more information on the construction of this set.
Workable outlier set	This is the subset of the full outlier set on which both morph server processing and normal mode analysis were successful. It consists of 3,814 motion pairs.
Manual training set	This is the training set that was produced by examining the SCOP domains in the outlier set for matches against PDB IDs in the set of manually classified motions in the Database of Macromolecular Motions.[1] Matches received the same classification as in the database, which were determined by manual examination of the specific literature. Thus, confidence in the accuracy of these classification is high.
Extended training set	The outlier set was searched for pairs that shared the same SCOP fold family as pairs classified in the Manual Training Set; these then received an identical classification. We found empirically that, because proteins that share the same SCOP fold often share similar mechanisms, proteins with the same SCOP fold have a high probability of undergoing similar conformation change and, hence, sharing the same motion size classification. Consequently, these classifications should be accurate but are less reliable than the classification in the Manual Training Set.
Classified set	This is simply the entire workable outlier set (minus those already classified in the extended training set) run through the automatic classifier defined by the decision tree, which we produced when we analyzed the extended training set.

B. Term	Definition
Mode concentration	This is discussed extensively in the text. It is a simple measure of how much the protein's motion is concentrated into any single low-frequency normal mode.
No. of CAatoms	Number of C-alpha atoms in the protein
Residuals	This is the Euclidean length of the residual difference between the atomic displacements between protein pairs and the SVD fit of the normal modes to the atomic displacements (in Angstroms)
Norm0	Maximum Value of the SVD displacement vector (unitless)
Norm1	Mean of the SVD displacement vector (unitless)
Norm2	Root-mean-square of the SVD displacement vector (unitless)
Frequency	The frequency in relative units of the normal mode with the highest SVD coefficient.
Ranking overlap	Rank of the normal mode with the largest overlap (unitless). Overlap is defined in the caption to Figure 2.
Maximum overlap	Value of the largest overlap (unitless quantity). Overlap is defined in the caption to Figure 2.
Size of 2nd core	This is the number of residues in the 2nd core (the 2ndCoreCAs key in the database). This is typically related to the size of the protein, although in poorly matched protein pairs the number can be less.
Trimmed RMS	This is the trimmed RMS score, as defined in Wilson et al.[59] and Gerstein and Krebs.[1]
Maximum CA movement	This is the largest movement (in Angstroms) of any residue during the course of the motion, as computed by the Morph Server.
Number of atoms	This is the number of atoms in the protein as computed by the Morph Server. (Atoms in non-standard amino acids are excluded). This is a measure of the size of the protein.
Energy of frames	The Morph Server computes energies for the various intermediate structures. These show a strong relationship to the sequence similarity between the two structures, and are indicators of how "good" a given morph is. The relationship of intermediate energies (energy of 4th frame, for example) with endpoint frames (energy of 8th frame, for example) can sometimes provide a rough sense of activation energies.
Translation	In hinge motions, the approximate translation (in Angstroms) the moving domains undergoes in the course of the motion, as automatically computed by the morph server. (This number is also computed for non-hinge motions, where it is less meaningful.)
Hinge rotation	In hinge motions, the rotation (in degrees) of the moving domain around the screw axis in the course of the motion, as automatically computed by the morph server. (This number tends to be small in non-hinge motions.)
Number of hinges	The number of putative hinges, or flexible linkages involved in the motion, as determined by the Morph Server
Traditional RMS	This is simply the traditional RMS score between the domains.
Rank of Norm0 mode	This is a software index that identifies the normal mode contributing the most to the motion as computed within our SVD framework. (The same normal mode that sets norm 0.)

[†]Section A lists the various data sources used in this paper, giving the location of each, along with a brief explanation of its use or importance. Section B lists definitions of the key statistics and other terms used in subsequent tables as well as in the text of the paper.



COLOR

Fig. 1. Construction of full outlier set. The crosses on this page illustrate motion pairs plotted in terms of RMS structure alignment scores against sequence percent identity for the 30,000 SCOP domain pairs Wilson et al.[59] identified from the PDB. Data points were binned into one-percent-wide bins, and the mean RMS and standard deviation in each one-percent-bin was computed. Points more than two standard deviations above the mean were removed from the original 30,000 pair dataset (red crosses) and used to compose the full outlier set (green crosses), which ultimately consisted of 4,400 such pairs.

This set is intended as a comprehensive sample of protein flexibility in the PDB.

Workable outlier set

We ran the full outlier set through our protein morphing server.[6] We placed the resulting database of pre-processed PDB files, morph statistics, and movies, on the World Wide Web, organized by their scop fold classification. The new automated approach was able to process and generate several thousand new morph movies. As described below, the morph server acted as a filter, eliminating about 600 pairs in the full outlier set that corresponded to non-physical motions. Next, we applied the normal mode analysis described below on the successfully morphed pairs, to produce a set of about 3,800 motion pairs, the "outlier set". In this paper, we concentrate exclusively on this new "workable outlier set" data.

Manual set

In order to perform feature analysis, we classified two subsets of the workable outlier set (the "manual set" and the "extended set") into the classification schema of the Database of Macromolecular Motions[1] ("fragment," "domain," "subunit," "complex" on the basis of size and "hinge," "shear," "neither hinge nor shear," and "unclassifiable" on the basis of packing). Further details about this classification may be found in our previous paper.[1]

For the "manual set," we performed a database merge of the "outlier set" against the previously published set of manually classified motions in the Database of Macromolecular Motions,[1] the "1998 motions." The PDB identifiers in each motion pair in the outlier set were checked for matches against the PDB identifiers associated with the 1998 motions. When a match was found (meaning the protein had been manually classified), the motion pair was given the same classification as its constituent protein had been given in the database. Two hundred and forty-five motion pairs met this criterion and were classified accordingly. Classifications in this manual training are expected to be accurate. (There was, however, one issue in applying this merge: GroEL is classified both as a subunit and a fragment motion. Because the Morph server analyzes single domains, not entire subunits, the fragment classification was used in this isolated case.)

Extended set

To enlarge the training data for the supervised machine learning analysis, we constructed a second, larger training set (the "extended set"). For a variety of physical reasons, proteins sharing the same fold family generally share a similar motion classification—in particular, we have observed this in our manual surveys of motions.[1,6,60,66,67] Consequently, we constructed this set under the assumption that domains sharing a fold usually share a motion

classification. The outlier set is constructed in such a way that both pairs always belong to the same fold family. It was, therefore, necessary only to determine the scop fold classification[60,65] for each of the 245 motion pairs in manual training set and then assign the classification in the manual set to the entire scop fold family. Pairs in the outlier set belonging to this scop fold family then simply received the family's classification. In this way we identified a set of 1,670 motions, which we call the "extended training set." This set of classifications, although potentially less accurate than the manual training set, is still quite useful.

Preprocessing With Morph Server

We analyzed 3,814 proteins using this method from the full outlier set. Previously,[6] we modified the X-PLOR package[68] to homogenize the stored coordinates, a non-trivial problem.[69,70] Filling-in of missing non-hydrogen coordinates was necessary for the energy minimization subsystems to work robustly with a large number of PDB files and ensured consistent numbering of atoms so the PDB files for the starting and ending conformations had to be pre-processed ("homogenized") by the Morph Server.[6] Only pairs of protein conformations for which the Morph Server had successfully produced a movie were considered; this had the effect of filtering out pairs unlikely to involve a true motion, although no doubt some pairs that did not represent a true biological motion nevertheless did generate a plausible morph. The Morph Server also removes overall rotation and translation motions from the input structure.

High-Throughput Normal Mode Analysis of the Outlier Set

We used MMTK[71] to carry out normal mode analysis on the pre-processed PDB file pairs. The numerical Python module[72] made the linear algebra computations. A master Perl[73] script fed database information to the slave Python MMTK module. The results reported here were performed by computing the normal modes of the starting structures in each pair. Reversing the calculations by computing the normal modes of the ending structures did not appreciably alter the results.

Finding the normal modes themselves dominated the time and memory requirements of our analyses. In order to process the larger proteins in our database, we approximated each residue as a single, virtual atom centered at its C-α coordinate and selected the corresponding standard force field in MMTK.[71] This made the memory requirements of the normal mode analysis tractable on our systems. To further accelerate the computations, we restricted MMTK to compute only the twenty lowest-frequency normal modes.

We used the MMTK deformation force field model. In this model, the energy is computed as the difference between some displaced model and the experimental structure using the formula:

$$E_1 = \frac{1}{2}\sum_{j=1}^{N} k(\mathbf{R}_{ij}^{(0)})[\,|\mathbf{R}_{ij}^{(0)} + \mathbf{d}_i - \mathbf{d}_j| - |\mathbf{R}_{ij}^{(0)}|\,]^2 \qquad (2)$$

where k is a constant, $\mathbf{R}_{ij}^{(0)}$ is the vector from atom i to atom j in the experimental structure, $\mathbf{d}_i$ is the vector between the atom i in the displaced structure and the same atom in the ground-state experimental structure.

Each calculation averaged 20 seconds per protein pair on a 450-Mhz Pentium III processor with 0.7 Gigabytes of RAM running the Red Hat Linux operating system. An average analysis took about 100 Megabytes of memory to invert the matrix.

Theoretical Approach for Analysis of Normal Mode Statistics

We computed a number of key statistics on the normal modes (Table Ib), which we describe here.

Analysis of observed motion.

The lowest frequency normal modes determined by Normal Mode Analysis may be represented as an $m \times n$ matrix A, where m is three times the number of atoms in the system (one entry for each Cartesian axis), and n is the number of normal modes of interest. In this paper, n is twenty.

Imagine a vector $\bar{\mathbf{v}}$ of length n, specifying some interesting linear combination of normal modes. Then $\mathbf{A}\bar{\mathbf{v}}$ is a vector of length m, representing a trajectory of atoms. If we let the vectors c_i and c_f be the vectors of length m giving the positions of the m / 3 atoms in conformations C_i (starting) and C_f (ending), respectively. We determined these from our database of motion, which has such data, chiefly derived from experimental sources such as X-ray crystallography.

If we now define a new vector $b = c_f - c_i$, or the differences between the ending and starting positions of each of the atoms of the structure along all three Cartesian axes, then we can find optimal v so that

$$\mathbf{A}\bar{\mathbf{v}} = \bar{\mathbf{b}} \qquad (3)$$

In the normal case where dim $\bar{\mathbf{v}} < 3N - 6$, this represents an over-determined system of linear equations, and may be solved by an appropriate numerical technique for solving linear least squares, such as Single Value Decomposition (SVD).[74] In practice, this is a very quick calculation, nearly instantaneous to the user.

Analytic Measures

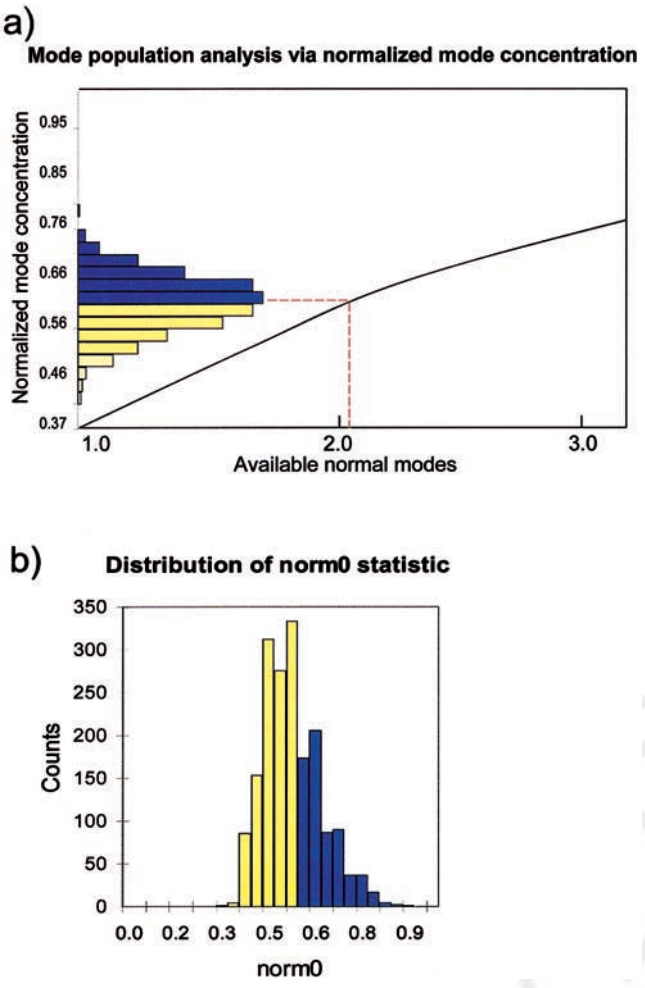
Overlap of each mode with direction of motion

For every motion pair, we computed the overlap of each normal mode against the vectors giving the differences between the structures corresponding to the motions. For one particular atom, we define the "overlap" O_{ij} as the cosine of the angle between the mode and the direction of motion,

$$O_{ij} \equiv \frac{\bar{\mathbf{b}}_i \cdot \bar{\mathbf{f}}_{ij}}{|\bar{\mathbf{b}}_i| \cdot |\bar{\mathbf{f}}_{ij}|}. \qquad (4)$$

In the above formula O_{ij} is represented as a normalized dot product between some reference vector $\bar{\mathbf{b}}_i$ (in this case, the displacement between the PDB structures of the motion pair in question) and $\bar{\mathbf{f}}_{ij}$, the jth normal mode displacement vector for the same atom.

tapraid5/7e-protein/7e-protein/7e-orig/7e1021-02a | austina | S=9 | 6/6/02 | 19:27 | Art: 01247R



COLOR

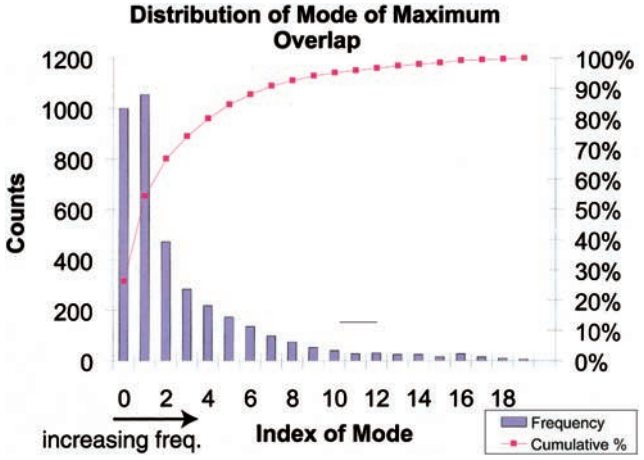
Fig. 2. **A:** Analysis of the normalized mode-concentration statistic to assess the normal modes populations. The center of the normalized mode-concentration histogram is traced to the number of available states (modes) using the Boltzmann logarithmic dependence relation. **B:** Histogram of norm0 statistic calculated over all entries in our database. The plot clearly shows that the large contributions (over 50%) from a single normal mode are not uncommon.

AQ: 16

For the ensemble of atoms in a structure, we can define "average overlap" O_j as the mean overlap averaged over all N atoms in the structure, i.e.,

$$O_j \equiv \frac{1}{n} \sum_{i=1}^{n} O_{ij}. \tag{5}$$

We can also calculate an average absolute value of the cosine $1/n\Sigma_{i=1}^{n}|O_{ij}|$, which provides a quantitative measure of the first-order overall deviation for a particular normal mode from the observed motion. The larger values of this quantity indicate that a given mode's atomic displacement vectors are more similar in directionality to the vectors giving the differences between the PDB files. The mode of "maximum overlap" is the mode with the greatest "absolute average overlap" and most matches the protein motion's directionality.



COLOR

Fig. 3. Our software places the twenty lowest-frequency normal modes in an array, thereby assigning each normal mode an index, from zero to nineteen. Increasing index numbers identify higher-frequency normal modes. We computed the overlap of each normal mode and recorded the index of the normal mode of greatest overlap. We plotted the number of times each index had greatest overlap in this histogram.

S-correlation

A means of quantifying the similarity of the displacement between the PDB structures and the normal mode displacement vectors can be also achieved by calculating the following quantity,

$$s = \sqrt{\sum_{j=1}^{n} j^2 O_j^2 - \left(\sum_{j=1}^{n} j O_j^2 \right)^2} \tag{6}$$

where O_j is defined as above. This formula, directly adapted from Hinsen's work[39] with a lowering of dimensionality, gives the s-correlation between the reference vector and the set of normal mode displacement vectors. This may be used to provide an overall quantitative measure of the similarity in directionality between the observed displacements and those of the various normal modes. Thus, the convention used to number the modes does not affect s-correlation in a meaningful way.

In the present work, we also utilize an interesting mathematical property of this statistic: its positive definite values imply that the displacement vectors from only the lowest two normal modes may coincide with the direction of the observed motion.

Mode concentration

Based on the fit of the modes to the observed motion, we calculate a number of statistics that show the degree to which the fit is dominated by a single mode. We define norm zero ("norm0") as simply the weight of the largest component (i.e., the largest value in the vector **v**), the one norm as the average component ("norm1"), and the two norm as simply the Euclidean mean ("norm2") of the component's weight.

All of these statistics give a measure of the degree to which the vector v is dominated by a single component. In somewhat more sophisticated fashion, we can measure this using information theory approaches.

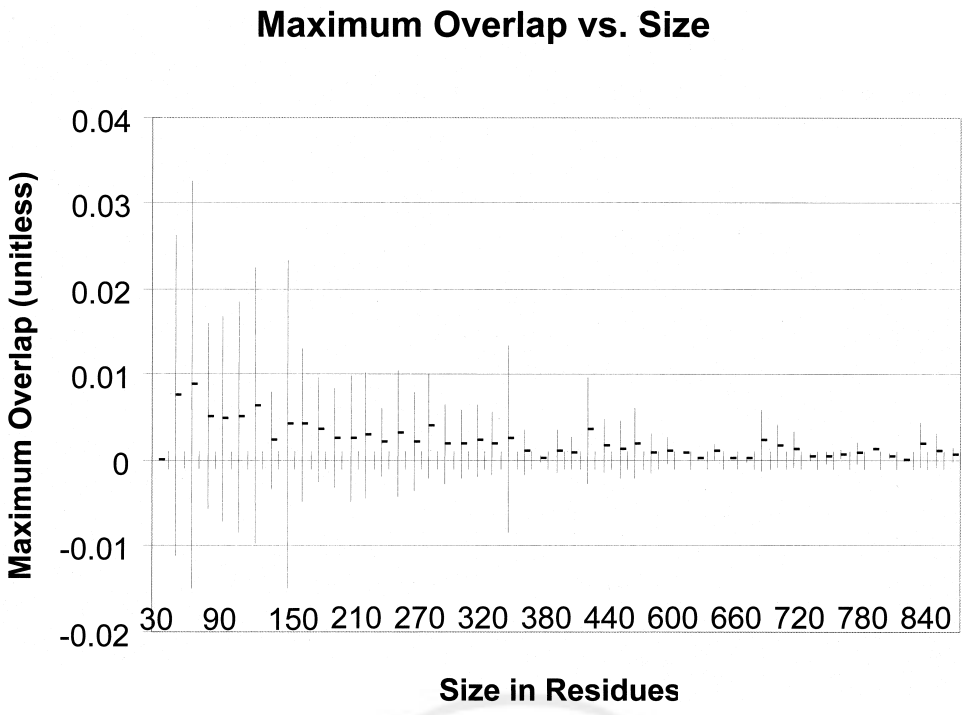


Fig. 4. Relationship between protein size and maximum overlap. To make the effect clearer, the y-values were binned into groups of 15 residues. The mean and standard deviation were computed for the values in each bin, with the results plotted. Each heavy horizontal bar indicates the mean in each bin, while the vertical bars indicate two standard deviations above and below the mean.

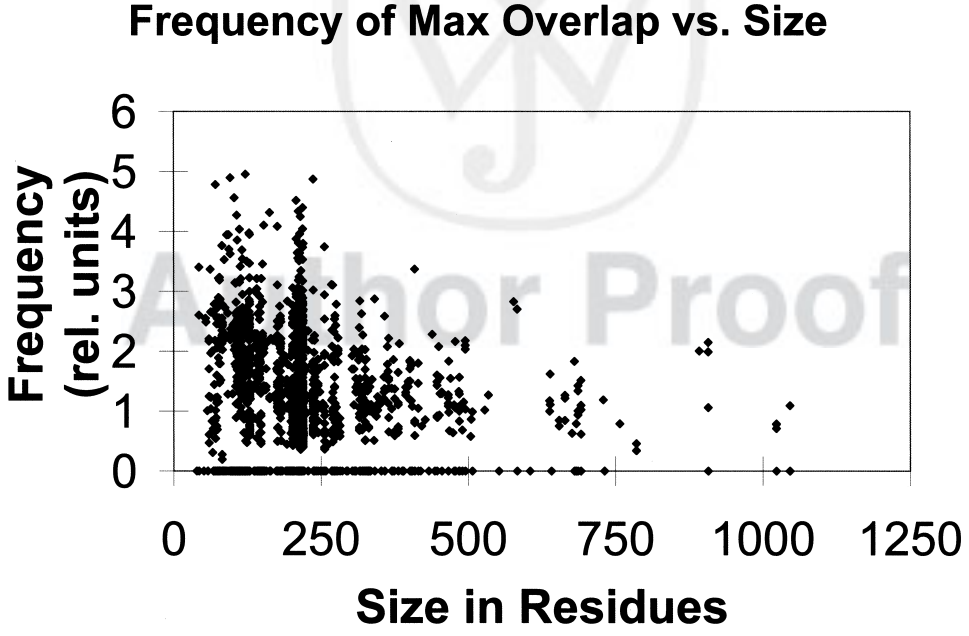


Fig. 5. Correlation between the frequency of the mode of maximum overlap and protein size.

In coding theory, information content is related to the negative entropy of a physical system. It specifies how much information is stored in a given set of numbers, and is typically used to compare the efficiencies of compression techniques. Therefore, once $\bar{\mathbf{v}}$ has been obtained, a statistic may be computed to summarize the information contained in the vector $\bar{\mathbf{v}}$:

$$I = \sum_{i=1}^{n} -|v_j|\ln|v_i| \tag{7}$$

TABLE II. Summary of New Statistics Added to Morph Server†

Key	No. of CAatoms	Residuals	Norm1	Norm2	Frequency	Ranking overlap	Maximum overlap
Mean	220	480	−0.001	540	3.1	2.7	0.0031
Std. dev.	110	660	0.051	360	0.89	3.6	0.005
Minimum	39	0.23	−0.14	15	4.2E–08	0	4.7E-5
Maximum	1,000	8,800	0.15	2,700	8.6	19	0.11
Median	210	330	0.00093	520	3.1	1	0.0017

†This table presents mean, standard deviation, minimum, maximum, and median values for the new statistics that were added to the database following normal mode analysis of approximately 3,800 motion pairs in the database. The statistics are defined in Table IB.

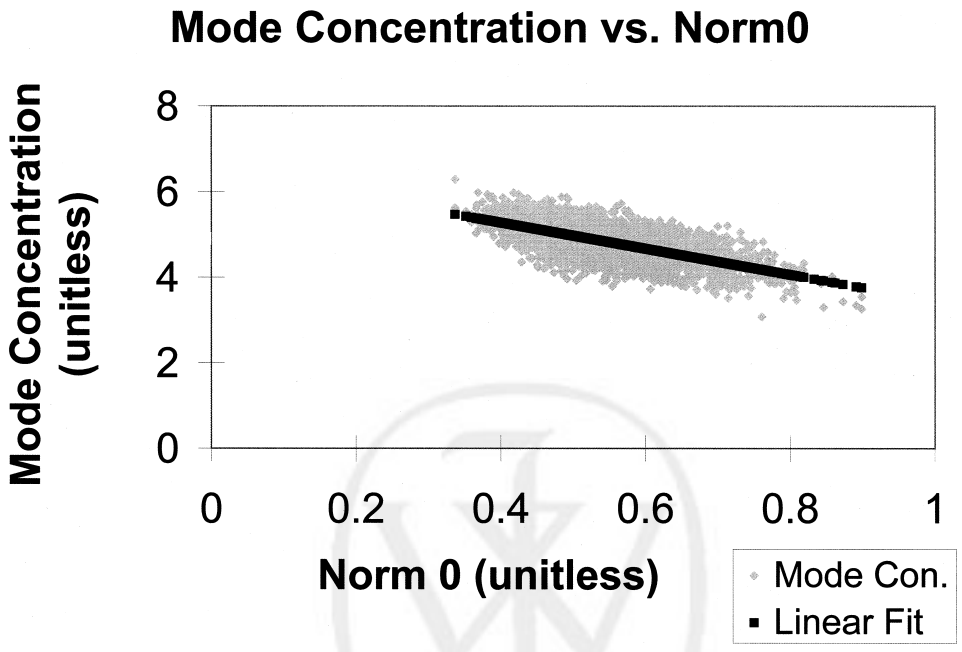


Fig. 6. Relationship between mode concentration and norm0 (concentration of motion in the mode with greatest concentration).

This statistic specifies how much movement is concentrated in any given mode, hence its name, "mode concentration."

We can normalize I to unity by dividing it over its maximal value, corresponding to the uniform movement distribution over all available modes, and obtain the "percentage mode concentration" statistic $\tilde{I}$, that specifies the degree to which a given motion is localized within a few modes relative to the uniform distribution (maximal disorder). As mentioned above, one can also directly relate information content (and, thus, also our normalized information content) to the well-known Boltzmann formula $S = k \ln N$ for the entropy (measure of the system disorder in statistical mechanics) expressed through the number of states N available to the system, i.e.,

$$\bar{I} \sim \ln N \tag{8}$$

The normalization ensures that $\bar{I}$ approaches zero if all movement is concentrated in only one normal mode ($N = 1$), whereas the value of $\bar{I} = 1$ corresponds to the even distribution of motion over all available normal modes (i.e., to the maximal value of I computed from Eq. (7)).

RESULTS

Application of These Statistics to the Outlier Dataset

Figures 2 through 5 illustrate some properties of the above statistics on the outlier dataset. F2-G5

Figure 2 shows distributions of the normalized mode concentration and norm0 statistics. Using the logarithmic dependence Eq.(8) of the normalized mode concentration with respect to the number of available modes, one can arrive at the number of most heavily involved modes. This would be the value of N, for which the value of $\bar{I}$ is most frequently observed. The observed peak in the normalized mode-concentration histogram at 0.6 [Fig. 2(A)] suggests that the actual direction of the motion lies most often along the direction of two modes. Analysis of norm0 histogram [Fig. 2(B)] further confirms this finding: the most commonly observed weight of the major contributing mode lies within the range 0.5–0.6 (i.e., there is usually one mode that dominates the motion fit) whereas the normal mode approximations with values of norm0 below 0.4 are quite rare (the latter would imply that there are usually more

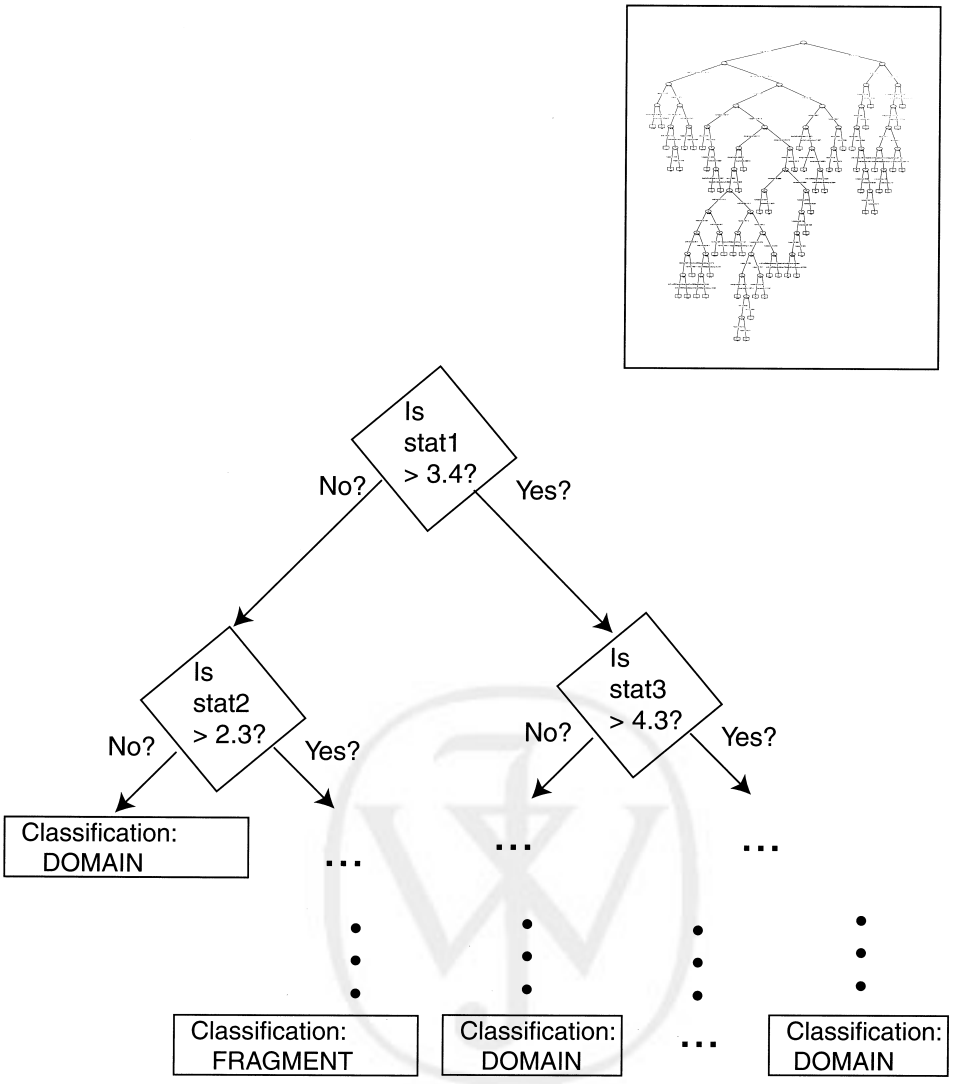


Fig. 7. Decision tree concepts. Two decision trees (not shown here) were generated by S-Plus (MathSoft, Inc.) using default parameters from the 245-element manual training set and the 1,670-element extend training set (defined in Table IA). These trees classify motions as "fragment," "domain," or "subunit." The decision tree associated with the extended training set defined an automatic classifier (implemented in Perl by examination of the tree) that produced the "classified set." This figure shows the conceptual operation of decision trees. At each node, the classifier chooses either the right or left branch, respectively, depending on whether or not the node's associated statistic is greater than the value associated with the node. **Inset:** Structure of an actual decision tree is shown in miniature. The classifier follows the decision tree until it reaches one of the terminal leaves, where a classification is made. A "training set" providing a set of examples and associated "correct" classifications is run through the S-Plus program, which generates a decision tree that can classify the training set correctly.

than two mostly contributing normal modes exist for each normal mode fit).

Figure 3 shows that most often the low-frequency modes tend to be the ones with maximum overlap with the actual direction of motion (Fig. 3). There is also a relationship between protein size (measured in number of residues), mode frequency, and maximum overlap (Figs. 4 and 5).

Protein size (measured in number of residues) is negatively correlated to maximum overlap (Fig. 4). Larger proteins have additional fragments that can be involved in a motion and, hence, additional degrees of freedom, decreasing the overlap between the tested normal modes and the observed motion. (An alternative explanation for this observation is that the various approximations used in normal modes approximation work less well for larger proteins.) Maximum overlap decreases with protein size, but the effect is not dramatic, so it should be possible to design a standard analysis that works well on proteins comparable to those in our database.

Increasing protein size (in residues) corresponds to modes of maximum overlap of decreasing frequency (Fig. 5). A standard analysis concerned with larger proteins may need to consider more low-frequency normal modes than would suffice for smaller proteins. It would be

TABLE III. Comparison of the Percentages and Absolute Counts of Domain, Fragment, and Subunit Motions in Each of the Classified, Extended, and Manual Training Sets†

	Predicted		Observed			
	Classified set		Extended set		Manual set	
Motion size	Count	Percent	Count	Percent	Count	Percent
Domain	2,165	95	1549	93	180	73
Fragment	94	4	107	6	50	20
Subunit	14	1	14	1	15	6
Totals	2,273	100	1670	100	245	100

†Definitions of the different sets in the header are given in the text as well as Table IA. "Count" gives the number of times the particular motion size classification (Domain, Fragment, and Subunit) occurs in that dataset. "Percent" is the percentage out of the total number ("Total") of domain, fragment, and subunit motions in the dataset. The two columns on the left for the auto-classified set ("count" and "percent") represent a prediction made by an auto-classifier; the remaining columns represent observations.

AQ: 14

AQ: 15

TABLE IV. †

Database statistic	Depth in tree built upon extended set	Depth in tree built upon manual set
Size of 2nd core	1	1
Trimmed RMS	3	2
Maximum CA movement	5	2
Number of atoms	4	3
Mode concentration	6	4
Energy of 2nd frame	6	4
Translation	4	5
Hinge rotation (degrees)	4	6
Number of hinges		6
Energy of 3rd frame		6
Norm0 (maximum value)	5	9
Energy of 9th frame	3	
Number of residues	5	
Frequency	5	
Residuals	6	
Norm1 (average norm)	6	
Rank of Norm0 mode	7	
Traditional RMS	8	
Norm2 (Euclidean norm)	8	
Energy of 4th frame	9	
Energy of 9th frame	9	
Energy of 8th frame	13	

†This table indicates the earliest depth of the supervised machine learning decision tree each statistic first occurs, thus quantifying the relevance of each statistic to the particular motion property at hand ("fragment," "domain," or "subunit" motion, in this case).

desirable, given a protein of specific size, to deduce a frequency cut-off value, above which normal modes could be expected to be less useful in an analysis of motion. Analyses of individual proteins in the literature support the existence of such a cutoff[46,75] showing a slight dependency on the force field used. Our results show that it is possible to determine such a cut-off frequency statistically from our database (Fig. 5) and thereby empirically deduce a reasonable number of normal modes to use in a given type of analysis. Researchers using an identical force field to the one used in this study may consult Figure 5 directly to determine the appropriate cut-off for their particular protein; researchers using slightly different force fields or dynamical methods may wish to obtain access to the database to compute a cut-off value appropriate for their specific dynamical analysis.

Validation of Mode Concentration With Feature Extraction Techniques

The physical and information theory basis of the mode concentration statistic suggested it might be useful in classification problems. Subsequent analysis via machine learning techniques (below) supports this.

Artificial intelligence feature analysis techniques, particularly supervised machine learning, provide one way of validating the usefulness of our mode concentration statistic. In general, the concept of supervised machine learning is that the system is "taught" to classify a given set of inputs by being given a "training set" that matches a sample set of inputs to a correct set of outputs.[76]

As described above, we created the manual and extended data sets as training sets to perform feature analysis. Using supervised machine learning techniques,[76,77] we constructed two decision trees in S-Plus (MathSoft, Inc.) using the software's default parameters[77–79] (one for each of the two training sets) to classify the statistics, including the new ones (Table II), in the morph server.[6] The use of S-Plus to construct decision trees from a specific training data set is a straightforward operation. T2

Decision trees, a form of supervised machine learning, attempt to partition the examples in the training set based on the values of individual statistics (Fig. 7). In the actual decision tree, each statistic used in the classification decision appears in at least one branch junction. Features more relevant to the classification problem tend to appear earlier in the decision-making process, corresponding to a higher-level branch in the trees. By recording the depth any statistic first appears in, decision trees may be used for feature analysis (Table III). Mode concentration ranks prominently with a low depth, indicating that it appears high in the tree and is, therefore, useful for classifying motions. AQ: 3 F6-F7 AQ: 4 T3

COLOR

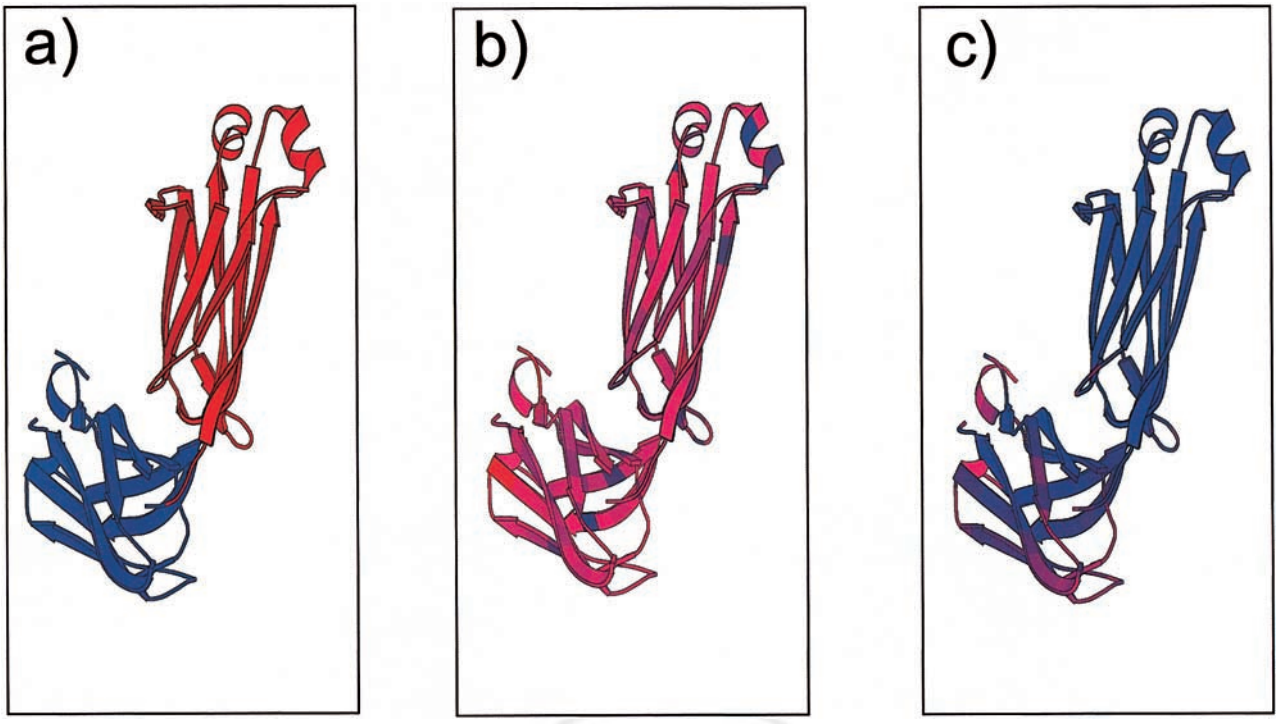


Fig. 8. Output of new set of Web tools associated with normal mode analysis that the user may request on any protein for which a PDB structure file is available. The URL for this server is http://www.molmovdb.org; these features may be accessed by browsing to a specific movie and selecting one of these analyses from the menu. **B:** Performs a normal mode flexibility analysis on the structure. Regions that are more flexible are colored in red, while less flexible regions are colored in blue. **A:** Similar information, using experimental temperature factors supplied in the PDB file, if available. **C:** The parts of the protein that actually move, as calculated from comparison of the starting and ending PDB structures for the motion. Areas that move are colored in red, while areas that remain stationary are colored in blue. The user may compare these three panels to deduce structural information. Hinge locations involved in the motion may be deduced, as these are highly flexible regions (as identified by A and B) located near the moving domains (show in red in C). The specific protein example shown is that of an immunoglobulin elbow joint motion (morph ID d2fb4l1-d1afvl1).

Using appropriate, simple physical and mathematical concepts (normal mode analysis, singular value decomposition), we postulated several statistics (mode concentration and the various analytic norm measures) and confirmed our initial hypotheses using artificial intelligence techniques. These culled the morph server's[6] output of 36 physically-motivated statistics down to a set of nine "essential" statistics that proved most useful in this particular classification problem (Table IV), which agree roughly with our own sense of the statistics most related to motion size. Similar databases of heterogeneous biological statistics may be "distilled" from a larger body of experimental data with these and similar techniques. In this case, the automatic classification features of the decision trees are only a side benefit. Feature analysis confirmed our earlier intuition that mode concentration can be useful for classifying motions.

Depending on the supervised machine learning technique used (decision trees), larger training sets can sometimes produce a more accurate automatic classifier than a smaller classifier. For this reason, it is possible that an automatic classifier produced from the larger extended training set may classify more accurately than one produced from the smaller, more accurate manual set, although this may seem counterintuitive. Comparing the results produced by the manual and the extended training sets thus will serve as a useful check.

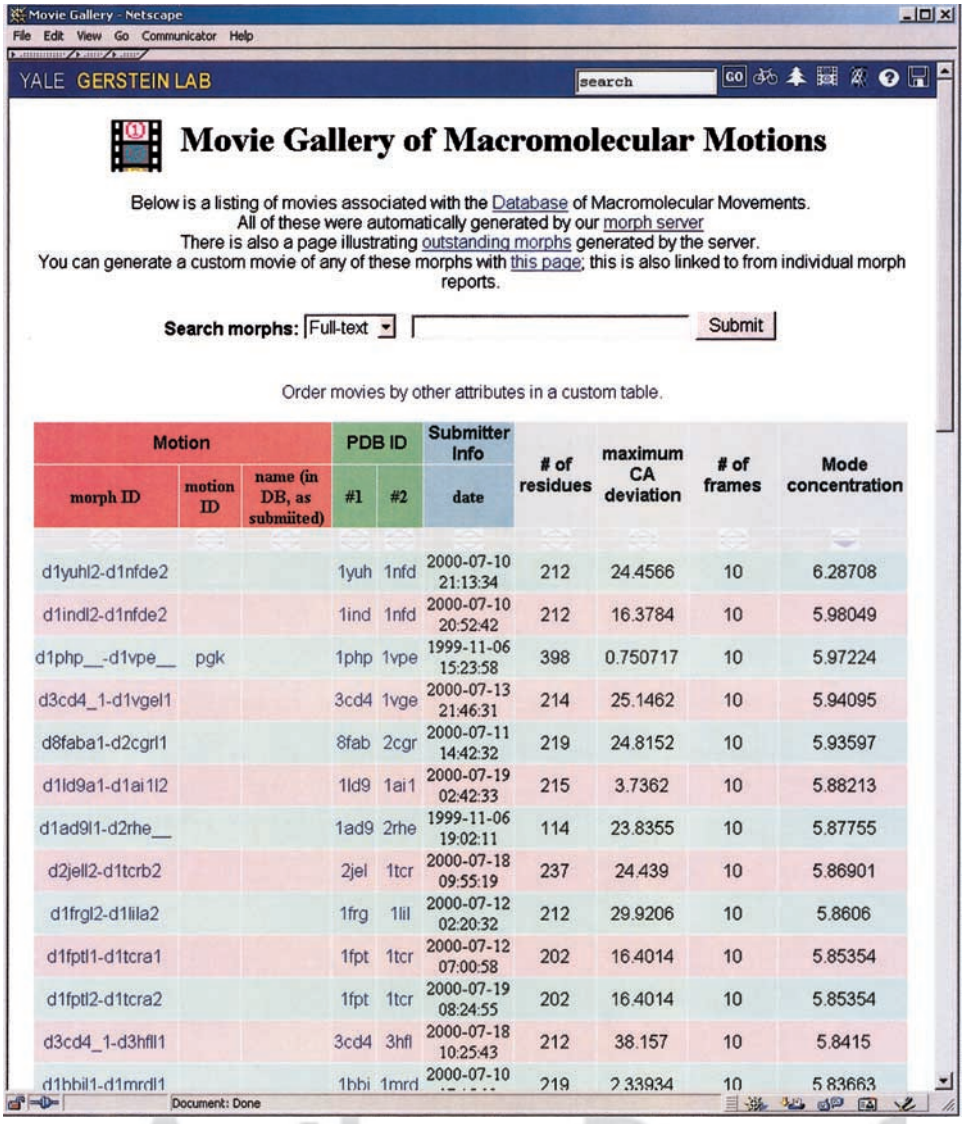
Web and Database Integration

We used the results of our decision tree analysis (Table III) to improve the ordering and presentation of statistics in Macromolecular Motions. Database web reports (http://molmovdb.org). In addition, a new web tool (Fig. 8) on this site graphically depicts output from the normal mode analysis as well as older experimental information.

The new data from normal mode analysis have been integrated into both the Macromolecular Motions Database and the Partslist Database (http://partslist.org/) as well.[80] This allows comparison by fold of motion and other data by a number of techniques, including regression analysis. Interactive users can test a number of statistics for correlation against the new data, as well as identify outlying folds that do not maintain the normal regression pattern by mouse over. Figure 9 gives a screen shot of motions ranked by average mode concentration in the Movie Gallery page of the Macromolecular Motions Database, which will show the animation of the corresponding motion on click.

DISCUSSION

Comprehensive structural studies within a database framework, such as the one described here, can complement more traditional computational studies of single molecules in a number of ways. The most immediate



Motion			PDB ID		Submitter Info	# of residues	maximum CA deviation	# of frames	Mode concentration
morph ID	motion ID	name (in DB, as submiited)	#1	#2	date				
d1yuhl2-d1nfde2			1yuh	1nfd	2000-07-10 21:13:34	212	24.4566	10	6.28708
d1indl2-d1nfde2			1ind	1nfd	2000-07-10 20:52:42	212	16.3784	10	5.98049
d1php__-d1vpe__	pgk		1php	1vpe	1999-11-06 15:23:58	398	0.750717	10	5.97224
d3cd4_1-d1vgel1			3cd4	1vge	2000-07-13 21:46:31	214	25.1462	10	5.94095
d8faba1-d2cgrl1			8fab	2cgr	2000-07-11 14:42:32	219	24.8152	10	5.93597
d1ld9a1-d1ai1l2			1ld9	1ai1	2000-07-19 02:42:33	215	3.7362	10	5.88213
d1ad9l1-d2rhe__			1ad9	2rhe	1999-11-06 19:02:11	114	23.8355	10	5.87755
d2jell2-d1tcrb2			2jel	1tcr	2000-07-18 09:55:19	237	24.439	10	5.86901
d1frgl2-d1lila2			1frg	1lil	2000-07-12 02:20:32	212	29.9206	10	5.8606
d1fptl1-d1tcra1			1fpt	1tcr	2000-07-12 07:00:58	202	16.4014	10	5.85354
d1fptl2-d1tcra2			1fpt	1tcr	2000-07-19 08:24:55	202	16.4014	10	5.85354
d3cd4_1-d3hfll1			3cd4	3hfl	2000-07-18 10:25:43	212	38.157	10	5.8415
d1hbil1-d1mrdl1			1hbi	1mrd	2000-07-10	219	2.33934	10	5.83663

COLOR

Fig. 9. Screenshot of the Movie Gallery web page. This shows a Movie Gallery page in the Macromolecular Motions Database that ranks different motions according to the average mode concentration.

benefit is that a database study makes more data available to researchers, and can sometimes make more general statements about trends and patterns in the results than would be possible from similar studies on a smaller sample of macromolecules. A disadvantage of the type of database study performed here is that they require greater computational resources than equivalent studies on single macromolecules. Also, the implementation of automatic methods to handle a large class of macromolecular may require somewhat greater algorithmic sophistication since steps requiring manual processing are less desirable when dealing with a large number of structures.

Researchers who have developed their own, novel computational structural studies may expand their computations from analyses of single molecules to a comprehensive study of an entire structural database, such as the Database of Macromolecular Motions. The results of such structural studies constitute databases in their own right. Artificial intelligence techniques can then be applied to such derived databases to append additional, useful statistics, "distill" a derived database down to a set of "essential" statistics, as well as construct automatic classifiers. This has obvious practical applications; e.g., pharmaceutical companies might mine existing biological databases and apply existing or new algorithmic techniques (e.g., variants on normal mode analysis) to generate derived databases describing potential drug targets within a statistical framework. Artificial intelligence techniques can be used to extract key features and empirically assess the validity of new statistical models.

CONCLUSIONS

We have developed a framework that allows for a statistical study, in combination with our Database of Macromolecular Motions, of the importance of normal mode vibrations in biologically significant macromolecular

motions. A statistic calculated from our analysis of normal mode displacements, mode concentration, is corroborated by feature selection as a useful statistic in classification. Feature selection techniques can be used to "summarize" databases of experimentally derived statistics into an especially salient set of "essential" statistics.

Examining the relationship between the aggregate directionality of the normal modes and structures' conformational change through a statistic such as mode concentration can be used to classify the motion ("fragment," "domain," or "subunit"). Normal modes have already been used[58] to identify dynamic protein domains. An analysis of the distribution of low-frequency normal mode trajectories should provide information about the type of protein motion and size of the domains involved in the motion. Our data empirically support earlier results[46] that analysis of only a small number of low-frequency modes should suffice for qualitative analysis of protein dynamics. The database can also be used to determine statistically the cut-off for normal modes computed using different force fields.

In addition to being made available through the Macromolecular Motions Database, our new data sets are integrated into the external Partslist database.[80] We have provided additional web tools associated with this paper that allow molecular biologists to perform flexibility analysis on structures with putative motions, thereby identify key residues involved in the motion, and compare the results with similar analysis on the over 4,000 new motions now available in the database.

ACKNOWLEDGMENTS

We thank Dr. Yuval Kluger for his help with machine learning, and Dr. Jiang Qian for the Partslist integration of the data. Numerous people have also either contributed entries or information to the database and morph server or have given us feedback on what the user community wants. The authors also thank Informix Software, Inc., for providing a grant of its database software.

REFERENCES

1. Gerstein M, Krebs W. A database of macromolecular motions. Nucleic Acids Res 1998;26:4280–4290.
2. Debrunner PG, Frauenfelder H. Dynamics of proteins. Annual Rev of Phys Chem 1982;33:283.
3. Lipscomb WN. Acceleration of reactions by enzymes. Accounts Chem Res 1982;15:232.
4. Gerstein MB, Jansen R, Johnson T, Park B, Krebs W. Studying macromolecular motions in a database framework: from structure to sequence. In Thrope MF, M. F. Thorpe and Duxbury PM., editors. Rigidity theory and applications 1999: Kluwer Academic/Plenum press. p. 401–442. AQ: 5
5. Gerstein M. A protein motions database. PDB Q Newsletter 1995;73:2 (July).
6. Krebs WG, Gerstein M. The morph server: a standardized system for analyzing and visualizing macromolecular motions in a database framework. Nucleic Acids Res 2000;28:1665–1675.
7. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
8. Isralewitz B, Gao M, Schulten K. Steered molecular dynamics and mechanical functions of proteins. Curr Opin Struct Biol 2001;11:224–230.
9. Young M, Kirshenbaum K, Dill KA, Highsmith S. Predicting conformational switches in proteins. Protein Sci 1999;8:1752–1764.
10. Shaknovich R, Shue G, Kohtz DS. Conformational activation of a basic helix-loop-helix protein (MyoD1) by the C-terminal region of murine HSP90 (HSP84). Mol Cell Biol 1992;12:5059–5068.
11. Dixon MM, Nicholson H, Shewchuk L, Baase WA, Matthews BW. Structure of a hinge-bending bacteriophage T4 lysozyme mutant, Ile3→Pro. J Mol Biol 1992;227:917–933.
12. Xu Z, Sigler PB. GroEL/GroES: structure and function of a two-stroke folding machine. J Struct Biol 1998;124:129–141.
13. Chik JK, Lindberg U, Schutt CE. The structure of an open state of beta-actin at 2.65 A resolution. J Mol Biol 1996;263:607–623.
14. Oka T, Yagi N, Fujisawa T, Kamikubo H, Tokunaga F, Kataoka M. Time-resolved x-ray diffraction reveals multiple conformations in the M- N transition of the bacteriorhodopsin photocycle. Proc Natl Acad Sci USA 2000;97:14278–14282.
15. Genick UK, Borgstahl GE, Ng K, Ren Z, Pradervand C, Burke PM, Srajer V, Teng TY, Schildkamp W, McRee DE, Moffat K, Getzoff ED. Structure of a protein photocycle intermediate by millisecond time- resolved crystallography. Science 1997;275:1471–1475.
16. Schlichting I, Almo SC, Rapp G, Wilson K, Petratos K, Lentfer A, Wittinghofer A, Kabsch W, Pai EF, Petsko GA, Goody RS. Time-resolved X-ray crystallographic study of the conformational change in Ha-Ras p21 protein on GTP hydrolysis. Nature 1990;345:309.
17. Volkman BF, Lipson D, Wemmer DE, Kern D. Two-state allosteric behavior in a single-domain signaling protein. Science 2001;291:2429–2433.
18. Tsai J, Levitt M, Baker D. Hierarchy of structure loss in MD simulations of src SH3 domain unfolding. J Mol Biol 1999;291:215–225.
19. Tang YZ, Chen WZ, Wang CX. Molecular dynamics simulations of the gramicidin A- dimyristoylphosphatidylcholine system with an ion in the channel pore region. Eur Biophys J 2000;29:523–534.
20. Van Belle D, De Maria L, Iurcu G, Wodak SJ. Pathways of ligand clearance in acetylcholinesterase by multiple copy sampling. J Mol Biol 2000;298:705–726.
21. Wlodek ST, Shen T, McCammon JA. Electrostatic steering of substrate to acetylcholinesterase: analysis of field fluctuations. Biopolymers 2000;53:265–271.
22. Daggett V, Levitt M. Realistic simulations of native-protein dynamics in solution and beyond. Annu Rev Biophys Biomol Struct 1993;22:353–380.
23. Berneche S, Roux B. Molecular dynamics of the KesA K(+) channel in a bilayer membrane. Biophys J 2000;78:2900–2917.
24. Gilson MK, Straatsma TP, McCammon JA, Ripoll DR, Faerman CH, Axelsen PH, Silman I, Sussman JL. Open "back door" in a molecular dynamics simulation of acetylcholinesterase. Science 1994;263:1276–1278.
25. Wriggers W, Schulten K. Investigating a back door mechanism of actin phosphate release by steered molecular dynamics. Proteins 1999;35:262–273.
26. van Aalten DM, Conn DA, de Groot BL, Berendsen HJ, Findlay JB, Amadei A. Protein dynamics derived from clusters of crystal structures. Biophys J 1997;73:2891–2896.
27. Wilson EB, Decius JC, Cross PC. Molecular vibrations. New York: McGraw-Hill. 1955. AQ: 6
28. Levy RM. Computer simulations of macromolecular dynamics: models for vibrational spectroscopy and X-ray refinement. Ann N Y Acad Sci 1986;482:24–43.
29. Levitt M, Sander C, Stern PS. Protein normal-mode dynamics; trypsin inhibitor, crambin, ribonuclease, and lysozyme. J Mol Biol 1985;181:423–447.
30. van der Spoel D, de Groot BL, Hayward S, Berendsen HJ, Vogel HJ. Bending of the calmodulin central helix: a theoretical study. Protein Sci 1996;5:2044–2053.
31. Ma J, Sigler PB, Xu Z, Karplus M. A dynamic model for the allosteric mechanism of GroEL. J Mol Biol 2000;302:303–313.
32. Brooks B, Karplus M, Normal modes for specific motions of macromolecules: Application to the hinge-bending mode of lysozyme. Proc Natl Acad Sci USA 1985;82:4995–4999.
33. Duncan BS, Olson AJ. Approximation and visualization of large-scale motion of protein surfaces. J Mol Graph 1995;13:250–257.
34. Hinsen K, Thomas A, Field MJ. Analysis of domain motions in large proteins. Proteins 1999;34:369–382.
35. Miller DW, Agard DA. Enzyme specificity under dynamic control: a normal mode analysis of alpha-lytic protease. J Mol Biol 1999;286:267–278.
36. Thomas A, Hinsen K, Field MJ, Perahia D. Tertiary and quater-

nary conformational changes in aspartate transcarbamylase: a normal mode study. Proteins 1999;34:96–112.
37. Thomas A, Field MJ, Perahia D. Analysis of the low-frequency normal modes of the R state of aspartate transcarbamylase and a comparison with the T state modes. J Mol Biol 1996;261:490–506.
38. Thomas A, Field MJ, Mouawad L, Perahia D. Analysis of the low frequency normal modes of the T-state of aspartate transcarbamylase. J Mol Biol 1996;257:1070–1087.
39. Hinsen K. Analysis of domain motions by approximate normal mode calculations. Proteins 1998;33:417–429.
40. Marques O, Sanejouand YH. Hinge-bending motion in citrate synthase arising from normal mode calculations. Proteins 1995;23: 557–560.
41. Smith JC, Cusack S, Pezzeca U, Brooks B, Karplus M. Inelastic neutron scattering analysis of low frequency motion in proteins: a normal mode study of the bovine pancreatic trypsin inhibitor. J Chem Phys 1986;85:3636–3654.
42. Smith J, Cusack S, Tidor B, Karplus M. Inelastic neutron scattering analysis of low-frequency motions in proteins: harmonic and damped harmonic models of bovine pancreatic trypsin inhibitor. J Chem Phys 1990;93:2974–2991.
43. Noguity T, Go N. Collective variable description of small-amplitude conformational fluctuations in a globular protein. Nature 1982;296:776.
44. Brooks B, Karplus M. Harmonic dynamics of proteins: Normal modes and fluctuations in bovine pancreatic trypsin inhibitor. Proc Natl Acad Sci USA 1983;80:6571.
45. Go N, Noguti T, Nishikawa T. Dynamics of a small globular protein, in terms of low-frequency vibrational modes. Proc Natl Acad Sci USA 1983;80:3696.
46. Levy R, Srinivasan A, Olson W, McCammon J. Quasi-harmonic method for studying very low frequency modes in proteins. Biopolymers 1984;23:1099–1112.
47. Levy RM, Rojas OdlL, Friesner RA. Quasi-harmonic method for calculating vibrational spectra from classical simulations on multidimensional anharmonic potential surfaces. J Phys Chem 1984;88:4233.
48. Henry ER. Molecular dynamics simulations of cooling in laser-excited heme proteins. Proc Natl Acad Sci USA 1986;83:8982–8986.
49. Gibrat JF. Normal mode analysis of human lysozyme: study of the relative motion of the two domains and characterization of the harmonic motion. Proteins 1990;8:258–279.
50. Cusack S. Temperature dependence of the low frequency dynamics of myoglobin. Measurement of the vibrational frequency distribution by inelastic neutron scattering. Biophy J 1990;58:243–251.
51. Ahn JS, Kanematsu Y, Enomoto M, Kushida T. Determination of weighted density states of vibrational modes in Zn-substituted myoglobin. Chem Phys Lett 1993;215:336–340.
52. Alden R, Schneebeck M, Ondrias M, Courtney S, Friedman J. Mode-specific relaxation dynamics of photoexcited Fe(II) protoporphyrin IX in hemoglobin. J Raman Spectrosc 1992;23:569–574.
53. Miller RJD. Energetics and dynamics of deterministic protein motion. Acc Chem Res 1994;27:145–150.
54. Durand P, Trinquier G, Sanejouand Y-H. A new approach for determining low-frequency normal modes in macromolecules. Biopolymers 1994;34:759–771.
55. Perahia D, Mouawad L. Computation of low-frequency normal modes in macromolecules: improvements to the method of diagonalization in a mixed basis and application to hemoglobin. Comput Chem 1995;19:241–246.
56. Amadei A, Linssen AB, Berendsen HJ. Essential dynamics of proteins. Proteins 1993;17:412–425.
57. de Groot BL, Vriend G, Berendsen HJ. Conformational changes in the chaperonin GroEL: new insights into the allosteric mechanism. J Mol Biol 1999;286:1241–1249.
58. Hayward S, Kitao A, Berendsen HJ. Model-free methods of analyzing domain motions in proteins from simulation: a comparison of normal mode analysis and molecular dynamics simulation of lysozyme. Proteins 1997;27:425–437.
59. Wilson CA, Kreychman J, Gerstein M. Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores [in process citation]. J Mol Biol 2000;297:233–249. **AQ: 7**
60. Hubbard TJP, Murzin AG, Brenner SE, Chothia C. SCOP: a structural classification of proteins database. Nucleic Acids Res 1997;25:236–239.
61. Brenner S, Chothia C, Hubbard TJP, Murzin AG. Understanding protein structure: using SCOP for fold interpretation. Methods Enzymol 1996;266:635–642.
62. Dubchak I, Muchnik I, Kim SH. Protein folding class predictor for SCOP: approach based on global descriptors [in process citation]. **AQ: 7**
Ismb 1997;5:104–107. **AQ: 8**
63. Gerstein M, Levitt M. Large-scale application of a structural alignment method to the SCOP classification of proteins: objective assessment of alignability. J Mol Biol 1997; **AQ: 9**
64. Gerstein M, Levitt M. Comprehensive assessment of automatic structural alignment against a manual standard, the Scop classification of proteins. Protein Sci 1998;7:445–456.
65. Murzin A, Brenner SE, Hubbard T, Chothia C. SCOP: A structural classification of proteins for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.
66. Janin J, Wodak S. Structural domains in proteins and their role in the dynamics of protein function. Prog Biophys Mol Biol 1983;42: 21–78.
67. Gerstein M, Lesk AM, Chothia C. Structural mechanisms for domain movements. Biochemistry 1994;33:6739–6749.
68. Brünger AT. X-PLOR 3.1, A System for X-ray crystallography and NMR. New Havens Yale University Press. 1993. **AQ: 10**
69. Hogue CW, Ohkawa H, Bryant SH. A dynamic look at structures: WWW-Entrez and the molecular modeling database. Trends Biochem Sci 1996;21:226–229.
70. Ohkawa H, Ostell J, Bryant S. MMDB: an ASN.1 specification for macromolecular structure. Ismb 1995;3:259–267.
71. Hinsen K. The molecular modeling toolkit: a new approach to molecular simulations. J Comp Chem 2000;79–85.
72. Ascher D, Dubois PF, Hinsen K, Hugunin J, Oliphant T. Numerical Python. Livermore, CA: Lawrence Livermore National Laboratory. 2000.
73. Wall L, Christiansen D, Schwartz R. Programming Perl. O'Reilly and Associates. 1996. **AQ: 11**
74. Press WH, Flannery BP, Teukolsky SA, Vetterling WT. Numerical recipes in C. Cambridge, MA: Cambridge University Press. **AQ: 12**
75. Levy R, Perahia D, Karplus M. Molecular dynamics of an ff-helical polypeptide: temperature dependance and deviation from harmonic behavior. Proc Natl Acad Sci USA 1982;79:1346–1350.
76. Christendat D, Yee A, Dharamsi A, Kluger Y, Savchenko A, Cort JR, Booth V, Mackereth CD, Saridakis V, Ekiel I, Kozlov G, Maxwell KL, Wu N, McIntosh LP, Gehring K, Kennedy MA, Davidson AR, Pai EF, Gerstein M, Edwards AM, Arrowsmith CH. Structural proteomics of an archaeon. Nat Struct Biol 2000;7:903–909.
77. Ripley BD. Pattern recognition and neural networks. Cambridge, MA: Cambridge University Press. 1996. **AQ: 12**
78. Venables WN, Ripley BD. Modern applied statistics with S-PLUS. New York: Springer. 1997. **AQ: 13**
79. Krause A, Olson M. The basics of S and S-Plus. New York: Springer, 2000. **AQ: 13**
80. Qian J, Stenger B, Wilson CA, Lin J, Jansen R, Teichmann SA, Park J, Krebs W, Yu H, Alexandrov V, Echols N, Gerstein M. PartsList: a web-based system for dynamically ranking protein folds based on disparate attributes, including whole-genome expression and interaction information. Nucleic Acids Res 2001;29: 1750–1764.

AQ1: Suitable running title? If not, please provide one ≤45 characters including spaces

AQ2: Change OK for clarity?

AQ3: Please cite Fig. 6 in "order" in text

AQ4: Change correct for sense?

AQ5: Same author 2 × ?

AQ6: Provide # of pages

AQ7: Delete "in process citation?"

AQ8: Correct journal abbreviation?

AQ9: update with vol#, pages

AQ10: Provide # of pages

AQ11: City?

AQ12: Correct city/state?

AQ13: NY correct?

AQ14: Pls verify table title

AQ15: Provide table title

AQ16: Uppercase or lowercase A + B?