Analysis of Protein Structures Reveals Regions of Rare Backbone Conformation at Functional Sites

ARTICLE *in* PROTEINS STRUCTURE FUNCTION AND BIOINFORMATICS · DECEMBER 2003

Impact Factor: 2.63 · DOI: 10.1002/prot.10484 · Source: PubMed

4 AUTHORS, INCLUDING:

Ivan Torshin
Lomonosov Moscow State University
54 PUBLICATIONS **364** CITATIONS

SEE PROFILE

PROTEINS: Structure, Function, and Genetics 53:872–879 (2003)

Analysis of Protein Structures Reveals Regions of Rare Backbone Conformation at Functional Sites

John M. Petock,[1] Ivan Y. Torshin,[1] Irene T. Weber,[1,2,*] and Robert W. Harrison[1,3]
[1]*Department of Biology, Georgia State University, Atlanta, Georgia 30303*
[2]*Department of Chemistry, Georgia State University, Atlanta, Georgia 30303*
[3]*Department of Computer Science, Georgia State University, Atlanta, Georgia 30303*

ABSTRACT **Regions of rare conformation were located in 300 protein crystal structures representing seven major protein folds. A distance matrix algorithm was used to search rapidly for 9-residue fragments of rare backbone conformation using a comparison to a relational database of encoded fragments derived from the database of nonredundant structures. Rare fragments were found in 61% of the analyzed protein structures. Detailed analysis was performed for 78 proteins of different folds. The rare fragments were located near functional sites in 72% of the protein structures. The rare fragments often formed parts of ligand-binding sites (59%), protein-protein interfaces (8%), and domain-domain contacts (5%). Of the remaining structures, 5% had a high average B-factor or high local B-factors. Statistical analysis suggests that the association between ligands and rare regions does not occur by chance alone. The present study is likely to underestimate the number of functional sites, because not all analyzed protein structures contained a ligand. The results suggest that rapid searches for regions with rare local backbone conformations can assist in prediction of functional sites in novel proteins. Proteins 2003;53:872–879.**

Key words: protein function prediction; ligand-binding sites; domain-domain contacts; subunit interfaces; local conformation; structural genomics

INTRODUCTION

The number of protein structures in the Protein Databank[1] (PDB) is growing rapidly, and the development of high-throughput methods will facilitate this growth. Many structural genomics groups target proteins of unknown function in order to gain more information about their function. Results from one study are described in Eisenstein et al.[2] Many studies have analyzed the properties of functional sites in protein structures.[3] Functional sites can be identified using sequence similarities and representative structures.[4] In the absence of sequence similarities with proteins of known function, the atomic structure can help define function when the protein fold is characteristic for a particular function. Two examples of functional characterization of proteins based on their structural fold are YecO from *Haemophilus influenzae*[5] and MJ1247 protein from *Methanococcus jannaschii*.[6] In the absence of a characteristic fold the novel protein structure can be analyzed for a potential ligand-binding site. This method was successful for *Bacillus subtilis* protein Maf.[7] Functionally important sites are generally located on the molecular surface of the protein. Active sites can be identified by molecular shape comparisons for similar proteins.[8] However, it can be difficult to determine their exact location on the surface in the absence of an obvious cleft or indentation suggestive of a ligand-binding site.[3] Therefore, it is important to develop new analytic methods that quickly identify unusual or important features in new protein structures that relate to the function.

Here, fragments of unusual backbone conformation have been identified in protein structures from different fold classes and analyzed in relation to observed functional sites. Local regions with regular secondary structure such as alpha helices or beta strands occur very frequently, whereas regions of irregular conformation are less common and may be only rarely observed. Typically, regions of unusual conformation are identified in order to correct potential errors during structure determination. One example is the program WHAT_CHECK, which finds 5-residue regions with rare conformations.[9] Such rare local conformations of proteins will occur by chance, or else they can arise because of errors in structure determination, low-resolution of the structure, or a local disorder. Alternatively, it is possible that some regions with rare conformations are critical for molecular recognition and hence can indicate ligand-binding sites, protein-protein interfaces, or other functionally important sites. For example, it is unusual to find sterically strained backbone conformation in proteins, but these strained regions were generally associated with the function.[10] However, no previous studies have addressed the functional importance of the backbone regions with rare conformations.

Grant sponsor: National Cancer Institute; Grant number: CA76259. Grant sponsor: Georgia State University Research Program Enhancement grant to J.M.P.

*Correspondence to: Irene T. Weber, Georgia State University, Biology Department, PO Box 4010, Atlanta, GA 30302-4010. E-mail: iweber@gsu.edu

Received 2 December 2002; Revised 12 March and 17 April 2003; Accepted 17 April 2003

Distance matrices can be used to find similar protein folds; one application to locate common substructures in proteins is described in Lesk.[11] This method facilitates rapid searches of large databases of protein structures. Here, a distance-matrix algorithm for finding 9-residue fragments of rare backbone conformation was used to analyze 300 protein structures representing the major fold classes in SCOP.[12] The locations of fragments with a rare conformation were analyzed in relation to functional sites observed in the protein structures.

MATERIALS AND METHODS

Algorithm Design

The aim was to select an algorithm for efficient searches in a large relational database of protein fragments. A problem of searching for regions of rare conformation is finding the closest match between one protein fragment and a large library of protein fragments. Several methods have been described (reviewed in Ref. 13). We chose the approach of matching distance matrices on a window of 9 alpha-carbon atoms, which can be calculated quickly and implemented in primitive operations of a relational database, unlike rigid body superposition. The algorithm used pairwise distances between alpha-carbon atoms. These distances are not directly restrained during normal crystallographic refinement, although indirect effects are possible. Therefore, the distance matrices will be largely independent of the normal geometric restraints in crystal structures. The selection criterion then becomes a relational comparison where the discrepancy in each interatomic distance is less than some tolerance.[11] The query on distance matrices can be expressed as a single relational query in SQL. In addition, distance matrices describe protein conformation more robustly than local measures like backbone torsion angles. The mathematical difficulties in using torsions to represent complicated geometric objects are described in Morgan[14] and Luciano et al.[15]

Distance matrixes are independent of the origin and orientation of the underlying structures. The distance matrix approach represents an approximation to the local curvature and torsion of a three-dimensional curve that follows the protein backbone. Because the (i, i + 1) alpha-carbon-to-alpha-carbon distance is nearly constant (~3.8 Å), the (i, i+n) distances correspond to a sequence of secant approximations to the three-dimensional curve. Formally, these distances define secant approximations to the three components of the Frenet frame of the curve.[16] Two curves are identical if they have identical Frenet frames. Matching Frenet frames is independent of the choice of origin and orientation of the curve in space.

Selection of the Test Set of Protein Structures

Protein crystal structures were chosen in seven major classes from SCOP[12]: (1) all alpha proteins, (2) all beta proteins, (3) alpha/beta (mainly parallel), (4) alpha and beta (mainly anti-parallel), (5) multidomain proteins, (6) membrane and cell surface proteins, and (7) small proteins (defined as mostly stabilized by SS-bonds, bound metals, and other ligands). A protein structure was selected from nearly every fold in each class. The resolution range of the 300 selected structures was 0.9 to 3.2 Å and the R-factors ranged from 0.10 to 0.26. The proteins were not selected on the basis of whether they had ligands or other functional features.

Distance Matrix Encoding of Protein Structures

Each protein chain was split into $N - 9$ overlapping fragments of 9-residues, where N is the number of residues in the chain. Each 9-residue fragment was encoded into the set of pairwise alpha-carbon distances. First, alpha-carbon distances from residue O to O+2, O + 3,..., O+8 were calculated. The distance between successive alpha-carbon positions (i and i + 1) was excluded because it is nearly constant at ~3.8 Å. The next set of distances was calculated by advancing one residue and calculating the alpha-carbon distances from O + 1 to O + 3, O + 4,..., O + 8. This distance calculation was repeated to the end of the window to form a set of 28 unique distances for each 9-residue fragment. Then, the window was advanced by one residue, and the calculation of 28 unique distances was repeated for each overlapping 9-residue window in the protein.

Searching the Database for Rare Regions

An updateable, nonredundant protein set[17] comprising 5117 (as of March 2002) crystal structures was used to prepare the relational database of all the protein fragments. A total of 689,875 encoded fragments were placed into an SQL-database to allow for efficient searches. The 300 protein structures were analyzed by calculating the 28 unique alpha-carbon alpha-carbon distances for each possible overlapping 9-residue fragment, as described above. The sets of pairwise alpha-carbon distances, which represent local backbone conformations of the protein, were compared to pairwise alpha-carbon distances in the nonredundant fragment database using the SQL database program Oracle.[18] Fragments were flagged as rare if any of the unique distances differed by >1.4 Å from the equivalent distance in any other fragment in the nonredundant database. The tolerance value of 1.4 Å was chosen as 3.5 times the average discrepancy of 0.4 Å between the positions of equivalent alpha-carbon atoms in independently refined crystal structures of identical proteins.[19,20] A smaller tolerance value would result in many more rare regions. A window size of 9 residues with this tolerance was found to give about 1 rare window in every 100 windows. However, using a window of <9 residues produced fewer rare regions.

Further Analysis of Structures Containing Rare Conformations

A subset of 78 crystal structures were examined in order to reveal factors associated with the rare fragments. Consecutive 9-residue rare fragments were combined into a single rare region. The overall B-factor was calculated as the average over all the individual atomic B-factors. The local B-factor was calculated as the average of the B-

TABLE I. Protein Distribution by Fold Class

	Total	All alpha	All beta	Alpha/beta	Alpha + beta	Multidomain	Membrane	Small
A. Number of Proteins Sampled in each SCOP Class[a]								
Test Set	300	68 (23%)	50 (17%)	50 (17%)	88 (29%)	14 (4%)	6 (2%)	24 (8%)
% SCOP	100%	23%	16%	16%	30%	5%	2%	8%
B. Protein Structures with Rare Conformations[b]								
Proteins	184	31 (17%)	29 (16%)	40 (22%)	58 (32%)	13 (7%)	5 (3%)	8 (4%)
% Test	61%	31/68	29/50	40/50	58/88	13/14	5/6	8/24
Set		46%	58%	80%	66%	93%	83%	33%
C. Structures with Rare Regions Selected for Detailed Analysis[c]								
Proteins	78	17 (22%)	13 (17%)	13 (17%)	23 (29%)	4 (5%)	2 (3%)	6 (8%)
# Rare regions	149	30 (20%)	20 (13%)	33 (22%)	40 (27%)	12 (8%)	2 (1%)	12 (8%)
# Rare Protein	1.91	1.76	1.54	2.54	1.74	3.0	1.0	2.0

[a]The number of proteins sampled in each of the 7 SCOP major classes ([12]) is shown for the test set of 300 proteins. The "small" protein class is defined as proteins mostly stabilized by disulfides or ligands. The test set includes "small" proteins up to 326 residues in length.
SCOP indicates the percentage of each class in the whole SCOP database.
[b]The distribution is shown for the 184 protein structures with at least one rare fragment.
[c]The distribution of the 78 structures that were selected for detailed analysis. The number of rare regions and the number of rare regions per protein structure are shown in each SCOP class.

factors for main chain atoms of the rare fragment. Protein structures were displayed with RasMol[21] in order to define residues forming alpha helices and beta strands using the algorithm in Kabsch and Sander[22] and to identify any ligand bound near the rare conformation. Waters were not considered. The distance between all the protein atoms in the rare region and the ligand atoms was measured (excluding hydrogen atoms). The ligand was considered to be in "shell 1" if at least one interatomic distance was $\leq$4.0 Å, and in "shell 2" if the shortest interatomic distance was between 4.0 and 6.0 Å.

Subunit-subunit interfaces and domain-domain contacts were also analyzed. Proteins were examined with Rasmol[21] to identify domain-domain contacts with an interatomic distance of $<$5.0 Å. Protein domains were verified with the 3Dee database.[23] Subunit-subunit interfaces were also identified using the distance criteria of $<$5.0 Å. Rare regions present in identical or nearly identical places in more than one subunit were counted as one rare fragment.

Statistical analysis was performed on 124 proteins with both nonprotein ligands and rare regions. In order to normalize for protein size, the fraction of all rare 9-residue windows in shell 1 was plotted against the fraction of all 9-residue windows in shell 1 of the ligand. Analysis of variance (ANOVA) was used to assess the statistical significance of the results.

RESULTS AND DISCUSSION

The 300 crystal structures in the test set were selected from the seven major SCOP classes of protein fold, as shown in Table I. The number of structures selected in each SCOP class was proportional to their relative distribution in the complete SCOP database.[12] The largest number of structures was in the alpha + beta class, and the smallest in the membrane protein class. These structures were analyzed for the presence of 9-residue fragments with a backbone conformation that was not observed in any other protein in the entire database of nonredundant protein structures. The distance matrix comparison showed that 184 (61%) protein structures had at least one fragment with rare conformation. The multidomain class contained the highest percentage (93%) of proteins with rare conformations. More than 80% of tested proteins in the alpha/beta, multidomain, and membrane protein classes showed at least one rare region. The fewest (33%) of proteins with rare regions were found in the small protein category.

A subset of 78 protein structures containing rare regions were examined to reveal the factors associated with rare regions. The proteins for initial analysis were selected in proportion to the class distribution in the complete SCOP database (Table IC). The 78 protein structures showed 149 regions of rare conformation with an average of 1.9 rare fragments per protein. Each protein structure had from 1 to 6 rare fragments. The most rare regions per protein were observed in the alpha/beta and multidomain protein classes, and the fewest in the membrane protein class.

Analysis of the 78 proteins and 149 regions of rare conformation is given in Table I of Supplementary Materials. The regions of rare conformation ranged in length from the minimum of 9 to a maximum of 24 residues (122–145 in PDB entry 1AOL). First, the 149 rare regions were examined for the presence of regular secondary structure (Table II). Few rare fragments included regular secondary structure: 12% contained some residues in alpha-helical conformation next to residues of irregular conformation, 8% included beta-strands, and $<$1% included both types of regular secondary structure. The majority (74%) of rare regions lacked regular secondary structure and occurred on the protein surface. These surface fragments were designated "loops." Different types of loop conformation were not considered, because the rare or unique regions are unlikely to fall into a specific structural classification. The 7 remaining rare fragments (5%) were located at the N- or C-termini. The result is not surprising because surface loops are more likely to have irregular and uncommon conformations compared with

TABLE II. Secondary Structure of Rare Regions[†]

All loop	110 (74%)
All beta	3 (2%)
Beta-loop-beta	2 (1%)
Beta-loop or loop-beta	7 (5%)
All helical	0 (0%)
Helix-loop-beta	1 (0.7%)
Helix-loop or loop-helix	17 (11%)
Loop-helix-loop	1 (0.7%)
N-terminus-helix	1 (0.7%)
N- or C-terminus	7 (5%)

[†]Loop indicates residues in a surface loop or turn. Beta indicates 3 or more consecutive residues in β-strand conformation. However, only 2 consecutive residues in beta conformation were required if a rare region consisted of beta-loop-beta. Helix indicates 4 or more consecutive residues in α-helical conformation.

TABLE III. Classification of the 78 Proteins with 149 Fragments of Rare Conformation[†]

Category	Proteins (% of 78)	Rare fragments (% of 149)
1. Ligand-Binding Site	46 (59%)	67 (45%)
2. Protein-Protein Interface	6 (8%)	10 (7%)
3. Domain-Domain Contact	4 (5%)	5 (3%)
4. High overall B-factor	3 (4%)	12 (8%)
5. Local B >2.0x Boverall	1 (1%)	4 (3%)
6. Unknown	18 (23%)	51 (34%)
Potentionally Functional (Total of categories 1–3)	56 (72%)	82 (55)%

[†]Because in some cases a protein or a fragment can be in more than one of the categories, to simplify the presentation this classification is hierarchical, e.g., if a protein fits into categories 1 and 2, then it is listed only in category 1. (4 rare fragments were near a ligand and subunit interface, and 3 were near ligand and at an interdomain contact. These fragments were classified as "ligand-binding.")

regions with regular secondary structure. However, surface loops are also very likely to be involved in protein function, or alternately, show local disorder.

Fragments of rare backbone conformation will occur by chance; however, two other possibilities have been examined. The rare fragment may be important for molecular recognition or else may arise due to disorder and experimental errors.

Potential Functional Significance of Peptides with Rare Conformation

Regions of rare backbone conformation in protein structures may be important for molecular recognition and, consequently, for protein function. Three indications of functional importance were examined: the presence of a ligand in the vicinity of the rare fragment, and location of rare regions at subunit-subunit interfaces and domain-domain contacts (see Table I in Supplementary Materials).

Protein structures containing fragments of rare conformation were examined for the presence of ligand near the rare fragment. A ligand was found within 6 Å of a rare conformation in 59% of the proteins and 45% of the rare fragments (Table III). The ligands ranged in size from metal ions such as zinc or magnesium to larger molecules including chlorophyll and DNA. There were 56 ligands in the inner shell with at least one atom within 4 Å of the rare fragment and 13 ligands in the 4–6 Å shell. A ligand in shell 1 is likely to form direct interactions with residues of the rare fragment. An average distance of 3.8 Å between protein and ligand atoms was found in a recent study.[24] An example of a protein with a ligand bound within 4.0 Å of atoms of a rare fragment was found in the all beta protein Cytochrome domain of Cellobiose Dehydrogenase (1D7B).[25] The residues 59–70 of both subunits form a loop that is part of the heme-binding site [Figure 1(a)]. Residues of this loop form specific interactions with the heme: the amide of Met 65 formed a hydrogen bond interaction with a carboxylate oxygen of Heme 401 with distance of 2.6 Å.

The presence of ligand in shell 2 indicates longer range electrostatic interactions or solvent-mediated interactions between ligand and the rare fragment. One example is shown for the protein Udp-*N*-acetylmuramoyl-L-alanine: D-glutamate ligase (2UAG)[26] in the alpha/beta class. This protein has a rare loop consisting of residues 413–421 that fold over the UMA ligand bound in a cleft of the protein [Figure 1(b)]. The shortest interatomic distance from Leu 416 to UMA 450 is 4.6 Å. The presence of a ligand near the rare fragment suggests that the rare conformation may be important for the recognition or binding of the respective ligand.

The location of rare backbone conformations was analyzed with respect to the intersubunit interface of multisubunit structures. There were 29 proteins with two or more subunits; 24 were oligomers of 2 to 14 identical subunits, 2 consisted of different types of subunits, and 3 were mixed with both identical and different subunits. Most had two subunits; however, one protein structure included 14 subunits. The analysis of protein structures with multiple subunits is complex, because oligomers do not always consist of subunits in identical conformations and rare regions could be important subunit-subunit contacts. Eleven multisubunit proteins had 38 rare fragments (or 14 unique regions per subunit) located at a subunit-subunit interface. The closest interatomic distance across the interface ranged from 2.5 to 4.2 Å, consistent with direct van der Waals or polar interactions. These distances are in agreement with analysis showing that both close packed protein atoms and interacting water molecules contribute to protein-protein interfaces.[27]

The location of the rare fragments was compared for each subunit of oligomeric proteins. Fourteen of the multisubunit structures containing identical subunits had all rare regions in the same place or overlapping regions in two or more identical subunits, 6 had rare regions in different locations, and 7 showed a mixture with rare fragments located in both identical and different regions of the subunits. The majority of rare fragments, consisting of 37 unique sites per subunit, were found in identical or largely overlapping regions, whereas 17 fragments were in different regions of identical subunits. The largest oligomer is the 14-mer of Clp Protease (1TYF),[28] which had a rare region at the interface in essentially the same place in all subunits [Fig. 2(a)]. The central pore region of Clp

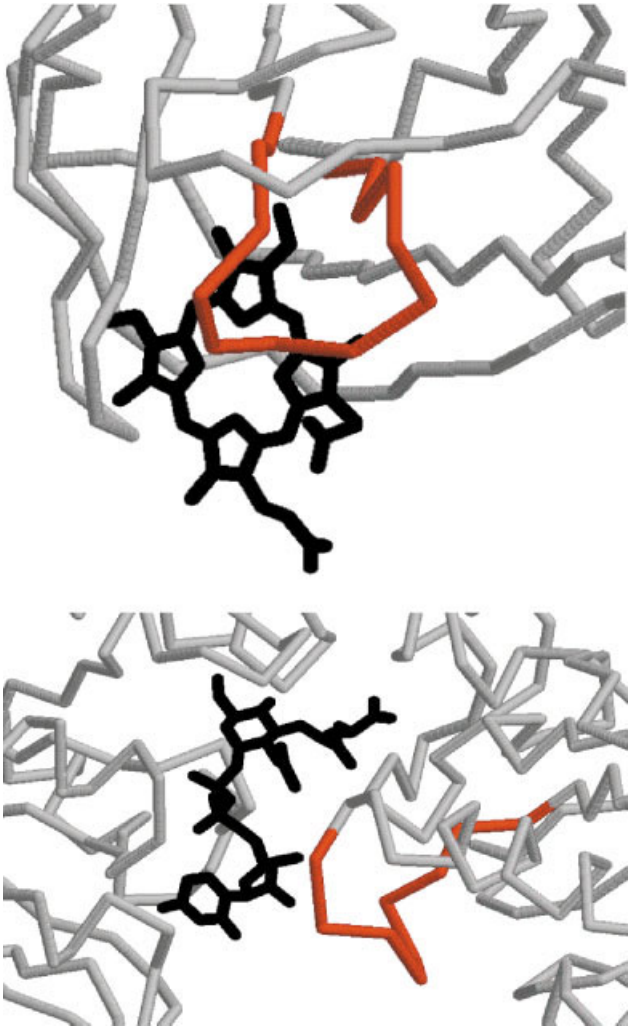

Fig. 1. (a) The rare conformation fragment in the all beta protein 1D7B (cytochrome domain of cellobiose dehydrogenase). The rare fragment consists of residues 58–70 of subunit A (red) that interact with hemoglobin (black). The protein backbone is shown in grey. (b) The rare conformation fragment of the alpha/beta fold class protein Udp-*N*-Acetylmuramoyl-L-Alanine: D-glutamate ligase (2UAG). Residues 413–421 (red) with a rare backbone conformation fold over the cleft in which in the ligand uridine-5′-diphosphate-*N*-acetylmuramoyl-L-alanine (black) binds. The shortest interatomic distance from Leu 416 to UMA 450 is 4.6 Å. The surrounding protein backbone is shown in grey.

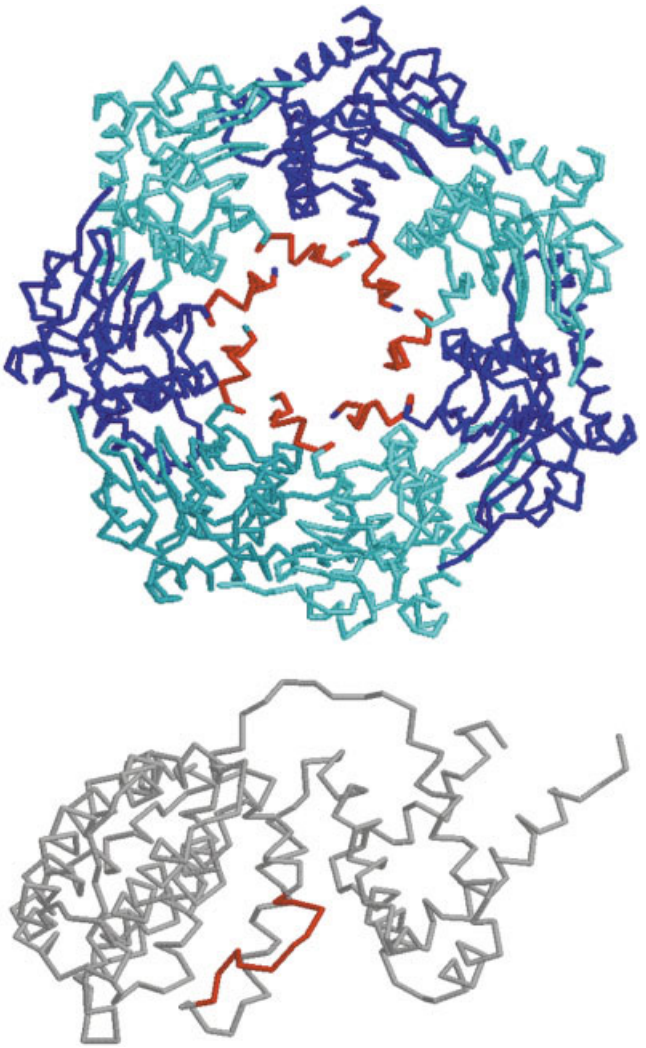

Fig. 2. (a) The rare conformation loop (red) consists of residues 12–21 in each of the 14 subunits of the alpha/beta protein Clp Protease (1TYF). The rare region is located at the interface between subunits. The 7 subunits A–G of one layer are shown in shades of blue. (b) The rare region (residues 786–795 in red) of the all alpha protein Ras-Gtpase-activating domain of human P120Gap (1WER) is located at the interface between two domains.

Protease is also the active site. Rare conformations may be observed in identical locations in several subunits either because the subunits share identical conformations, or artificially due to use of the same subunit model during determination of the crystal structure. Alternatively, rare fragments can be found at different regions of identical subunits due to experimental errors, or real structural differences in the subunits. However, 38% of multisubunit proteins had rare fragments located at subunit interfaces, which suggests a role in protein recognition, rather than a random occurrence.

The location of the rare fragments was analyzed with respect to the interface between structural domains in proteins. A total of 7 proteins were observed to have 8 rare regions forming domain-domain contacts. The rare region (residues 786–795) of the all alpha fold protein, Ras-Gtpase-activating domain of human P120Gap (1WER)[29] is shown in Figure 2(b). In P120Gap, the closest distance between atoms of the two domains was 3.8 Å. Overall, the shortest interatomic distances ranged from 2.6 to 4.6 Å between the domains of all proteins. In fact, the interdomain contacts were all <4.2 Å, except for one fragment in 2UAG, which showed a separation of 4.6 Å between domains. These distances are consistent with close-packed van der Waals and polar interactions between the domains, similar to the results for intersubunit interactions. The domain interface is critical for domain movements, which are important for the function of many multidomain proteins.[30] In addition, the interdomain regions are likely to be involved in stabilizing the tertiary structure and to form ligand-

TABLE IV. Resolution, *R*-factor, and *B*-factor for the Analyzed Crystal Structures

	Resolution (Å)		*R*-factor		*B*-factor (Å^2)	
	Range	Average	Range	Average	Range	Average
Test set (300)	0.9–3.2	2.1	0.10–0.26	0.21	7.2–89.1	28.3
Proteins with rare regions (184)	1.3–3.2	2.1	0.11–0.26	0.22	7.3–89.1	29.3
Proteins without rare regions (116)	0.9–3.2	2.0	0.10–0.25	0.20	7.2–70.5	26.6

binding sites. Consequently, the domain interface residues are important for the function.

The categories that are potentially important for protein function are summarized in Table III. Only functional sites that could be identified from the analyzed crystal structures were included. Therefore, the analysis represents the information likely to be present in the structure of a protein of unknown or poorly defined function. It is likely to be an underestimate of the actual number of rare regions associated with functional categories of ligand-binding site, subunit interface, or domain-domain contact. Proteins that contained fragments of rare conformation included 59% with a ligand bound near a rare region; an additional 8% had rare regions at subunit interfaces, and 5% had rare regions at domain-domain contacts. A total of 55% of all the rare fragments can be classified as functional. However, these potentially functional regions were identified in 72% of the proteins with rare conformations. Therefore, in the majority of proteins with a region of rare conformation, the unusual backbone conformation is involved in function. This result is consistent with previous observation that sterically strained backbone regions were generally associated with the function.[10]

Quality of Protein Structures containing Rare Conformations

The selected protein structures were examined to determine if the presence of rare conformations was indicative of possible problems with quality of structures. Identification of rare backbone conformations over 5-residue windows is part of the WHAT_CHECK structure validation process; however, the rare conformations can occur by chance as well as by experimental error.[9]

The resolution, crystallographic *R*-factor and overall *B*-factor were compared for structures with and without rare regions (Table IV). The proteins with rare fragments showed an average resolution of 2.1 Å, with the range of 1.3–3.2 Å, whereas the proteins without rare fragments had the range of 0.9–3.2 Å and average of 2.0 Å. The *R*-factors had a similar range (0.11–0.26 and 0.10–0.25) with average of 0.22 for structures with rare regions, and 0.20 for structures without rare regions. Therefore, the range and average for resolution and *R*-factor was essentially the same for structures in both categories. Moreover, the presence of rare fragments did not show significant correlation with values of either *R*-factor or resolution.

Another measure of the quality of a crystal structure is the average atomic *B*-factor, because a high *B*-factor suggests high thermal disorder. The 185 protein structures containing rare fragments had average *B*-factor ranging from 7.3 to 89.1 Å^2, which is higher than observed for the structures without rare regions (7.2–70.5 Å^2; Table IV). In fact, 14 (18%) proteins containing rare regions had unusually high overall *B*-factors of >40 Å^2, which is consistent with over-refinement of *B*-factors and may indicate poor quality of phases. However, 15% of proteins without rare conformations also showed an overall *B*-factor of >40 Å^2. The same proportion (15%) of crystal structures in the complete PDB was considered to have over-refined *B*-factors.[9] Therefore, the structures with rare regions had only a small excess (18 compared with 15%) of high overall *B*-factors.

The local disorder was estimated by calculating the average *B*-factor for the backbone atoms of the rare fragment (see Table I in Supplementary Materials). These local backbone *B*-factor values were examined as a fraction of the overall *B*-factor for each structure. Only a few (7/149 or 5%) rare fragments showed a local backbone *B*-factor of more than twice the overall *B*-factor, as an indication of high local disorder. Therefore, the fragments of rare conformation are generally well defined in protein structures, despite their predominance in surface loops.

Finally, 23% (18/78) of the proteins had none of the previously examined features: the overall *B*-factor was <40 Å^2, the local main chain B-factor was less than twice the overall B-factor, and the rare regions were not located near a ligand and did not form a domain-domain contact or subunit interface. These proteins contained 35% of the rare fragments. However, this category is overestimated, because not all tested protein structures were determined in the complex with ligand or in the appropriate oligomeric state. In some proteins (e.g., 1CHD), two rare regions were close to each other, which suggested a possible role in maintaining tertiary structure. In addition, these rare fragments may occur by chance, experimental error, or arise from some unknown or unexamined role in function.

Statistical Analysis of Rare Regions in Ligand Binding Sites

The analysis of the subset of 78 proteins showed that the largest number of rare regions were associated with ligands. Therefore, the 124 proteins with both nonprotein ligands and rare regions from the complete test set of 300 were examined for the presence of rare fragments near ligands (Table II of Supplementary Materials). The proteins ranged in size from 51 to 1612 residues, and the ligands ranged from metal ions to nucleic acids. Protein fragments were considered to be in shell 1 of the ligand if at least one protein atom was within 4.0 Å of a ligand atom. The null hypothesis is that rare backbone conformations

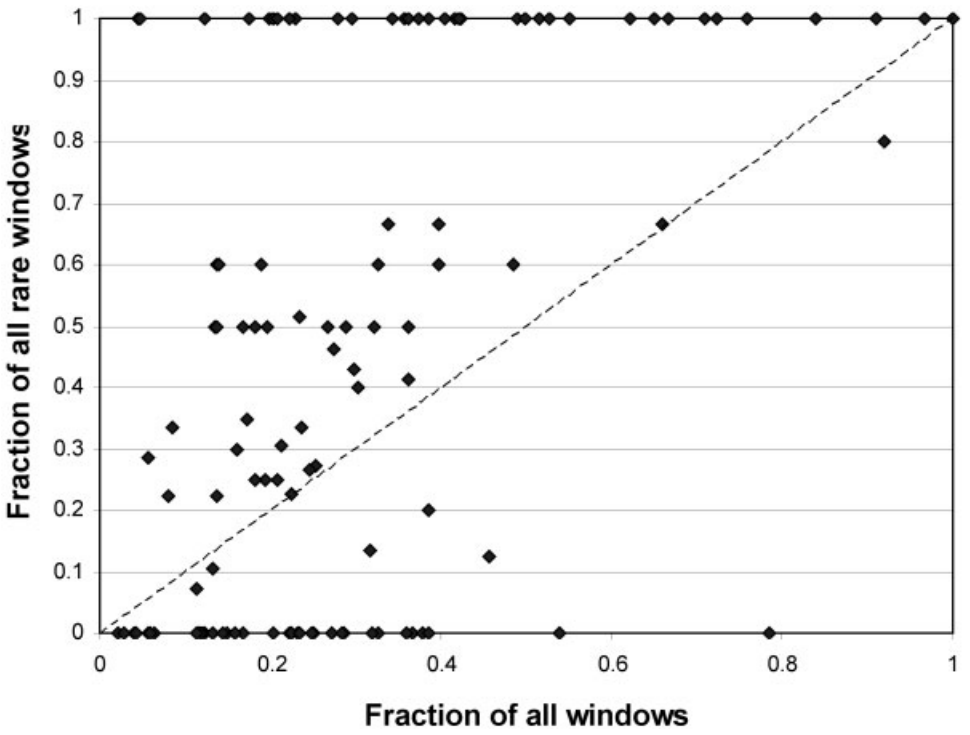

Fig. 3. Analysis of windows near ligand for 124 proteins with rare backbone regions and ligands (Table II of Supplementary materials). The fraction of all rare 9-residue windows near ligand is plotted against the fraction of all 9-residue windows near ligand (at least one interatomic distance of ≤4.0 Å). The "null hypothesis" diagonal is shown as a dashed line.

occur by chance alone. In that case, the distribution between shell 1 and the rest of the protein will be the same for rare 9-residue regions and all 9-residue regions. On average, there were 1.7 rare 9-residue fragments in every 100 windows or in a 109-residue protein. In order to normalize for different protein sizes, the fraction of rare 9-residue windows in shell 1 was plotted against the fraction of all 9-residue windows in shell 1 of the ligand (Fig. 3). If rare regions were distributed evenly as would be expected by chance alone, the points would cluster around the diagonal from (0,0) to (1,1). However, the points do not follow the diagonal, suggesting that more than one factor is involved. ANOVA was performed between the groups consisting of the fraction of rare conformation windows near ligand and the fraction of all windows near ligand. The results ($F = 5.2$, $p = 0.023$) strongly suggest that at least two distinct groups are present. 34.7% proteins had no rare regions near the observed ligands. Interestingly, 30.8% of the proteins had all the rare regions next to ligands. Another 35.5% had at least one rare region next to the ligand. Not all of the rare regions were near a nonprotein ligand in the complex. However, not all of the ligand-binding sites may be established for the 124 proteins, so these numbers are likely to underestimate the association of rare fragments with functional sites.

CONCLUSION

Efficient distance matrix searches of a large database of protein fragments have been used to identify protein fragments with rare backbone conformations. Rare conformations were found in 61% of the 300 tested protein structures in all major folds. As expected, the majority of rare regions were in surface loops without regular secondary structure. Detailed analysis of ligand-binding sites, protein-protein interactions, and domain-domain contacts has shown that 55% of all the individual rare fragments can be classified as functionally significant. These potentially functional regions were identified in 72% of the proteins with rare conformations. In another 5% of the structures, the rare conformation regions are likely to appear because of high average *B*-factors or local backbone disorder. In the remaining 23% of the proteins with regions of rare conformation, none of these features was found in the structure. This "unknown" category is likely to be overestimated, because the rare regions in these proteins were not correlated with all known functional sites, only those observed in the particular crystal structure. Statistical analysis has shown that the rare regions are not distributed evenly in the protein structure. Therefore, the existence of a rare backbone region in a protein suggests that a functional site might be nearby, although the possibility that the rare region is due to chance cannot be excluded. Unexpectedly, the rare regions are not common in areas of local high *B*-factor. This suggests that unique backbone conformations may be important for protein recognition of ligands. Therefore, rare conformations should not be automatically eliminated during crystallographic refinement. The rapid distance matrix search

can assist in prediction of potential sites of biological function in a novel protein.

ACKNOWLEDGMENTS

We are grateful for discussions with Dr. Bi Cheng Wang and other members of the South Eastern Collaboratory for Structural Genomics.

REFERENCES

1. 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. http://www.rcsb.org/.
2. Eisenstein E, Gilliland GL, Herzberg O, Moult J, Orban J, Poljak RJ, Banerjei L, Richardson D, Howard AJ. Biological function made crystal clear—annotation of hypothetical proteins via structural genomics. Curr Opin Biotechnol 2000;11:25–30.
3. Thornton JM, Todd AE, Milburn D, Borkakoti N, Orengo CA. From structure to function: approaches and limitations. Nat Struct Biol 2000;7 Suppl:991–994.
4. Landgraf R, Xenarios I, Eisenberg D. Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. J Mol Biol 2001;307:1487–1502.
5. Lim K, Zhang H, Tempczyk A, Bonander N, Toedt J, Howard A, Eisenstein E, Herzberg O. Crystal structure of YecO from Haemophilus influenzae (HI0319) reveals a methyltransferase fold and a bound *S*-adenosylhomocysteine. Proteins 200;45:397–407.
6. Martinez-Cruz LA, Dreyer MK, Boisvert DC, Yokota H, Martinez-Chantar ML, Kim R, Kim SH. Crystal structure of MJ1247 protein from M. jannaschii at 2.0 Å resolution infers a molecular function of 3-hexulose-6-phosphate isomerase. Structure (Camb) 2002;10: 195–204.
7. Minasov G, Teplova M, Stewart GC, Koonin EV, Anderson WF, Egli M. Functional implications from crystal structures of the conserved *Bacillus subtilis* protein Maf with and without dUTP. Proc Natl Acad Sci USA 2000;97:6328–6333.
8. Rosen M, Lin SL, Wolfson H, Nussinov R. Molecular shape comparisons in searches for active sites and functional similarity. Protein Eng 1998;11:263–277.
9. Hooft RWW. Vriend G, Sander C, Abola EE. Errors in protein structure. Nature 1996;381:272–272.
10. Herzberg O, Moult J. Analysis of the steric strain in the polypeptide backbone of protein molecules. Proteins 1991;11:223–229.
11. Lesk AM. Extraction of geometrically similar substructures: least-squares and Chebyshev fitting and the difference distance matrix. Proteins 1998;34:317–32.
12. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540. http://scop.mrc-lmb.cam.ac.uk/scop/.
13. Holm L, Sander C. Mapping the protein universe. Science 1996;273: 595–602.
14. Morgan A. Solving polynomial systems using continuation for engineering and scientific problems. New York: Prentice-Hall; 1987. p 228–263.
15. Luciano C, Banerjee P, Mehrotra S. 3-D animation of telecollaborative anthropomorphic avatars. Commun ACM. 2001;44:65–67.
16. Farin G. Curves and surfaces for computer-aided design. New York: Academic Press, 1998. p141–150.
17. Madej T, Gibrat JF, Bryant SH. Threading a database of protein cores. Proteins 1995;23:356–369.
18. Oracle Corporation, Oracle http://www.oracle.com/corporate/.
19. Flores TP, Orengo CA, Moss DS, Thornton JM. Comparison of conformational characteristics in structurally similar protein pairs. Prot Sci 1993;2:1811–1826.
20. Betts M.J, Sternberg MJE. An analysis of conformational changes on protein-protein association: implications for predictive docking. Prot Eng 1999;12:271–283.
21. Sayle RA, Milner-White JE. RASMOL: biomolecular graphics for all. Trends Biochem Sci 1995;20:374–376.
22. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.
23. Siddiqui AS, Dengler U, Barton GJ. 3Dee: a database of protein structural domains. Bioinformatics 2001;17:200–201.
24. Moreno E, Leon K. Geometric and chemical patterns of interaction in protein-ligand complexes and their application in docking. Proteins 2002;47:1–13.
25. Hallberg BM, Bergfors T, Backbro K, Pettersson G, Henriksson G, Divne C. A new scaffold for binding haem in the cytochrome domain of the extracellular flavocytochrome cellobiose dehydrogenase. Structure Fold Des 2000;8:79–88.
26. Bertrand JA, Auger G, Martin L, Fanchon E, Blanot D, Le Beller D, Van Heijenoort J, Dideberg O. Determination of the MurD mechanism through crystallographic analysis of enzyme complexes. J Mol Biol 1999;289:579–590.
27. Lo Conte L, Chothia C, Janin J. The atomic structure of protein-protein recognition sites. J Mol Biol 1999;285:2177–2198.
28. Wang J, Hartling JA, Flanagan JM. The structure of ClpP at 2.3 A resolution suggests a model for ATP-dependent proteolysis. Cell 1997;91:447–456.
29. Scheffzek K, Lautwein A, Kabsch W, Ahmadian MR, Wittinghofer A. Crystal structure of the GTPase-activating domain of human p120GAP and implications for the interaction with Ras. Nature 1996;384:591–596.
30. Gerstein M, Lesk AM, Chothia C. Structural mechanisms for domain movements in proteins. Biochemistry 1994;33:6739–6749.