PROTEINS: Structure, Function, and Genetics 34:317–332 (1999)

A Rapid Method for Exploring the Protein Structure Universe

Malin M. Young, A. Geoffrey Skillman, and Irwin D. Kuntz*
Department of Pharmaceutical Chemistry, University of California, San Francisco, San Francisco, California

ABSTRACT **We have developed an automatic protein fingerprinting method for the evaluation of protein structural similarities based on secondary structure element compositions, spatial arrangements, lengths, and topologies. This method can rapidly identify proteins sharing structural homologies as we demonstrate with five test cases: the globins, the mammalian trypsinlike serine proteases, the immunoglobulins, the cupredoxins, and the actinlike ATPase domain-containing proteins. Principal components analysis of the similarity distance matrix calculated from an all-by-all comparison of 1,031 unique chains in the Protein Data Bank has produced a distribution of structures within a high-dimensional structural space. Fifty percent of the variance observed for this distribution is bounded by six axes, two of which encode structural variability within two large families, the immunoglobulins and the trypsinlike serine proteases. Many aspects of the spatial distribution remain stable upon reduction of the database to 140 proteins with minimal family overlap. The axes correlated with specific structural families are no longer observed. A clear hierarchy of organization is seen in the arrangement of protein structures in the universe. At the highest level, protein structures populate regions corresponding to the all-alpha, all-beta, and α/β superfamilies. Large protein families are arranged along family-specific axes, forming local densely populated regions within the space. The lowest level of organization is intrafamilial; homologous structures are ordered by variations in peripheral secondary structure elements or by conformational shifts in the tertiary structure. Proteins 1999; 34:317–332.** © **1999 Wiley-Liss, Inc.**

Key words: fold space; protein family; structural comparison; protein fingerprinting, principal component analysis

INTRODUCTION

Protein structures are windows that give us a glimpse into the distant evolutionary past. However, the relationships among protein families are complex, and the customary measures, which depend on primary sequence similarity, have well-known limitations. Seminal work on pairs of homologous proteins showed that amino acid sequence identities may drop dramatically without a significant increase in structural root-mean-square deviations .[1] The tolerance of structure to sequence variability is the major reason why primary sequence alignments are not always sensitive enough to detect the distant evolutionary relations between proteins that can be observed by structural comparison techniques. Sequence-independent structural comparisons not only can give us insights into protein evolutionary relations, but they enable us to define conserved folding elements and structural cores. Identification of key structural elements found in natural proteins may lead to deeper insights into how to design protein structures de novo.

The benefits of protein structural comparisons are apparent; however, the growth of the structure databases presents a significant computational challenge for structural similarity searching and clustering techniques. Although lagging far behind the number of known sequences, the number of protein tertiary structures in the Protein Data Bank (PDB) has grown exponentially in the past 5 years, from 657 entries released in 1992 to the present total of over 7,500 released entries, with a current doubling time of approximately 18 months.[2,3] The need to stay abreast of the ever-increasing structural data flow has recently driven the development of rapid and automatic structural comparison methods. In this paper, we present a fast method for defining structural similarities and its application to the automatic organization of a large set of protein structures based upon their structural similarities.

The first attempts to automate protein structural alignments applied rigid-body rotations and translations to superimpose the atomic positions of one structure onto another.[4] As expected, this type of approach was reasonably successful when applied to the alignment of similar structures. However, the obvious drawback to these methods is their sensitivity to insertions and deletions, which can obscure matches between structures sharing common substructural elements.

Later work has focused on overcoming this sensitivity to insertions and deletions. The challenge of protein structural comparison became the identification of the subset of structurally correlated atoms or fragments, termed the

Grant sponsors: National Science Foundation, General Medical Sciences Institute of National Institutes of Health; Grant numbers: GM07075, GM08388, GM31497.
*Correspondence to: Dr. Irwin D. Kuntz, Department of Pharmaceutical Chemistry, Box 0446, University of California, San Francisco, San Francisco, CA 94143–0446. E-mail: kuntz@cgl.ucsf.edu
Received 3 April 1998; Accepted 19 October 1998

equivalence set, shared by two structures. Once an equivalence set is assigned, optimization techniques can be used to generate a final structural alignment. Dynamic programming,[5,6] distance matrices,[7,8] fragment matching,[9,10] hashing,[11] maximal common subgraph detection,[12,13] and simulated annealing[14] methods were all applied singly and in combination to the problem of defining the initial equivalence set. Optimization of the equivalence sets generally has been performed using cycles of multiple dynamic programming and structural alignment, Monte Carlo algorithms, or simulated annealing. Optimizations of this sort are computationally intensive, which has previously restricted the utility of these methods.

Recently, protein structural comparison algorithms and computational resources have progressed sufficiently so that searching and clustering the PDB has become computationally feasible.[15–19] Secondary structure-based prescreening techniques have made the process of database searching and clustering more rapid.[9, 20–24] Gibrat et al.[25] noted that all of these algorithms have approached protein structure comparisons in the same way, by treating each structure as an assembly of secondary structure element (SSE) pairs. To compare structures, SSE pairs are exhaustively examined, and similar SSE pairs are stored in a pairs list. The pairs list is used to calculate a maximum common substructure of SSE pairs.

Two algorithms, DALI/3D lookup and VAST, use this type of SSE-based representation to perform rapid protein structural comparisons. They have been employed to automatically classify sets of protein structures and domains. The results of their analyses are found in two databases: FSSP (at the European Molecular Biology Laboratory) and MMDB (at the National Center for Biotechnology Information).[26,27] These databases differ in their approaches to organizing the similarities shared by protein structures. The FSSP database summarizes the results from two all-by-all structural similarity calculations: the similarities between sequence-unique representative structures (with sequence identities $\leq$25%), and the similarities between the members of each sequence family and their representative structure. Criticisms of this approach are directed at the method used to select the representatives from sequence-based groups.[18] MMDB provides a look at the protein structural neighborhood around each tertiary structure at five different sequence redundancy levels, based on a precalculated all-by-all structural comparison of domains and tertiary structures.

Two other classification schemes, SCOP and CATH, classify protein structures based on a combination of automatic methods and manual inspection.[28,29] The organization of structures in both databases is hierarchical. The SCOP classification places proteins into tertiary structural classes, superfamilies, and folds, based on similarities in function, overall structure, and domain folds. The CATH classification is at the domain level, with an emphasis on organizing domains into classes of distinct architectures.

All-by-all comparisons of protein structures can be used to visualize the distribution of protein structures in a theoretical high-dimensional space through statistical methods such as principal components analysis. Recent attempts have been made to map the protein universe[19,30] with limited subsets of the PDB. In each case, an all-by-all comparison was performed on a representative domain database. A two-dimensional[30] or a three-dimensional[19] projection of the fold distribution was calculated using multidimensional scaling or principal components analysis, respectively. Holm and Sander[30] concluded from their two-dimensional projection that fold "attractors" exist in the protein fold space. Sowdhamini's[19]three-dimensional distribution showed no such attractor regions, but rather a continuous triangular space with the extremum points containing all-alpha, all-beta, and α/β structures.

The results of the two studies differ, but whether the differences are due to the projections produced by each study, to the representative domains chosen, or to the differences in the methods used in the structural comparisons is unclear. It remains uncertain whether two- and three-dimensional projections are sufficient to portray accurately the domain space as defined by the representative structures. In addition, as the issues of complexity, organization, and sensitivity of the structural spaces to the contents of the datasets were unaddressed, it becomes difficult to evaluate the usefulness of these representations in biological terms.

This paper introduces a novel and rapid structural comparison algorithm. The algorithm, called protein fingerprinting, evaluates structural similarities based on the arrangement, topology, type, and length of protein secondary structure elements. While the method can identify similarities between protein structures, it does not generate structural alignments, so it is fast enough to complete an all-by-all comparison of the entire PDB in a day on an R4400 SGI workstation. Therefore its strength (i.e., speed) is complementary to slower comparison techniques which generate protein structural alignments. A likely use of this method would be as a database prescreen, prior to more computationally intense protein structural alignment methods.

In addition to being well suited to structural database prescreening, we assert that the rapidity of this method better qualifies it for mapping large regions of the structural space than current structural comparison methods. Because the number of pairwise comparisons increases quadratically with the number of structures, the rapid growth of the PDB puts great pressure on any comparison algorithm. Our method, which takes approximately 0.04 seconds per pairwise comparison, is capable of performing all-by-all calculations very large structural datasets and therefore may provide a useful global view of the structural space.

In this paper, we first calibrate the procedure by carrying out similarity searches for five structural families that span a range of difficulty. We then apply our method towards the generation of a high-dimensional depiction of the protein structural universe by performing an all-by-all comparison of a large set of protein structures and a subsequent principal components analysis. Last, we project the protein universe into three dimensions, corresponding

to the top three principal components, to analyze the organization of structures in greater detail.

MATERIALS AND METHODS

Binary Representation of Protein Structure

At the onset, we have two decisions to make: 1) Shall we treat entire proteins or structural domains?, and 2) What level of resolution shall we use to descibe proteins? Briefly, we have elected to examine entire proteins because such a representation allows us to map similarity relations within and between protein structural families. With both intra- and interdomain level similarity information, our method should be able to evaluate subtle structural changes within protein families and define the relative similarities of protein families sharing common substructures such as domains.

We have chosen to follow the common practice of treating protein structures as ensembles of secondary structures. While our method is not limited to secondary structures,[31] there are several advantages to this representation. First, it is less sensitive to structural perturbations because the positions and types of secondary structures tend to be more conserved among distantly related proteins than either atomic positions or residues. Second, it is faster to find equivalences among secondary structure elements than between atoms or amino acids, enabling an exhaustive search of the equivalence space. Last, a secondary structure-based representation is appropriate because secondary structure arrangements have historically been the basis for empirical protein structure classifications.[32]

Each SSE within a protein structure is assigned a type, a center-of-mass point, and a directional unit vector (Fig. 1). Two types of secondary structure elements, β strands and α helices, are defined using the DSSP algorithm developed by Kabsch and Sander.[33] We chose an automatic procedure to assign secondary structures rather than the PDB structure file assignments for several reasons. One reason was to ensure uniformity. Since our method uses secondary structure assignments to characterize and compare protein structures, we wished to ensure that secondary structure assignments were similar for members of the same structural family. Additionally, we chose to use DSSP simply because not all protein structure files contain secondary structure assignments. By using an automated method, we were free to add any protein structure we wished to our dataset. On the other hand, we are aware that there are peculiarities to the DSSP algorithm that can result in missed, shortened, or fractured secondary structure elements. We encountered this difficulty in many of the test cases we present in this paper. However, during the course of our investigation we have found that our algorithm is capable of recognizing sufficient levels of SSE similarity to identify homologous proteins in spite of DSSP assignment discrepancies.

The point representing each SSE is the center of mass, or positional average, of the SSE C_α atoms. The third descriptor is a unit vector that lies along the axis of rotation of the secondary structure element. These three descriptors are used in the rapid prescreening stage of our algorithm to calculate a binary string containing information about the distance and angular relations between a protein's SSEs, which we call that protein's fingerprint. We chose a binary representation of protein tertiary structure to parallel successful small molecule fingerprinting schemes which have been developed to search small molecule databases containing upwards of hundreds of thousands of structures.[31]

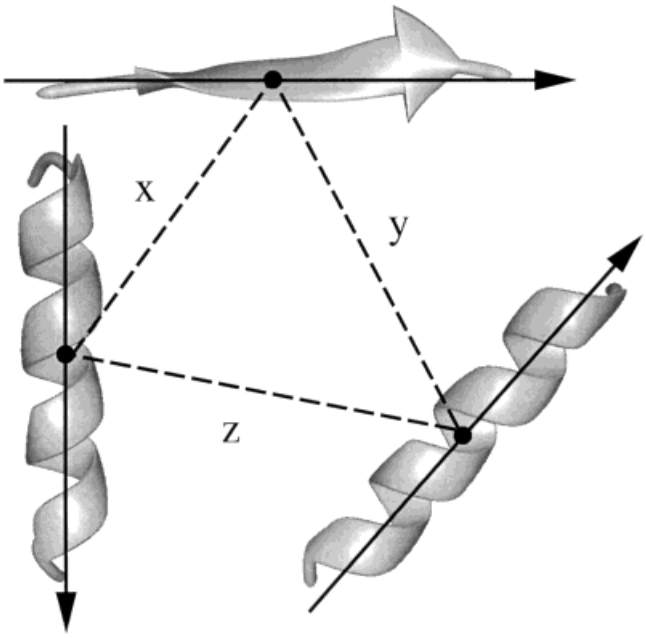

Fig. 1. A triplet of secondary structure elements, consisting of three secondary structure types, centers of mass, and unit vectors. The three descriptors are used to roughly encode the angular and distance relations defined by each triplet into a binary representation of the protein structure. *All* secondary structure triplets in a protein tertiary structure are used in this encoding procedure.

The generation of the fingerprint involves examining all possible triplets of SSEs. This, obviously, limits the precise method presented here to the representation and comparison of protein structures with more than two secondary structure elements. However, as the number of proteins with less than three secondary structure elements is relatively small, we feel that this representation does not substantially limit the scope of our similarity calculations. Further, the fingerprinting approach could be implemented with finer resolution, even to the atomic level, if detailed study of smaller proteins was of particular interest.

An SSE triplet is the smallest similarity unit in the fingerprint. Four triplet types are possible, ααα, ααβ, αββ, and βββ, each of which hashes to a different subregion in the fingerprint. We encode the three distances relating the triplet centers of mass into the appropriate SSE-defined subregion. The three continuous distances (shown as *x*, *y*, and *z* in Figure 1) are ordered in ascending magnitude, placed into discrete distance bins, and their bin numbers are used to calculate a hash index to the subregion within the fingerprint. The bin widths were calibrated to enclose

10% of the observed distance distribution across a diverse protein test set, thereby ensuring uniform population of the bins. The hash index, z, is calculated by Equation (1):

$$z = c + bn + an^2 \quad (1)$$

where c is the bin number for the longest triplet side distance, a is the bin number for the smallest triplet side distance, b is the bin number for the third distance, and n is the total number of bins. Theoretically, then, the length of the distance subregion is equal to n^3. However, not all z values are geometrically possible due to the triangle inequality, therefore z is assigned a new value in order to compress the fingerprint size. The final z value indicates the position within the subregion of the fingerprint that corresponds to the distance information in the SSE triplet. A counter is incremented at this position to reflect that a triplet with the corresponding distance characteristics exists within the protein structure.

The three vectors arising from the three SSEs in a triplet are used to derive an index to a volume subregion of the fingerprint. We encode the volume of the parallelepiped defined by the three vectors by computing the vector triple product. Since the vectors are unit vectors, the volume ranges between -1 and 1, where the sign distinguishes the chirality of the vectors. The volume is used to calculate a bin number. This bin number is the index to the appropriate subregion of the fingerprint, and the counter at that position is incremented. Again, the volume bins are calibrated to span 10% of the observed volume distribution.

We use a binary encoding method to represent the frequency of occurrence of each fundamental feature (i.e., a triplet). A count is made of the times a particular position in the volume and distance regions of the fingerprint is flagged. The frequency of hashing (f_h) to a position is encoded in a binary fashion by flagging the number of bits (m) per position as calculated by the simple equation:

$$m = \text{int}\,([\ln(f_h)/\ln(2)] + 1) \quad (2)$$

which emphasizes the importance of rare triplets with a low f_h in the fingerprint. Thus, this technique encodes information about both the types of triplets and their frequency distribution. After all SSE triplets have been processed, the final result is a binary fingerprint encapsulating the distance and angular relations of a protein's secondary structure elements. A fingerprint is calculated for every protein in the PDB. Our method takes approximately 80 seconds to fingerprint our database of 1,031 protein structures (≈0.078 seconds per structure) on a common desktop workstation, the SGI R4400 workstation. These fingerprints can be used to rapidly prescreen the database as a first step in a searching or clustering calculation.

Fingerprint-Based Database Prescreening

It is known that the similarity between two binary fingerprints can be assessed by calculating a Tanimoto coefficient.[34] The Tanimoto coefficient for a pair of fingerprints A and B is defined as:

$$T = B_c/(B_A + B_B - B_c)$$

where B_c is the number of bits set (bit positions containing 1's) in the same positions in fingerprints A and B, B_A is the number of bits set in fingerprint A, and B_B is the number of bits set in fingerprint B. The Tanimoto coefficient is equal to 1 for identical fingerprints, and 0 for completely disparate fingerprints. Separate Tanimoto scores are calculated for the volume and distance regions of the fingerprint, and the weighted sum of the two scores is used to calculate the final similarity score. The weights for the volume and distance regions were determined empirically, and were calibrated using the globin, mammalian trypsinlike serine proteases, and the immunoglobulin heavy chain test cases. They are 0.3 and 0.7, respectively. The final similarity score, S_f, becomes then Equation (4):

$$S_f = 0.30T_v + 0.70T_d \quad (3)$$

where T_v is the Tanimoto similarity between the volume subregions, and T_d is the Tanimoto similarity between the distance subregions of the pair of fingerprints under comparison. This measure of similarity between fingerprints provides a rapidly calculable estimate for protein structural similarity, which allows rapid prescreening of the protein database.

A similarity score can be calculated for a search key and all the proteins in the database very rapidly because the calculation involves simple binary comparisons. Unlike other methods, the size of the comparison does not increase with the number of SSEs because the length of the fingerprints remains constant. For every protein, then, the average time per pairwise comparison is the same. The prescreening step makes approximately 67 comparisons per second on a Silicon Graphics R4400 workstation.

Triplet-Based Structural Comparisons

The low-resolution representation of protein structure preserved in the fingerprints is sufficient to identify proteins potentially sharing structural similarities. However, because the distance and angular information is decoupled in the fingerprint representation, the resolution is insufficient to rank-order structures by their degree of similarity to a search key. The purpose of the fingerprint comparison step, therefore, is to rapidly discard all protein structures dissimilar to a search key (the protein of interest), and to pass the most promising structures to a more rigorous structure comparison algorithm. This triplet-based comparison stage compares all of the SSE triplets in each of the two proteins under consideration explicitly, and it attempts to match as many as possible using a greedy algorithm. The features used to compare SSE triplets are length of SSEs, their types, the three pairwise angles relating the SSEs in each triplet, the three triplet center-of-mass distances, and implicitly, the topological positions of the triplets. All of the explicitly defined features must

TABLE I. User-Defined Parameters[†]

Parameter	Parameter subtype	Globin, serine protease, IgG light chain, and all-by-all calculation parameters	Azurin and hexokinase B similarity search parameters
Midpoint–midpoint distance tolerances (Å)	HH	5	6
	HS	3	4
	SS	2	3
Length difference tolerances for each triplet type (no. of residues)	HHH	8	10
	HHS	7	8
	HSS	7	8
	SSS	5	6
Angle tolerances for each triplet type (degrees)	HHH	25	35
	HHS	20	25
	HSS	15	20
	SSS	10	20

[†]User-defined parameters used in the similarity searches and in the all-by-all structural comparison. H, helix; S, strand. Angular and length tolerances are set at varying levels for each triplet type to enhance user control over how stringently different triplet types are matched.

match to within user-defined tolerances to qualify a match (Table I).

Unlike in the fingerprint comparison stage, some attention is given in the triplet-based comparison stage to the topological order of secondary structure elements. The greedy algorithm iterates from the "most N-terminal-SSE-containing" triplets towards the C-terminal triplets in the two proteins, and makes as many matches as possible during the iteration. In a greedy algorithm, there is no optimization of the matches; therefore there may be suboptimal equivalences found by this method. We find that the speedup with a greedy algorithm is substantial, resulting in comparison times of a few seconds for large α/β barrel-containing proteins and other highly symmetric structures.

The final similarity score assigned to a structural pair is based on the number of matching triplets. The similarity score equation we have chosen is the Dice coefficient. In our experience, the Dice coefficient is less sensitive to simple differences in structural size than the Tanimoto score and is therefore more appropriate to use at this stage of the comparison (see below). The equation for the Dice coefficient is

$$D_{A,B} = (2 * N_m)/(N_A + N_B) \quad (4)$$

where N_m is the number of matching triplets, N_A is the total number of triplets in protein A, N_B is the number of triplets in protein B, and $D_{A,B}$ is the Dice coefficient, which ranges between 1 (identical) and 0 (completely different).

The Dice similarity score represents the percent of secondary structure triplets that match out of the total number of possible matches. As the number of triplets is dependant on the number of secondary structure elements, this score is correlated to the percent of matching secondary structure elements out of the total number of SSEs in the two structures under comparison. The Dice score is informative when we consider protein structures as a whole rather than as assemblages of domains. Two similar structures which are approximately equal in size will have a Dice score near 1. However, in cases where there is a substantial size mismatch between two structures sharing a common substructure, the Dice score should be lower. Those structures that share substantial degrees of similarity, therefore, will have Dice scores higher than those that share a common subdomain (which in turn will have scores higher than those without any similarity).

Construction of the Protein Structure Universe

The first step in mapping the structural universe was to prepare a representative protein database. We constructed a large database from the set of all nonidentical chains in the July 1996 release of the PDB. Identical chains with and without substrate bound were included in the database as well, in order to address the possibility of structural rearrangements upon ligand binding. The final database contains 1,031 protein chains. We performed an all-by-all pairwise comparison of the chains in the database. For each structure, all other structures in the database were passed through the prescreening step of the algorithm to the triplet-based comparison stage, and ranked according to their Dice similarities. The Dice similarity scores, subtracted from 1, can be thought of as "dissimilarity distances" separating each protein from every other protein in the structural space. The similarity distance relations were compiled into a pairwise distance matrix, which was used as input to a principal components analysis to calculate the distribution of structures within the protein universe.

RESULTS

Three-Dimensional Similarity Searches

Five test cases, ranging in difficulty from easy to challenging, have been used to assess the accuracy of the similarity searching algorithm. For each test case, one member of the structural family is used as a search key in a database search. Although these results are limited to specific examples, we feel the selectivity of the algorithm can be evaluated from the rankings obtained by the members of the search key's structural class. The test cases examined in this study were: the globins, the trypsinlike mammalian serine proteases, the immunoglobulin G Fab variable heavy chains, the cupredoxin fold family, and the actinlike ATPase domain-containing proteins. The user-defined tolerances used for each test case are listed in Table I.

The globin fold has a structural core of six helices in a partially opened folded-leaf pattern. The protein used as the search key in this test case was sperm whale myoglobin (1mbd). The prescreen enrichment plot is shown in Figure 2A. The enrichment of the database in globin structures by the prescreening stage is near-perfect. This

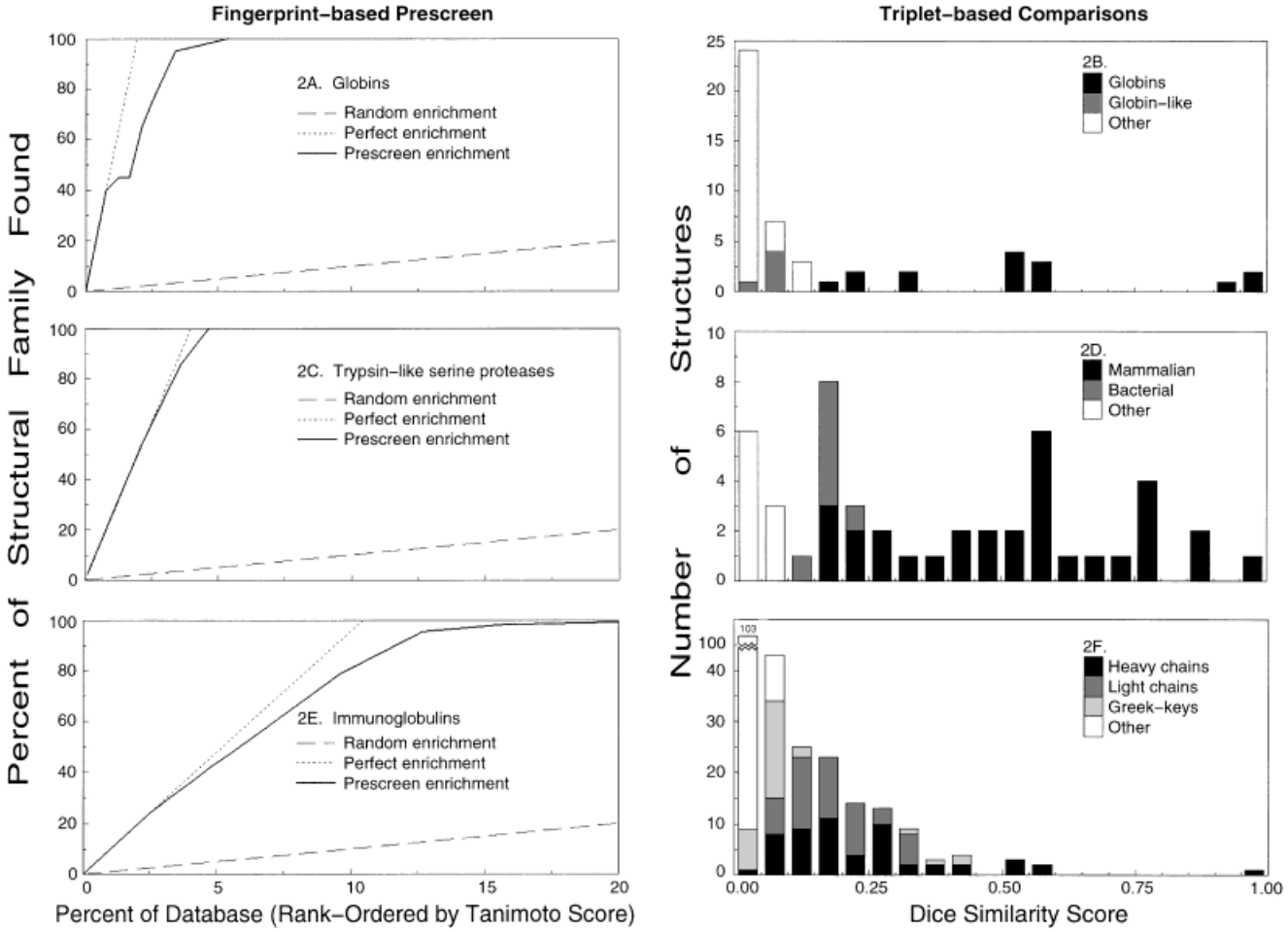

Fig. 2. Prescreen enrichment plots and similarity score distributions for three test cases: the globins (**A, B**), the trypsinlike serine proteases (**C, D**), and the immunoglobulins (**E, F**). The prescreen stage is nearly perfect in ranking members of each structural family higher than other structures in the database (panels A, C, E). These structures, when passed to the triplet-based comparison stage, can be distinguished from nonmember structures (panels B, D, F).

stage showed exceptional specificity for globin structures, although it had difficulty distinguishing some all-alpha nonglobin structures from true members of the globin family. The secondary stage was much more sensitive and specific; the 15 globins in the database were ranked as the top 15 scoring proteins in the database (Fig. 2B).

The trypsinlike serine protease fold exhibits a greek-key domain duplication. Two quite similar all-beta domains are arranged asymmetrically in this class of structures. A search was performed with bovine pancreas trypsinogen (1tgn). The results of this search are shown in Figure 2C,D; 982 of 990 proteins dissimilar to 1tgn (99.2%) were filtered out during the rapid prescreening stage (Fig. 2C). The second stage not only scores all mammalian trypsin-like serine protease structures higher than all other proteins in the database, but it also identifies the bacterial trypsinlike serine proteases as members of this structural family (Fig. 2D).

The immunoglobulins are largely all-beta proteins, with seven core strands forming two sheets in the greek-key β-sandwich fold. The database was searched with the heavy chain of an Fab (BV04–01) mouse autoantibody (Fig. 2E,F). The immunoglobulins as a class exhibit more structural variability than the first two test cases, with heavy chains, light chains, and other greek key-containing proteins populating the database. As a result, there was more spread in the score distributions for both the first and the second stages of the similarity calculations. However, the high similarity score region of the secondary stage distribution was populated predominantly by immunoglobulin heavy chains. Light chains were also identified by the algorithm as being highly homologous in structure to the search key, and this is reflected by the distribution of light chains with lower Dice similarity scores. The lectins were identified as the highest-scoring nonimmunoglobulins with Dice scores equal to approximately 0.06. This result is not surprising. The lectins share a consensus motif of five β strands and a turn with the immunoglobulin variable domain.[35]

We next move to proteins sharing more remote structural homologies. Searches were performed for members of the cupredoxin fold family and for proteins sharing a

TABLE II. Azurin Similarity Search Results†

PDB Id	Name	Dice score	RMSD to 1aizA	DALI Z score	Alignment length
1aizA	Azurin	1.00	0.0	30.2	129
1aac	Amicyanin	0.60	2.4	7.1	80
2pcy	Apo-plastocyanin	0.37	3.6	4.6	87
1tsrA	p53 tumor suppressor	0.32	3.9	2.0	84
1tupB	p53 tumor suppressor	0.31	3.5	2.7	82
1tupA	p53 tumor suppressor	0.32	3.9	2.0	85
1xsoA	Cu,Zn SOD	0.32	3.6	2.1	80
1sxaC	SOD (Cu+)	0.32	3.5	1.8	71
2pcdA	Protocatechuate 3,4-dioxygenase	0.27	3.3	1.3	61
1cyx	Quinol oxidase	0.26	2.3	6.6	89
1cwpC	Cowpea chlorotic mottle virus	0.25	3.4	2.1	74
1ncg	*N*-cadherin	0.23	3.1	2.1	62
1cauB	Canavalin	0.23	3.7*	—	25*
1pmy	Pseudoazurin	0.21	2.8	6.2	83
2pcdQ	Protocatechuate 3,4-dioxygenase	0.20	3.7	1.1	65
2pabA	Prealbumin	0.20	2.9	2.5	65
1roxA	Transethyretin	0.20	3.0	2.6	65
1rsy	Synaptotagamin I	0.19	2.9	3.3	75
1htp	H-protein	0.19	2.1*	—	35*
1f3g	Phosphocarrier III	0.19	4.3*	—	28*

†Twenty structures most similar to azurin (1aizA) in the database of 1,031 structures (based on Dice similarity score). Topological alignments to azurin were made using DALI.[13] Nontopological alignments were made manually and are indicated with an asterisk. Structure pairs with DALI z scores <2 are considered dissimilar.

ribonuclease H-like folding motif. As the proteins in these two cases are more distant structural relatives, we relaxed the user-defined tolerances to perform the similarity searches (Table I). The range of Dice scores, therefore, for the following two examples should not be directly compared with those for the globins, serine proteases, and IgG structures as they result from different user-defined tolerances.

Azurin, plastocyanin, and amicyanin are all monodomain cupredoxins (blue or type-1 copper proteins). Members of this family are defined by SCOP to share a common structural core—a seven-stranded, two-sheet greek-key β sandwich and a conserved copper binding site.[36] However, they can share as little as 10% sequence identity.[37,38] The monodomain cupredoxins exhibit a great deal of structural variation, especially in peripheral secondary structure elements. Searching the database with azurin identified amicyanin and plastocyanin as the most structurally similar proteins, with similarity scores of 0.60 and 0.37, respectively (Table II). Other members of the monodomain cupredoxin family, quinol oxidase (1cyx) and pseudoazurin (1pmy) were ranked, respectively, at 10 and 14 with scores of 0.26 and 0.21.

All 20 top scoring structures contained a 7- to 9-stranded β-sandwichlike fold. The high-scoring proteins lacking the cupredoxin fold contained either a 7- to 9-stranded 2-sheet greek-key β-sandwich (1xso, 2pdcA, 2pcdQ, 1ncg, 2pabA, 1roxA, 1rsy, 1tupA 1tsrA, 1tupB), a 7- to 8-stranded 2-sheet half-β-barrel (1htp, 1f3g), an 8-stranded 2-sheet jellyroll β-sandwich (1cwpC), or a double-stranded β helix (1cauB), which is a distinctive β-sandwichlike architecture with a jelly-roll topology.[27] These architectures are similar in their SSE compositions, arrangements, and root-mean-square (RMS) distances to azurin. For example, Ryden[36] aligned the structure of Cu-Zn superoxide dismutase (SOD) to poplar plastocyanin with a RMS distance of 2.99 Å over 68 of 99 total alpha-carbons, which supported speculations that Cu-Zn SOD is distantly related to the blue copper proteins . Also, many top-scoring proteins with the greek-key β-sandwich fold can have substantial portions of their backbone aligned to azurin with RMS distances of approximately 3 Å (Fig. 3 and Table I). Since our algorithm does not encode functional information (e.g., the presence of copper binding sites), these proteins were considered similar to members of the monodomain cupredoxin structural family.

Hexokinase, glycerol kinase, actin, and the heat shock protein 70 kDa ATPase fragment all contain a duplication of the ribonuclease H-like motif of three α/β/α layers.[38] The common fold consists of two α/β domains, each containing topologically identical five-stranded β sheets. Each mainly parallel β sheet is made up of five strands with strand 2 antiparallel with respect to the four other strands. The structural similarities shared by proteins in this superfamily are more difficult to detect, and this is the level of structural homology at which threading methods can do poorly.[39] A search was conducted with hexokinase B as the search key. The top 20 scoring proteins in the database are shown in Table III. Five members of this structural family are represented in the top 20. One member, hexokinase A, did not score highly because DSSP did not define six strands in the hexokinase A structure that were defined for hexokinase B. The two structures found to be the most similar to hexokinase B were glycerol kinase and actin, both of which contain a duplication of the ribonuclease H-like motif. Ribonuclease H is not considered to be a

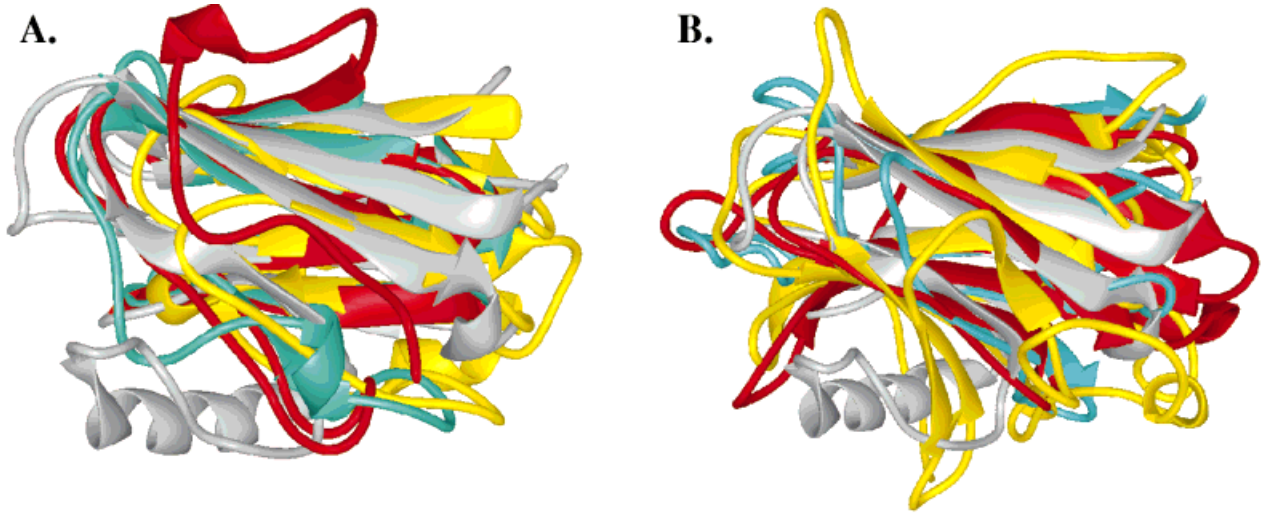

Fig. 3. **A:** Multiple structure alignment of members of the cupredoxin fold family. Azurin is indicated in gray, plastocyanin (3.6 Å RMSD/87 residues) in cyan, amicyanin (2.4 Å/80) in red, and pseudoazurin (2.8 Å/83) in yellow. **B:** Multiple structure alignment of azurin and noncupredoxin structures ranked in the top 20. Azurin is shown in gray, prealbumin (2.9 Å/65) in cyan, synaptotagmin I (first C2 domain, 2.9 Å/75) in red, and tumor suppressor p53 (3.5 Å/82) in yellow.

TABLE III. Hexokinase B Similarity Search Results†

PDB Id	Name	Dice score	RMSD to 2yhx	DALI Z score	Alignment length
2yhx	Hexokinase B	1.000	0.0	65.5	457
1glaG	Glycerol kinase	0.385	4.6	9.3	240
2btfA	β-Actin	0.345	5.0	8.5	230
1gtrA	Glutaminyl-tRNA synthetase	0.344	5.4	1.9	82
1bncA	Biotin carboxylase	0.342	3.9	1.7	69
1ace	Acetylcholinesterase	0.341	4.5	0.7	114
1dge	Dialkylglycine decarboxylase	0.333	5.2	0.6	80
6taa	α-Amylase	0.329	4.4	0.5	102
2dkb	2,2-Dialkylglycine decarboxylase	0.328	5.2	0.6	80
1scuB	Succinyl-CoA synthetase	0.326	4.1	2.4	81
1gal	Glucose oxidase	0.318	3.2*	—	59*
1hpm	44K ATPase fragment, 70 kDa heat shock protein	0.316	5.1	9.2	227
2dln	D-alanine-D-alanine ligase	0.308	4.0	0.5	53
1minA	Nitrogenase Mo-Fe protein	0.307	4.6	1.0	149
1gsh	Glutathione biosynthetic ligase	0.304	4.4	0.7	69
3hsc	44K ATPase fragment, 70 kDa heat shock protein	0.302	5.2	9.0	229
5rubA	Rubisco	0.302	3.9*	—	48*
2aaa	Acid α-amylase	0.302	5.7	0.7	102
1adeB	Adenylsuccinate synthetase	0.300	4.2	1.0	79
1nhp	NADH peroxidase	0.299	4.0	0.9	53

†Twenty structures most similar to hexokinase B (2yhx) in the database of 1,031 structures (based on Dice similarity score). Topological alignments to hexokinase B were made using DALI.[13] Nontopological alignments were made manually and are indicated with an asterisk. Structure pairs with DALI z scores <2 are considered dissimilar.

member of this structural family, since it contains one, and not two, domains with the ribonuclease H-like motif.

The 15 structures in the top 20 hits that lack an actinlike ATPase domain represent 4 distinct architectures. Group I contains 3 α/β/α layers, with a parallel 5- to 7-stranded central sheet (1scuB, 1minA, 1ade, 1gal, 1gtrA). The second group also shares the 3 α/β/α-layer architecture, but the central sheet is mixed and ranges in size from 4 to 8 strands (1dge, 2dkb, 1ace,1bncA, 2dln). Group III consists of α/β (TIM) barrel-containing proteins (2aaa,6taa, 5rubA) consisting of closed 8-stranded β-sheet barrels. Group IV has only one member: the lowest scoring structure of the top 20 ranking proteins, NADH peroxidase (1nhp). NADH peroxidase has a β/β/α architecture with a central 5-stranded parallel β sheet.

The β-sheet structures of hexokinase B and these 15 proteins accounts for the majority of their structural similarity. The two β sheets in hexokinase B are curved in a near-barrellike manner. Also, each sheet is surrounded by helices. The location and curvature of these two β

sheets and the positions of several peripheral helices overlap in many of the 15 proteins without an actinlike ATPase domain listed in Table III. However, a direct correlation between the similarity score and the RMS distance for the alignment of each protein and hexokinase B is not observed. Since our method weights helices and strands equally, the similarity scores we calculate do not correlate well with residue-based measures such as RMS distance, which implicitly weight helices more highly than strands due to the fact that helices generally contain more residues than strands.

Tables II and III indicate that, although our method is capable of recognizing homologous proteins, it is not as sensitive as the DALI method for discriminating homologous and analogous folds. For example, in the cupredoxin case, structural "second cousins" (defined by a DALI z score ≥ 4) were not clearly distinguished from other members of the 7-stranded, 2-sheet greek- key β-sandwich fold family (Table II). Analogously, for the ATPase domain-containing proteins, structural "cousins" (DALI z score ≥ 8) were not always considered to be more similar than analogous protein structures (Table III). Since our method treats protein structures at lower resolution, it is likely that residue-level details important for distinguishing structural subclasses are lost when comparisons of remote homologs are made. Also, for extremely remote homologs, variability in the SSE arrangements of these structures may be sufficently pronounced to escape their detection as being more structurally homologous than a few analogous folds. For these reasons, we feel that the value of our method at its current level of resolution resides primarily as a prescreening tool to be used prior to more detailed structural comparison techniques.

Analysis of the Protein Structure Universe

After performing the calibration experiments, we applied our method to the characterization of the protein structure universe. To this end, an all-by-all comparison of the 1,031 chains was performed, and for each chain in the database a near-neighbor list was constructed containing all the structures in the database, ranked in descending order by their Dice similarity score. We have used the $1{,}031 \times 1{,}031$ distance matrix defined by this calculation in clustering analyses and principal components analysis to extract relations shared by subpopulations of protein structures.

The user-defined tolerance values used for this calculation were the same as those used for the immunoglobulin, trypsinlike serine protease, and globin similarity searches. We chose these more conservative values to promote the clustering of structural nearest-neighbors in the space. Of course, the tolerances could be relaxed, and the universe recalculated, in an attempt to map more distant similarity relations in greater detail.

A Scree plot shows that 50% of the variance is encapsulated by the top six principal components (Fig. 4). A total of 94 principal components are required to represent 95% of the variance in the set of protein structures. The principal components analysis substantially reduced the complexity of the space, since the maximum number of orthogonal axes possible for a set of size N is $N - 1$ axes (if all members of the set are unrelated), which in this case would be 1,030 axes. The reduced dimensionality we observe is due to the fact that there are strong interrelations among existing protein structures.

Figure 5 is a three-dimensional projection of the protein structural universe as defined by its top three principal components. Each point represents one protein structure; its placement within the space is determined by the distance relations between that structure and all other structures in the dataset. All points have been colored based on their SSE composition, with the all-beta structures in red, all-alpha structures in blue, with a gradient of shades for α/β and α + β structures. The three-dimensional projection shows that principal component 1 (PC1) serves to separate α/β and α + β proteins from the all-alpha and all-beta superfamilies. Segregation of the all-alpha and all-beta structures occurs primarily along principal component 2 (PC2).

PC2 also partitions the immunoglobulin G (IgG) structural family from the remainder of the universe (Fig. 6A). The immunoglobulins are seen populating a distinctive "plume" extruding from the center of the universe along PC2. The immunoglobulin light chains and heavy chains are in adjacent plumes parallel to PC2 and perpendicular to PC1, with their relative PC1 positions determined by the SSE composition of each family. The spread of the IgG light chain family along PC2 is due to structural variations within the family itself (Fig. 6B,C). The angle between the Fc and V domains of the structures in clusters 1 and 2 differs by approximately 30 degrees.

This unusual degree of flexibility results in an uncommonly large spread of IgG light chain structures along PC2. The relatively invariant intradomain SSE triplets shared by the IgG light chains are sufficient to cluster them along PC2. However, unlike domain-based clustering techniques, which would represent the immunoglobulin structural family as a point in a theoretical fold space, the variation we observe is due to the loss of matching interdomain triplets. The information contained in the intra- and interdomain SSE triplets resulted in two levels of organization: the classification of the IgG light and heavy chains as two distinct, but similar structural families, and the organization of the structures within each family on the basis of interdomain positional variations. We feel that this detailed organization of the IgG chains demonstrates that our approach can automatically define a reasonable intrafamilial classification of proteins based upon interdomain variations in tertiary structure.

A similar partitioning is observed for principal component 3 (PC3, Fig. 7A). The trypsinlike serine proteases are distributed away from the main body of structures along PC1 and PC3. Again there is a structural gradient which organizes structures within the trypsinlike serine protease family, but in this case, the separation is due to the varying number of peripheral helices observed in structures of this family. Those structures with four helices are positioned closer to the α/β extremum, and those struc-

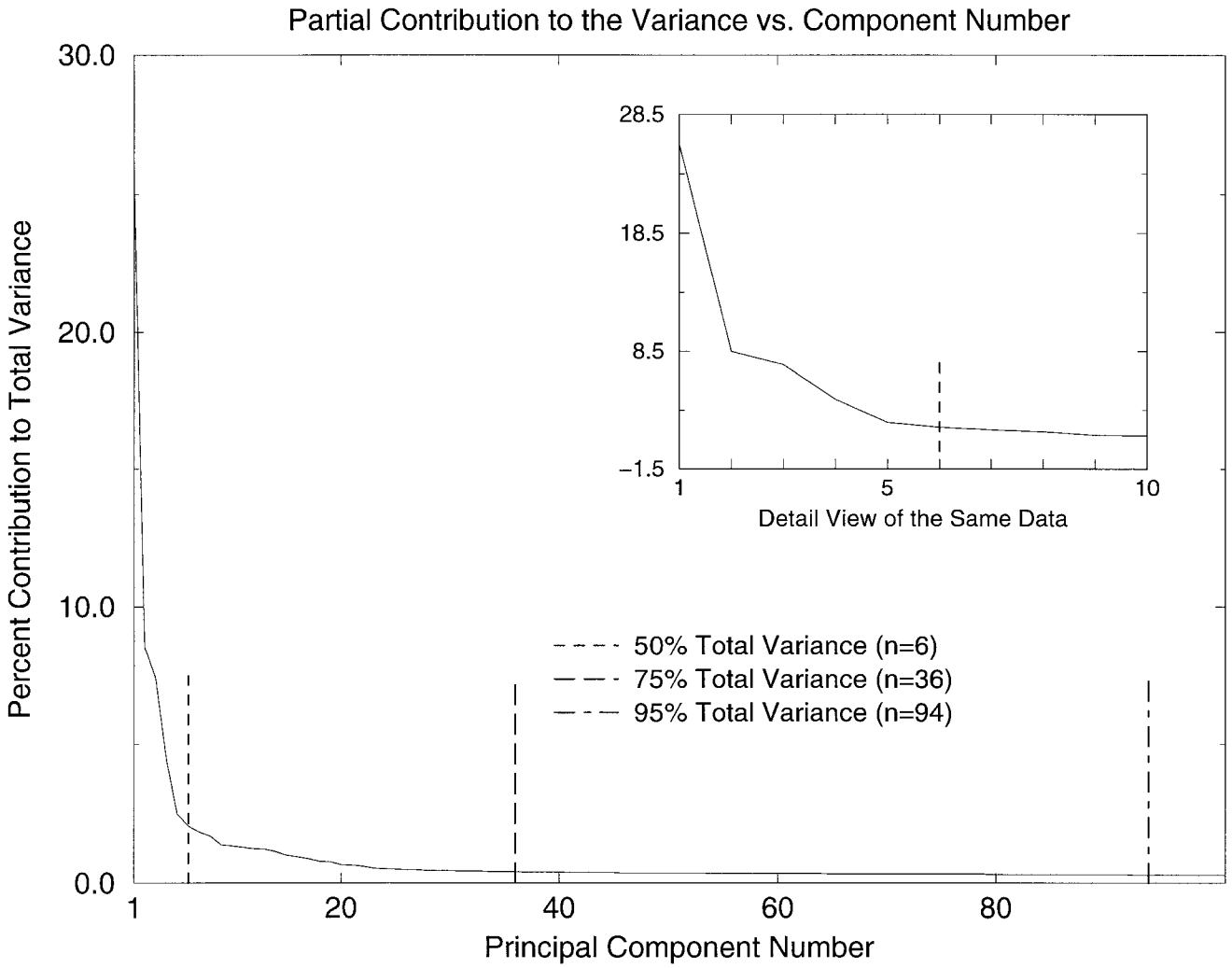

Fig. 4. Scree plot of the percent variance captured by each principal component. The three vertical lines indicate the number of axes needed to encapsulate 50%, 75%, and 90% of the total variance.

tures with one helix are near the all-alpha/all-beta extremum of PC1 (Fig. 7B,C). The organization of these structures along PC3 appears to reflect the subtle structural variations within each subgroup in a manner similar to that observed for the IgG light chain family along PC2.

To address the robustness of these results, we decided to reduce the effect of large families of similar structure by using a structure-based clustering of the protein database. We extracted 140 clusterheads at a Dice similarity threshold of 0.10 to produce a representative spanning subset in order to remap the protein structure universe using this diverse sample. Upon recalculation of the distance matrix defined by the 140 representatives and subsequent PCA, we were able to define their positions within a simplified high-dimensional structure space (Fig. 8A). The Scree plot for this analysis (Fig. 8B) shows that many more axes, 49 to be precise, are required to contain 50% of the variance. This is directly due to removing multiple representatives of specific structures, leaving a largely orthogonal set. Interestingly, the arrangement of structures along PC1 appears to be relatively unchanged. While PC2 and PC3 are not the same axes as those observed for the larger universe, simply because the overrepresentation of the immunoglobulins and trypsinlike serine proteases has been eliminated from the representative set, the positions of outliers in both the full and reduced structural universes along PC1 remain essentially constant. This result indicates that the basic organization of the universe has remained unchanged as the axes associated with specific families were modified.

DISCUSSION

Our protein structural comparison algorithm is a rapid method for protein structure-based similarity analysis. It can identify both obvious structural similarities, such as seen in the globin fold family, and those that are more subtle, as the cupredoxins. The rapidity of our approach

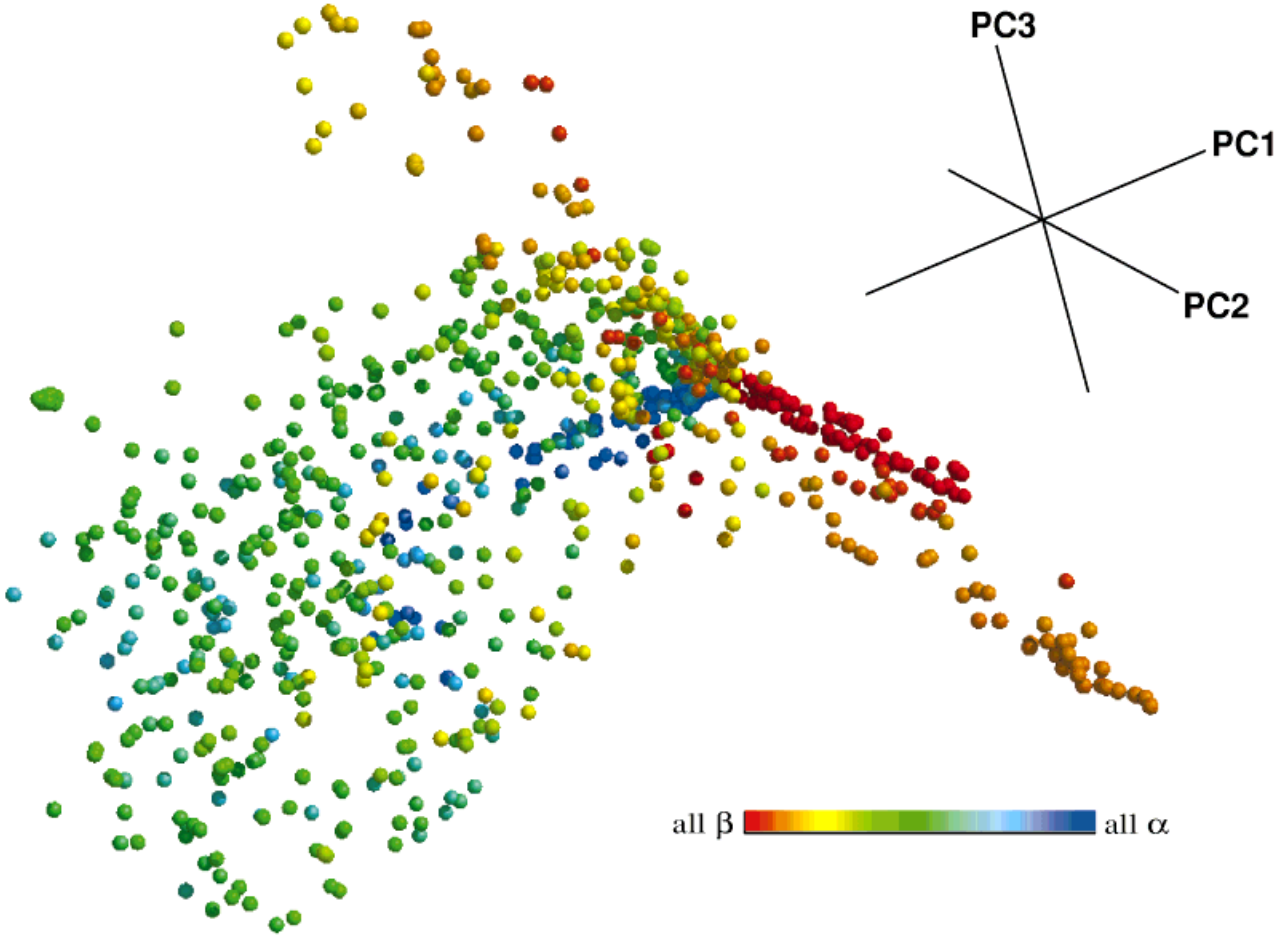

Fig. 5. Three-dimensional projection of the protein structural space as defined by the top three principal axes (PC1, PC2, PC3). Protein SSE compositions are shown, ranging from red (all-β) to blue (all-α). PC1 is an all-α/all-β → α/β axis. PC2 primarily separates the immunoglobulin light and heavy chains from the rest of the protein structure universe. PC3, in part, serves to segregate the trypsinlike serine proteases.

permits the entire PDB to be clustered at the tertiary structure level without requiring the prior reduction of the database to a small representative set of structures. We have presented here an application of the method to mapping the structure space as defined by the similarity relations shared by 1,031 structures. Principal components analysis of the dataset shows that half of the variance can be represented by six principal components, and 95% of the variance is contained within 94 principal components. By these similarity measures, the protein structural universe is a high-dimensional object. Therefore, two- or three-dimensional representations of the fold universe may be overly simplistic portrayals of the protein fold distribution, unless some other organizing measures can be identified.

Clearly the overall shape of the protein universe and many of the detailed features will be functions of the protein structures and the mathematical machinery used in the analysis, for example, the choice of domain or full-protein structures, the secondary structure assignment algorithm, the hashing algorithm, the choice of fundamental descriptors, and the similarity measure. We chose a secondary structure-based representation of protein tertiary structures because it both captures salient information about protein architecture at the fold level and it reduces the computational complexity of comparing protein structures. Our method treats protein structures as sets of SSE triplets. Triplets contain correlations that are not present in the pairwise description of protein structures used by VAST. The enrichment of information is important because SSE interactions are necessarily descriptions of protein structures.

The similarity measure we use in our analysis is the Dice score, which is the percent SSE triplet identity. We feel the Dice score ranks protein structure similarities appropriately when protein structures are viewed as a unified whole rather than as an assemblage of domains. The important question of the best basis set for representing protein structural diversity has no simple answer. Choosing domains to be the elementary units will provide fingerprints that recognize these substructures but which are less effective at matching entire proteins. Either approach has important uses.

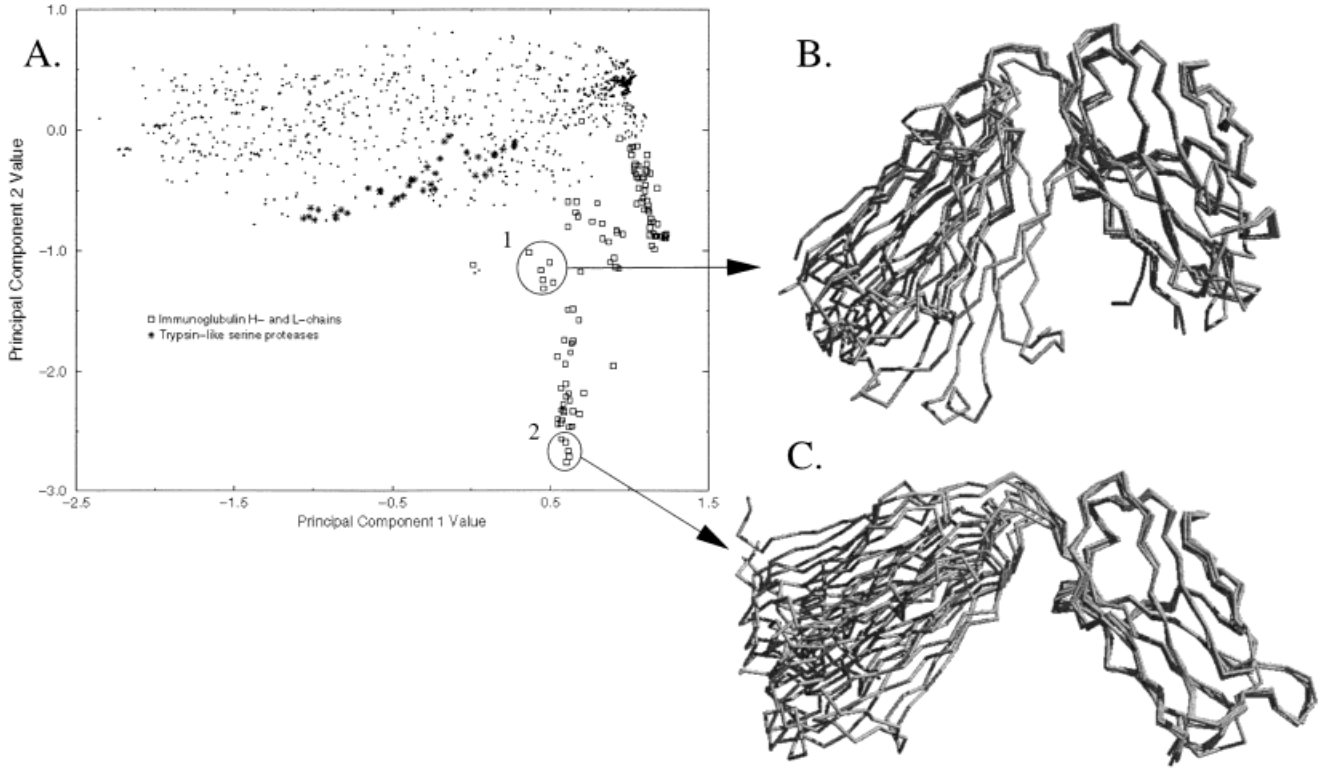

Fig. 6. Structural variability within the family of immunoglobulin G light chains. **A:** Two-dimensional projection of the protein universe as defined by principal components 1 and 2. Open squares represent the immunoglobulin light and heavy chains; stars mark the location of trypsinlike serine proteases. **B:** Multiple alignment of the variable domains of four structures in cluster 1 (1bjmA, 1indL, 1mfeL, 1ngpL). Two other structures (8fabA, 1indL) contain constant domains that are twisted nearly perpendicular to the other cluster 1 and all cluster 2 structures. **C:** Multiple alignment of the variable domains from structures in cluster 2 (1dbjL, 1igiL, 1bafL, 2cgrL, 1fvdA). Cluster 2 contains structures which are more extended than those in cluster 1.

We find that with the protein structural representation and similarity score we have chosen, some general features are seen that are likely to emerge from any reasonable representation. We note, for example, that a hierarchy of structural organization exists in our model of the protein structure universe. At the highest level, there is a partitioning of structures into all-alpha, all-beta, and α/β superfamily regions in the three-dimensional projection defined by the top three principal components. This result is similar to that observed by Sowdhamini et al.[19] in their calculation of a protein domain universe. A lower level of organization can be seen as well; principal components 2 and 3 separate specific fold families from the body of the structural universe. The families appear as elongated "plumes" along these axes in a manner reminiscent of the "fold attractor" hypothesis introduced by Holm and Sander.[30] The lowest level of organization is intrafamilial, and is represented by the spread of structures along a family-specific axis. For the immunoglobulin light chains, the variation in structural positions along the axis reflects the spectrum of relative positions of the Fc and V domains (Fig. 6). Analogously, the gradient of structures along the trypsinlike serine protease axis orders the proteins based on the number of peripheral helices (Fig. 7). Defining the organization of the protein structural universe, however, is but one of many interesting applications of our approach. Currently, the algorithm encodes information about the spatial arrangement and topology of α helices and β strands. However, the set of essential structural features that define biologically significant similarities are much broader and richer in scope. Since our method encodes a set of simple descriptors as a binary fingerprint, we are relatively free to extend our algorithm to include additional structural features such as loop centroids, atomic positions, and metal binding sites. This will give us the opportunity to explore the relative importance of each descriptor in defining biologically relevant structural and functional similarities.

Methods such as ours, which can compare large and complicated protein structures rapidly, may also be able to address a variety of questions concerning the organization of protein structural families, such as: How are protein families interrelated at a purely structural level? What are the boundaries of structural families? Are families tightly or loosely knit? How are families related and what appears to be the evolutionary mechanism(s) responsible for these interrelations? With rapid structural comparison approaches it is possible to address these questions quantita-

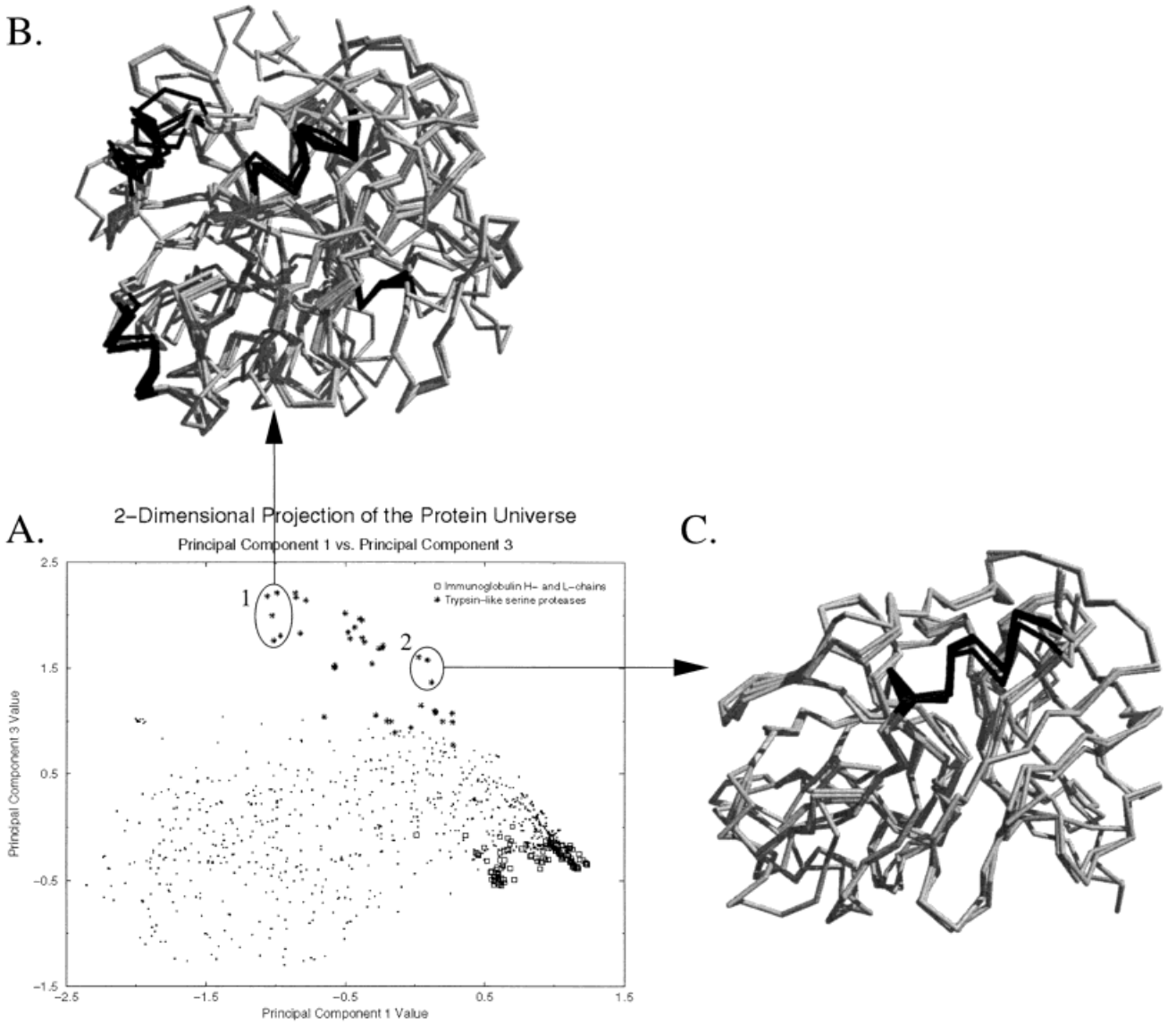

Fig. 7. **A:** Two-dimensional projection of the protein universe as defined by principal components 1 and 3. Open squares are the immunoglobulin light and heavy chains; stars are the trypsinlike serine proteases. Two clusters of trypsinlike serine protease structures located near the termini of the trypsinlike serine prtease "plume" were aligned using DALI for further analysis. **B:** The structural alignment of the five structures in cluster 1 (1bbrK, 1hahH, 1hcgA, 1hyl, 1hagE). Members of this cluster all contain four DSSP-assigned peripheral α helices, which are shown in black. **C:** Structural alignment of three structures in cluster 2 (1hneE, 3rp2, 1ppfE). All three structures contain only 1 α helix, shaded black for clarity. Structures intermediate to clusters 1 and 2 have two or three helices, with those structures containing three helices located near cluster 1, and those with two helices near cluster 2.

tively. Furthermore, the sensitivity of our algorithm to structural variations within families may give us the opportunity to organize families into structural subgroups to define subtle evolutionary and/or functional relations. Studies such as these may provide a quantitative method to deconvolute the structure–function interrelationship on a per-family basis.

Defining the boundaries of protein structural space should also allow us to quantify the population density of regions within the universe. Highly populated regions may correspond to Holm and Sander's "attractor" regions, populated either because they correspond to minima in the folding energy landscape, or because they are evolutionary "sinkholes," which are occupied by structures that diverged from a ancient common ancestor. Those regions that are unpopulated may offer us many intriguing clues to the rules governing protein structure and folding. Some unpopulated regions may be energetically or kinetically inaccessible. Others may correspond to acceptable structures which have not been successful evolutionarily, or

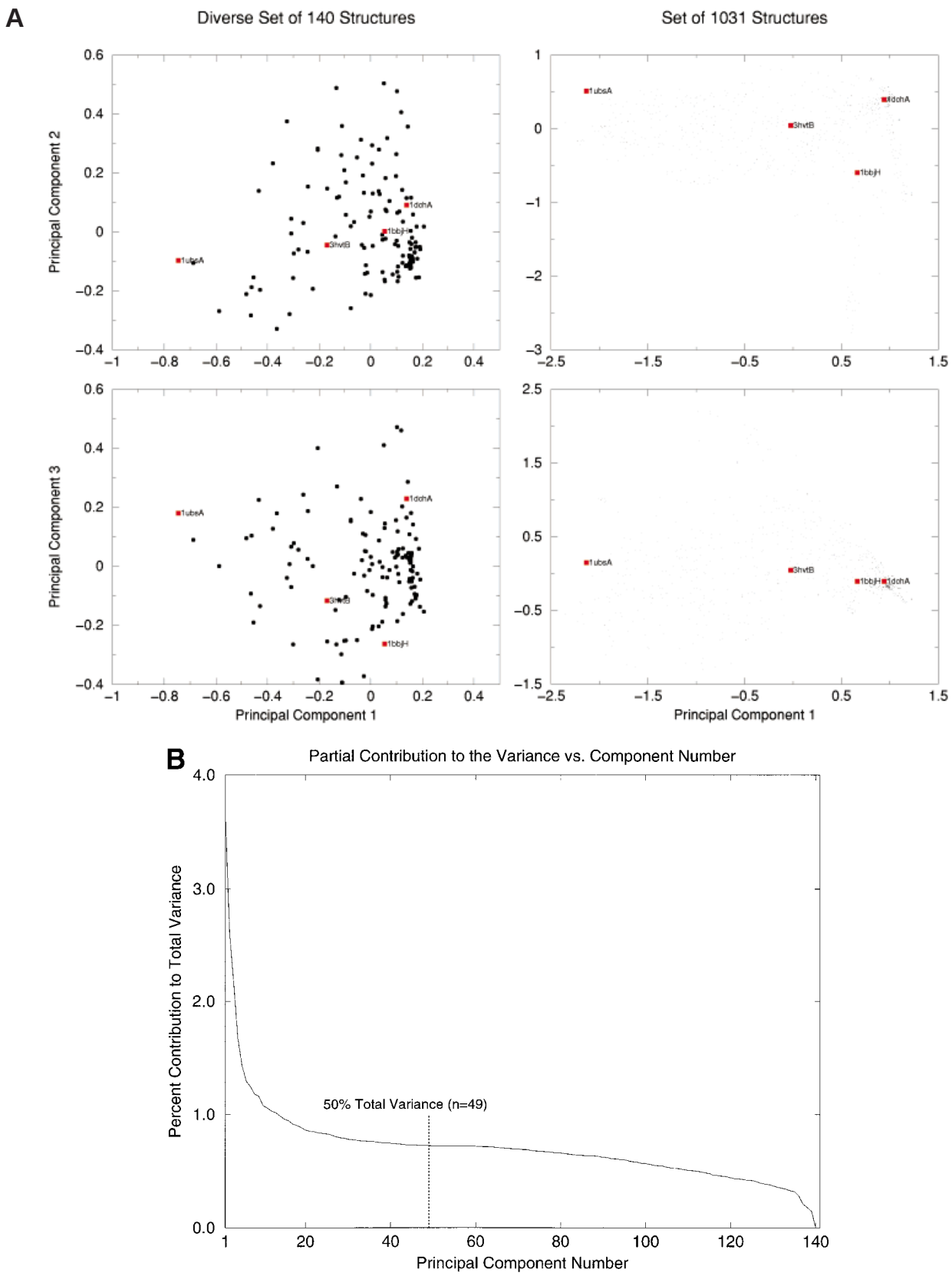

Fig. 8 **A:** Structural distributions for the reduced (140) and the full (1,031) sets of protein tertiary structures. The positions of four structures are labeled and indicated in red to facilitate comparison of the two-dimensional projections. **B:** Screen plot of the percent variance captured by each principal component for the set of 140 structural representatives. The vertical line marks the number of axes needed to capture 50% of the total variance.

which may not have evolved at all. We may be able to discern the type and number of structures that could theoretically inhabit these empty areas and perhaps understand why we have not yet seen examples of these structures.

Finally, since we use a low-resolution representation of protein tertiary structure in our algorithm, we may be able to apply it to the validation of models generated by *in silico* folding experiments. If a model can be placed within a populated region of the known protein structure universe, it may indicate that the model may be closer to a folding minima than a model that occupies an isolated position in the universe. In future work, we plan to pursue a population analysis of the protein universe and several practical applications of our algorithm.

CONCLUSION

We have developed an automatic protein structural comparison method that uses secondary structure element compositions, spatial arrangements, lengths, and topologies to identify tertiary structure similarities. Our method identified proteins sharing structural homologies in five test cases: the globins, the mammalian trypsinlike serine proteases, the immunoglobulins, the cupredoxins, and the actinlike ATPase domain-containing proteins. Principal components analysis (PCA) of a similarity distance matrix calculated from an all-by-all comparison of 1,031 nonidentical chains from the PDB produced a distribution of structures within a high-dimensional structural space. Six axes were required to encapsulate 50% of the variance, two of which were associated specifically with the immunoglobulins and the trypsinlike serine proteases. Some features of the universe remained stable upon reduction of the database to 140 structures and subsequent PCA, although the axes correlated with specific families were no longer observed. Upon further analysis of the universe, we found that the structures were organized hierarchically. At the highest level, protein structures were segregated into all-alpha, all-beta, and α/β superfamily regions. An intermediate level was composed of family-specific axes that partitioned large structural families from the bulk of the universe, forming local highly populated regions within the space. The lowest level of organization was intrafamilial. Homologous structures were arranged along a gradient defined either by variations in peripheral secondary structure elements or by conformational shifts in the tertiary structure.

ACKNOWLEDGMENTS

We thank Patricia Babbitt for many informative discussions, and Todd Ewing for the use of his software.

REFERENCES

1. Chothia C, Lesk A.M. The relation between the divergence of sequence and structure in proteins. Embo J 1986;5:823–826.
2. Allen FH, Bergerhoff G, Sievers R. Protein data bank. In: Crystallographic databases: information content, software systems, scientific applications. Bonn: Data Commission of the International Union of Crystallography. 1987:107–132.
3. Bernstein FC, Koetzle TF, Williams GJB, Meyer EF, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. The protein data bank: a computer-based archive file for macromolecular structures. J Mol Biol 1977;535–542.
4. Remington SJ, Matthews BW. A systematic approach to the comparison 112:535–542, of protein structures. J Mol Biol 1980; 140:77–199.
5. Taylor WR, Orengo, CA. Protein structure alignment. J Mol Biol 1989;208:1–22.
6. Sali A, Blundell TL. Definition of general topological equivalence in protein structures: a procedure involving comparison of properties and relationships through simulated annealing and dynamic programming. J Mol Biol 1990;212:403–428.
7. Holm L, Sander C. Protein structure comparison by alignment of distance matrices. J Mol Biol 1993;233:123–138.
8. Yee DP, Dill KA. Families and the structural relatedness among globular proteins. Prot Sci 1993;2:884–899.
9. Alexandrov NN, Go N. Common spatial arrangements of backbone fragments in homologous and non-homologous proteins. J Mol Biol 1992;225:5–9.
10. Vriend G, Sander C. Detection of common three-dimensional substructures in proteins. Proteins 1991;11:52–58.
11. Fischer D, Bachar O, Nussinov R, Wolfson H. An efficient automated computer vision based technique for detection of 3-dimensional structural motifs in proteins. J Biomol Struct Dyn 1992;9: 769–789.
12. Mitchell EM, Artymuik PJ, Rice DW, Willett P. Use of techniques derived from graph theory to compare secondary structure motifs in proteins. J Mol Biol 1989;212:151–166.
13. Grindley HM, Artymiuk PJ, Rice DW, Willett P. Identification of tertiary structure resemblance in proteins using a maximal common subgraph isomorphism algorithm. J Mol Biol 1993;229: 707–721.
14. Bakarat MT, Dean PM. Molecular structure matching by simulated annealing. III. The incorporation of null correspondences into the matching problem. J Comput Aided Mol Design 1991;5: 107–117.
15. Holm L, Sander C. DALI: a network tool for protein structure comparison. Trends Biochem Sci 1995;20:478–480.
16. Orengo CA, Brown NP, Taylor WR. Fast structure alignment for protein databank searching. Proteins 1992;14:139–167.
17. Alexandrov NN. SARFing the PDB. Prot Eng 1996;9:727–732.
18. Fischer D Tsai CJ, Nussinov R, Wolfson H. A 3D sequence-independent representation of the Protein Data Bank. Prot Eng 1995;8:981–997.
19. Sowdhamini R, Rufino SD, Blundell TL. A database of globular protein structural domains: clustering of representative family members into similar folds. Fold Design 1996;1:209–220.
20. Madej T, Gibrat JF, Bryant SH. Threading a database of protein cores. Proteins 1995;23:356–369.
21. Alexandrov NN, Fischer D. Analysis of topological and nontopological structural similarities in the PDB: new examples with old structures. Proteins 1996;25:354–365.
22. Alexandrov NN, Go N. Biological meaning, statistical significance, and classification of local spatial similarities in nonhomologous proteins. Prot Sci 1994;3:866–875.
23. Mizuguchi K, Go N. Comparison of spatial arrangements of secondary structure elements in proteins. Prot Eng 1995;8:353–362.
24. Rufino SD, Blundell TL. Structure-based identification and clustering of protein families and superfamilies. J Comput Aided Mol Design 1994;8:5– 27.
25. Gibrat JF, Madej T, Bryant SH. Surprising similarities in structure comparison. Curr Opin Struct Biol 1996;6:377–385.
26. Holm L, Sander C. Touring protein fold space with Dali/FSSP. Nucleic Acids Res 1998;26:316–319.
27. Ohkawa H, Ostell J, Bryant S. MMDB: an ASN.1 specification for macromolecular structure. Ismb 1995;3:259–267.
28. 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.
29. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH: a hierarchic classification of protein domain structures. Structure 1997;5:1093–1108.
30. Holm L, Sander C. Mapping the protein universe. Science 1996;273: 595–602.

31. Bemis G, Kuntz ID. A fast and efficient method for 2D and 3D molecular shape recognition. J Comput Aided Mol Design 1992;6: 607–628.
32. Richardson, JS. The anatomy and taxonomy of protein structure. Adv Prot Chem 1981;34:167–339.
33. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.
34. Willett P. Similarity and clustering in chemical information systems. New York: Wiley, 1987.
35. Boutonnet NS, Rooman MJ, Ochagavia ME, Richelle J, Wodak SJ. Optimal protein structure alignments by multiple linkage clustering: application to distantly related proteins. Prot Eng 1995;8:647–662.
36. Ryden L. Evolution of blue copper proteins. Prog Clin Biol Res 1988;274:349–366.
37. Ryden LG, Hunt LT. Evolution of protein complexity: the blue copper-containing oxidases and related proteins. J Mol Evol 1993; 1993;36:41–66.
38. Bork P, Sander C, Valencia A. An ATPase domain common to all prokaryotic cell cycle proteins, sugar kinases, actin, and hsp70 heat shock proteins. Proc Natl Acad Sci USA 1992;89:7290–7294.
39. Bryant, SH. Evaluation of threading specificity and accuracy. Proteins 1996;26:172–185.