Analysis of Protein Sequence/Structure Similarity Relationships

Hin Hark Gan,*†‡ Rebecca A. Perlow,§ Sharmili Roy,§ Joy Ko,† Min Wu,* Jing Huang,* Shixiang Yan,* Angelo Nicoletta,* Jonathan Vafai,¶ Ding Sun,* Lihua Wang,§ Joyce E. Noah,* Samuela Pasquali,‖ and Tamar Schlick*†‡

*Department of Chemistry, †Courant Institute of Mathematical Sciences, ‡The Howard Hughes Medical Institute, §Department of Biology, ¶New York University Medical School, and ‖Department of Physics, New York University, New York, New York 10012 USA

ABSTRACT Current analyses of protein sequence/structure relationships have focused on expected similarity relationships for structurally similar proteins. To survey and explore the basis of these relationships, we present a general sequence/structure map that covers all combinations of similarity/dissimilarity relationships and provide novel energetic analyses of these relationships. To aid our analysis, we divide protein relationships into four categories: expected/unexpected similarity (**S** and **S**$^?$) and expected/unexpected dissimilarity (**D** and **D**$^?$) relationships. In the expected similarity region **S**, we show that trends in the sequence/structure relation can be derived based on the requirement of protein stability and the energetics of sequence and structural changes. Specifically, we derive a formula relating sequence and structural deviations to a parameter characterizing protein stiffness; the formula fits the data reasonably well. We suggest that the absence of data in region **S**$^?$ (high structural but low sequence similarity) is due to unfavorable energetics. In contrast to region **S**, region **D**$^?$ (high sequence but low structural similarity) is well-represented by proteins that can accommodate large structural changes. Our analyses indicate that there are several categories of similarity relationships and that protein energetics provide a basis for understanding these relationships.

INTRODUCTION

Proteins display diverse sequence/structure similarity relationships. Understanding protein similarity relationships is vital for the annotation of genome sequences (Andrade et al., 1999; Pearl et al., 2000; Wilson et al., 2000; Todd et al., 2001). Proteins with high sequence identity and high structural similarity tend to possess functional similarity and evolutionary relationships, yet examples of proteins deviating from this general relationship of sequence/structure/function homology are well-recognized. For example, high sequence identity but low structure similarity can occur due to conformational plasticity, mutations, solvent effects, and ligand binding. Despite this protein diversity most current surveys have focused on the expected similarity relationship where the proteins have significant sequence and structural similarity (Wilson et al., 2000; Chothia and Lesk, 1986; Russell et al., 1997; Levitt and Gerstein, 1998; Wood and Pearson, 1999). Furthermore, the physical basis of the expected sequence/structure similarity relationship remains unexplored. To survey and examine the basis of protein relationships, we report here a representative, broader sequence/structure map that captures known similar/dissimilar protein relationships. [This paper stemmed from an assignment given in the graduate course on Molecular Modeling (Schlick, 2002), taught by T. Schlick, at New York University. The students were challenged to find and analyze the protein pairs with the following relationships: high sequence and structural similarity; high sequence/structural similarity but markedly different biological or functional properties; low sequence similarity but high structural similarity; and high sequence similarity but low structural similarity. The assignment is available from the online textbook of the Biophysical Society (http://www.biophysics.org/) and http://monod.biomath.nyu.edu/index/course/IndexMM.html.] Based on this survey, we introduce four categories of similarity relationships. We analyze and derive the similarity relationships using energetic considerations and illustrate observed relationships with interesting examples.

To aid our analysis, we partition the map into four regions of similarity relationships in a sequence identity (I) versus root-mean-square-deviation (RMSD or R) map: expected similarity (region **S**); unexpected similarity (region **S**$^?$); expected dissimilarity (region **D**); and unexpected dissimilarity (region **D**$^?$). Our scheme distinguishes the expected (**S** and **D**) from the unexpected (**S**$^?$ and **D**$^?$) relationships for the analysis of energetic factors and characteristic structural/functional features represented in different similarity regions.

In the expected similarity region **S**, the RMSD generally rises as I decreases, reflecting increasing structural deviations as the degree of sequence similarity declines. Although several empirical formulas have been used to parametrize this trend (Chothia and Lesk, 1986; Wilson et al., 2000; Russell et al., 1997; Wood and Pearson, 1999), to the best of our knowledge, no physical derivations of the observed trend are available. Here, we formulate the trend in region **S** based on the requirement of protein stability and the energetics of sequence and structural changes. Essentially, we propose, using energetic estimates, that since

Submitted December 7, 2001, and accepted for publication June 25, 2002.

Address reprint requests to Tamar Schlick, Courant Institute of Mathematical Sciences, 251 Mercer Street, New York, NY 10012. Tel.: 212-998-3116; Fax: 212-995-4152; E-mail: schlick@nyu.edu.

Joyce E. Noah's present address is Department of Chemistry, Stanford University, Stanford, CA 94305.

© 2002 by the Biophysical Society

0006-3495/02/11/2781/11 $2.00

CRL Exhibit 1008
CRL v. Seikagaku

native proteins have evolved to approximate the free energy minimum (Kuhlman and Baker, 2000), structural changes are accompanied by sequence changes to maintain stability. Specifically, we derive an approximate sequence/structure formula, $R \sim f^{1/(2\alpha)}$, where $f = 100 - I$ and the exponent α is a measure of the stiffness of proteins. Thus, the expected sequence/structure trend depends on sequence diversity f and stiffness or flexibility of proteins. For the sparsely populated region $\mathbf{S}^?$, unfavorable energetics may explain the absence of aligned protein pairs.

The unexpected dissimilarity region $\mathbf{D}^?$ in our map contains many interesting sequence/structure/function relationships of complex multidomain proteins. These proteins can accommodate large structural changes arising from various biological and geometric factors: existence of flexibly linked regions of secondary structure facilitating large relative (rigid-body) motions of different domains; conformational changes induced by ligands; mutations in linker regions; and conformational plasticity/flexibility vital to protein function.

In the sections that follow, we define the four regions of similarity, analyze the physical basis of the observed sequence/structure patterns, and briefly survey structural and functional properties of protein pairs in region $\mathbf{D}^?$. Under Materials and Methods, we discuss our selection of probe proteins, structure alignments, and aligned protein pairs.

MATERIALS AND METHODS

Our sequence identity/RMSD map, spanning various possible combinations of protein relationships, is based on 465 probe proteins (see Table 1) representing 19 PROSITE functional (enzyme and nonenzyme) classes (Hofmann et al., 1999; http://www.expasy.ch/cgi-bin/prosite-list.pl). The PROSITE database lists groups of functionally related proteins in each functional class. To ensure a broad representation of probe proteins used, we select one protein from each entry of each PROSITE functional class where a three-dimensional structure is available. As shown in Table 1, the number of probes available is uneven across the functional classes, which is likely caused by inhomogeneity of functional family size and/or biases in structural databases. Although our probe selection procedure minimizes such problems, biases in databases are generally difficult to remove completely (Levitt and Gerstein, 1998).

Our probe proteins are tested against the protein structures in the Protein Data Bank (PDB) (http://www.rcsb.org/pdb/) to generate 53,383 aligned pairs using the FSSP method (Holm and Sander, 1996; Fold classification based on Structure/Structure alignment of Proteins). The FSSP performs an all-against-all structure comparison on a representative PDB set where no two sequences have over 30% sequence identity; the representative set of June 22, 1999 has 967 PDB chains. Because short alignments (<50 residues) are likely to be fortuitous (i.e., correspond to unrelated proteins) (Brenner et al., 1998), we select alignments with a minimum of 75 residues; longer cutoff lengths exclude significant alignments of protein fragments. For each probe protein, the number of generated alignments varies (from 10 to over 500) depending on the size of the protein family/superfamily in the database.

The FSSP algorithm generates alignments that are generally comparable to other structure alignment algorithms (Shindyalov and Bourne, 1998), such as the combinatorial extension (CE) and vector alignment search tool (VAST). Although individual hits by FSSP and CE algorithms can differ, we observe that the overall patterns generated in the sequence/structure map are similar. Differences in the internal (cutoff) parameters of alignment algorithms generate some variations in alignment hits, especially for dissimilar proteins. However, our selection of alignments greater than 75 residues helps to reduce sensitivity to the choice of cutoff parameters in alignment algorithms.

In several studies of sequence/structure plots, different probe selection procedures and numbers of aligned protein pairs have been used. Wood and Pearson's (1999) set of probes spanned 36 families; Wilson et al. (2000) used ~30,000 fold-related pairs of protein domains from the SCOP (Murzin et al., 1995) database (Structural Classification of Proteins, http://scop.mrc-lmb.cam.ac.uk/scop/); and Levitt and Gerstein (1998) used 2017 pairs of evolutionarily related proteins (i.e., belonging to the same superfamily) to test the significance of the SCOP classification. Unlike Levitt and Gerstein's work, we do not impose a restriction on the selection of aligned protein pairs, and use raw scores (R and I).

The finding of Wilson et al. (2000) that sequence identity (I) appears to be a more robust measure of functional conservation than statistical scores supports our use of traditional (I and R) rather than statistical measures of similarity. Unlike traditional measures, statistical scores measure the statistical significance of a structure and/or sequence alignment from a reference random mean (Holm and Sander, 1996). As found by Levitt and Gerstein (1998), statistical scores have a more systematic dependence on the alignment length than R.

In their structural alignments, Chothia and Lesk (1986) and Wilson et al. (2000) used a "trimming" procedure to include only residues in the conserved core; we do not impose this procedure here. The present paper also differs from Levitt and Gerstein's (1998) work for protein domains because both protein domains and multidomain proteins are used in our alignments.

RESULTS AND DISCUSSION

Four regions of sequence/structure similarity

The 53,383 aligned protein pairs in the R versus I map (Fig. 1) are concentrated in two regions: high structural similarity region ($R < 2$ Å, horizontal strip) and low sequence identity region ($I < 20\%$, vertical strip). In contrast, an uneven distribution of points is seen in the region of $R > 2$ Å and $I > 20\%$, where proteins with large conformational flexibility are found, such as proteins in the immunoglobulin superfamily (*red symbols*) and calcium-binding protein family (*green symbols*). Although protein pairs in this region have significant sequence similarity, their RMSD values can be large (>5 Å). The region of $R < 2$ Å and $I < 20\%$ is marked by the absence of aligned protein pairs, and this remains largely unchanged by increasing the number of probes used for alignment. Broadly speaking, the I/R map displays the preferences and the diversity of sequence/structure relationships in the protein structure database. As with other mapping schemes, a partitioning illustrates different types of relationships between proteins found in various functional classes or families.

We thus suggest a simple partitioning of the I/R map into the following four regions in terms of the combinations of R (Å) and I (%) values:

$$\begin{aligned}
&\text{Expected similarity, } \mathbf{S}\text{: } R \le 2 \quad \text{and} \quad I > 20,\\
&\text{Unexpected similarity, } \mathbf{S}^?\text{: } R \le 2 \quad \text{and} \quad 0 \le I \le 20,\\
&\text{Expected dissimilarity, } \mathbf{D}\text{: } R > 2 \quad \text{and} \quad 0 \le I < 20,\\
&\text{Unexpected dissimilarity, } \mathbf{D}^?\text{: } R > 2 \quad \text{and} \quad I \ge 20.
\end{aligned} \tag{1}$$

TABLE 1 List of 465 probe proteins for sequence/structure alignments

Class	PDB Codes of Probes
1.	**40 RNA- or DNA-associated proteins**
	1a04A,1a12A,1a26,1a32,1a3qA,1a5j,1ad2,1al3,1an2A,1au7A, 1bqv,1bxeA,1dipA,1div,1dpsA,1grj,1huuA,1iml,1lfb,1mh1, 1mnmA,1pkp,1qpzA,1rip,1ris,1rmd,1rss,1seiA,1sig,1smtA, 1tfb,1tum,1whi,2arcA,2hfh,2hts,2irfG,2reb,3erdA,3ullA
2.	**8 Chaperones**
	1a6dA,1ba1,1derA,1dkzA,1lepA,1qpxA,1qupA,1yer
3.	**16 Cytokines**
	1aly,1ax8,1b5l,1bfg,1bgc,1dptA,1lki,1pdgA,1rcb,1rh2F, 2gmfA,2ila,2ilk,2tgi,3inkC,1bndA
4.	**33 Domains**
	1a1z,1agrE,1aohA,1b3uA,1b4rA,1bak,1bbzA,1bd8,1bkrA,1bw4, 1cds,1ddf,1efcA,1egf,1fbr,1hfc,1klo,1kwaA,1lckA,1lvk, 1mh1,1qfhA,1rtm1,1sfp,1sra,1tf4A,1tsg,1x11A,3bct,3crd, 8kme2,9wgaA,1cll
5.	**17 Electron transport proteins**
	1a6l,1aac,1awd,1ayfA,1be3A,1bjx,1efvA,1fvkA,1gox,1hpi, 1kte,1plc,1rcf,1rcy,1rie,1ycc,5nul
6.	**83 Hydrolases**
	1a17,1a2zA,1a4mA,1a6f,1a6q,1a7tA,1a8rA,1abv,1agjA,1ah7, 1ako,1aohA,1aq0A,1auiA,1auk,1axkA,1aye,1ayx,1az9,1b3rA, 1b8jA,1b9zA,1bbzA,1be3A,1bglA,1bhe,1bmfG,1bolA,1btl,1bvqA, 1c24A,1cem,1cex,1chkA,1cnsA,1cs8A,1ctn,1ctt,1d3vA,1eceA, 1frpA,1gci,1gpl,1hfc,1hjrA,1ihp,1inp,1l92,1lam,1lpbA, 1mpgA,1pauA,1pmaA,1poa,1pscA,1qadA,1qaeA,1qsaA,1qvbA,1skyB, 1snc,1taxA,1tf4A,1tml,1tyfA,1ubpC,1uch,1ugiA,1vhrA,1whtB, 2abk,2acy,2bce,2eng,2masA,2prd,2pth,2rspB,3dni,3lzt, 4pgaA,5ptp,7rsa
7.	**8 Inhibitors**
	1ayoA,1gci,1hfc,1hssA,1ovaA,1stfI,3ssi,1wba
8.	**20 Isomerases**
	1a0cA,1a36A,1a41,1aj6,1amk,1bd0A,1bif,1deaA,1ecl,1etb1, 1fkj,1mucA,1opy,1pinA,1pmi,1reqA,1rpxA,2cpl,2sqcA,4xis
9.	**14 Ligases**
	12asA,1a0i,1a48,1a4iA,1adeA,1atiA,1bncA,1cdzA,1gln,1gsoA, 1iow,1lgr,1u9aA,2scuA
10.	**30 Lyases**
	1aj8A,1ak1,1amj,1auwA,1ax4A,1ayl,1azsA,1b4kA,1b66A,1b93A, 1burA,1cl2A,1dciA,1dhpA,1dosA,1ecmA,1fbaA,1fiy, 1jenA,1ordA, 1pda,1pii,1qcxA,1qipA,1qnf,1tdj,1uroA,2cba,2tysA,4enl
11.	**22 Other transport proteins**
	1a44,1aca,1aw0,1bcfA,1bj5,1bp1,1bxwA,1cb6A,1dpe,1hmt, 1lla,1lst,1mrp,1rzl,1sbp,1swuA,1utg,2fha,2gdm,2mhr 2omf,2vhbA
12.	**7 Other enzymes**
	1amuA,1b10A,1bdo,1bk0,1htp,1zpdA,2rslA
13.	**29 Other proteins**
	1a8y,1agdB,1ap8,1bkb,1boy,1brt,1c1lA,1cf1A,1cof,1ct5A, 1d2nA,1dar,1efcA,1hurA,1kpf,1lfdA,1maz,1nls,1ounA,1plq, 1qjaA,1sacA,1sfp,1tbgA,1thv,1tif,1vin,1wer,2pii
14.	**47 Oxidoreductases**
	1a4iA,1ak5,1an9A,1aoeA,1aozA,1b2nA,1b8bA,1b8sA,1bdb,1bmdA, 1bpwA,1cf9A,1cp2A,1dmr,1dpgA,1drw,1dssG,1dxy,1frb,1fvkA, 1gox,1han,1hyhA,1iso,1k89,1lla,1lox,1lucA,1lvl,1oacA, 1phd,1phm,1qguA,1rlr,1ryc,1soxA,1trb,1xikA,1yaiA,2aop, 2frvA,2occA,2ohxA,2pgd,3mdeA,3pchA,4pah
15.	**3 Post-transcriptional modifications**
	1am2,1amuA,2af8
16.	**7 Receptors**
	1aoxA,1c3wA,1fepA,1fmk,1vls,3ebx,6prcH
17.	**15 Structural proteins**
	1amm,1auvA,1avc,1bg2,1bkq,1cunA,1dynA,1edhA,1fsz,1neu, 1smvA,1tubA,1yagA,3nul,8fabA
18.	**11 Toxins**
	1a87,1a8d,1acc,1agjA,1ddt,1mrj,1pfo,1poa,1preA,3ebx 3seb
19.	**55 Transferases**
	1fmtA,16pk,1a49A,1a7j,1a8i,1afwA,1aj2,1ajsA,1b8oA,1bkpA, 1booA,1bpyA,1brwA,1c3oB,1cezA,1cjwA,1clqA,1daaA,1dik,1ecpA, 1fl3A,1fkj,1fmk,1ft1A,1gdoA,1glcG,1gpr,1guqA,1hka,1kwaA, 1lckA,1mxb,1nmtA,1npk,1onrA,1pgtA,1qh4A,1qk3A,1rhs,1rkd, 1sfe,1shkA,1trkA,1uby,1xwl,1zin,1zymA,2dkb,2dpmA,2ercA, 2yhx,3tdt,3tmkA,6mhtA,1bpyA

The proteins are from 19 PROSITE protein (enzyme and nonenzyme) classes (http://www.expasy.ch/cgi-bin/prosite-list.pl); "Other enzymes" refers to enzymes not in the other enzyme classes; "other proteins" includes those belonging to families that are still uncharacterized; and the "domains" class contains functional domains in proteins (e.g., EF-hand calcium-binding, actin-binding, cellulose-binding). We choose one protein from each entry in PROSITE representing a group of closely related proteins; short peptides such as hormones are excluded. Some functional classes are larger than others for biological reasons and factor-related experimental limitations; the number in each class is indicated. In PROSITE, some proteins fall into more than one functional class.

This general description broadly partitions high structural (**S**) and low sequence (**D**) similarity regions, exclusion region ($\mathbf{S}^?$), and the unexpected dissimilarity region ($\mathbf{D}^?$). The breakdown of alignments (in percentages) in the four regions is as follows: 15.75% in **S**, 0.09% in $\mathbf{S}^?$, 80.91% in **D**, and 3.25% in $\mathbf{D}^?$.

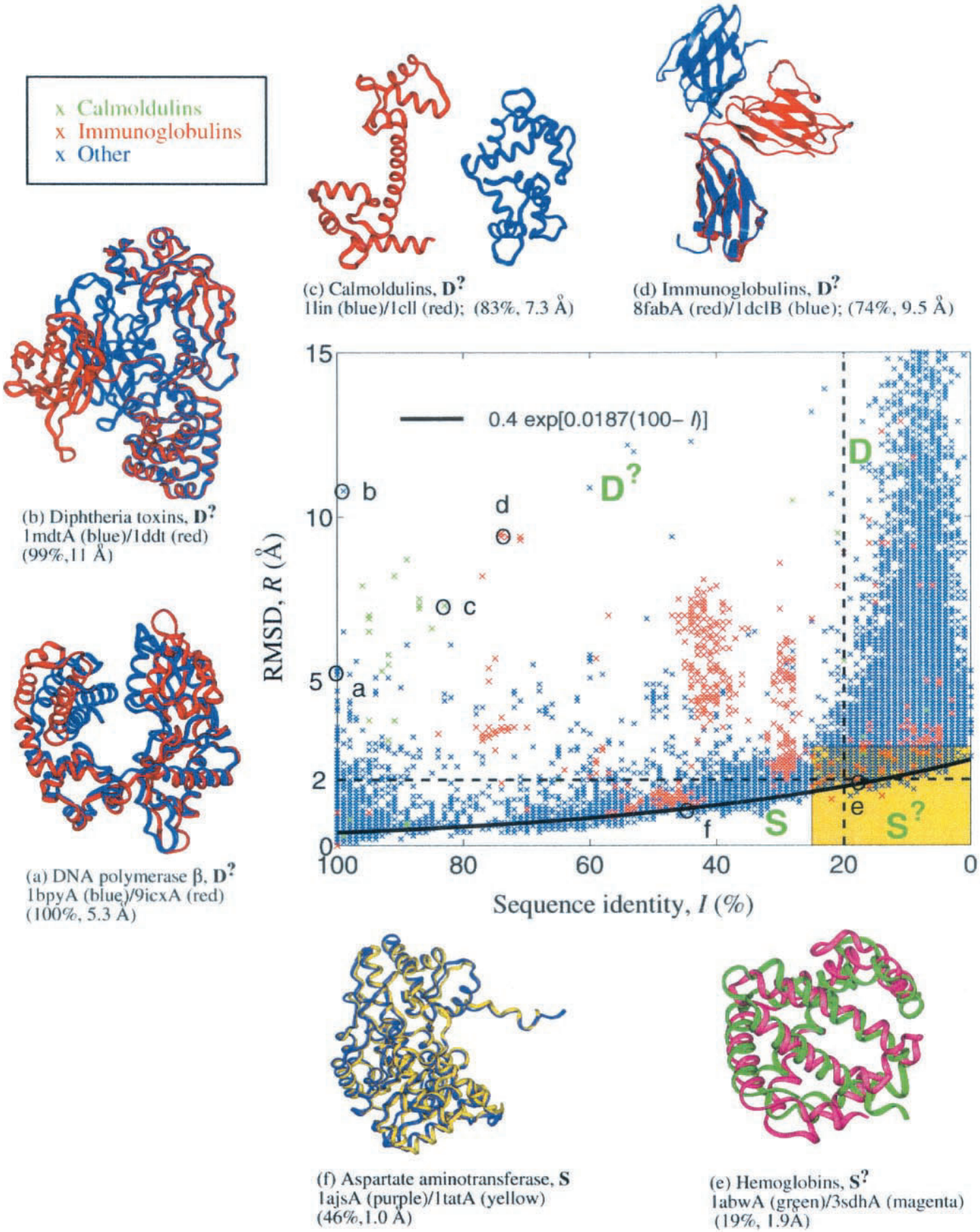

FIGURE 1 (caption on next page)

We found that the boundaries between regions remained largely unchanged when the size of the datasets increased from ~7000 to ~53,000 points. However, the boundaries at the intersection of the four regions are somewhat ill-defined because of the continuous distribution of points in this area. We define the boundary or "twilight" region as the area where $R < 3$ Å and $I < 25\%$ (highlighted in yellow in Fig. 1). Prominent examples of protein pairs in the twilight region are shown in Fig. 2. Overall, our definitions for regions **S**, **S**$^?$, **D**, and **D**$^?$ indicate a tentative but meaningful categorization of protein relationships.

Energetics of sequence/structure relations

In the following three subsections, we provide two alternate derivations (in two subsections) of the sequence/structure relation in region **S** based on protein energetics and suggest the energetic factors that likely influence the alignment results in region **S**$^?$.

Derivation of the sequence/structure relation in region S

Currently, statistical and empirical analyses are used in large-scale surveys of protein relationships (Abagyan and Batalov, 1997; Levitt and Gerstein, 1998; Wilson et al., 2000). In particular, the empirical Chothia-Lesk relation between sequence and structure states that RMSD, R, increases with decreasing sequence identity, I:

$$R = a \exp(bf), \tag{2}$$

where $f = 100 - I$, $a = 0.4$, and $b = 0.0187$ (thick line in Fig. 1). Below, we show that this sequence/structure trend (Eq. 2) can be analyzed using the energetics of sequence variation and structural perturbation, and the consideration of protein stability in the presence of sequence/structural changes. This is made possible by estimating the energetics associated with the changes, and by the use of solvation free energies in water (Chiche et al., 1990) to estimate protein stabilization energy.

To derive the sequence/structure relation, we investigate the energetics of the process by which one sequence/structure transforms into another for a pair of aligned protein structures with given R and I values. Because we only consider protein energetics, the procedure for transforming a protein into another does not necessarily follow the steps in protein evolutionary history. We propose a two-step procedure as follows: first, the probe protein is "deformed" to the structure of the target protein while keeping the sequence fixed; second, we hypothesize sequence adjustment to optimize the energy for the deformed structure. Below, we use this procedure to deduce relations among energy E, R, and I.

Imagine deforming the probe protein S_P to the target protein structure S_T. We assume that the energetic cost associated with this process, δE, increases as some power of R as follows:

$$\delta E/N_l = AR^{2\alpha}, \tag{3}$$

where R^2 is given by

$$R^2 = N_l^{-1} \sum_{i=1}^{N_l} (\mathbf{R}_i^P - \mathbf{R}_i^T)^2. \tag{4}$$

Here, N_l is the alignment length; $\mathbf{R}_i^P$ and $\mathbf{R}_i^T$ are the coordinate centers of residues for probe and target proteins, respectively; and A and α are positive constants. The exponent α determines the character of the energy deviation, and its value depends on the model chosen. Energetically stable or stiff proteins have larger α values than structurally flexible proteins. For $\alpha = 1$, δE rises quadratically with R, an approximation that may be adequate for small R values. The quadratic energy function of structural deviations has been used to interpret neutron scattering data on protein flexibility (Zaccai, 2000).

Following the structural deformation, the native sequence of the probe protein is no longer optimal from the viewpoint of energy. Thus, the probe sequence must change to minimize the associated energy. This problem is equivalent to finding the native sequence for a given structure (inverse folding). We assume that the target sequence, evolved by nature, approximates the optimal solution (Kuhlman and Baker, 2000). The sequence optimization process described above therefore lowers the energy of the target structure. If $-S(f)$ is the energy change per residue caused by the sequence change f, then the energy difference per residue between the probe and target protein becomes

$$\delta E'/N_l = AR^{2\alpha} - S(f). \tag{5}$$

For globular proteins, the stabilization energy per residue is typically 1–2 k_BT at room temperature. We thus expect the energy difference $\delta E'/N_l$ between probe and target structures to be small and independent of alignment length. This is the case for the solvation free energy per residue (Chiche

FIGURE 1 Sequence/structure map of 53,383 protein pairs displayed as a root-mean-square deviation (RMSD or R) versus percent sequence identity, I, plot. The 465 probe proteins listed in Table 1 were used to generate these pairs using the FSSP algorithm. Only alignments >75 residues are included. Sequence/structure coordinates are marked by blue × symbols except for those resulting from calmodulin (1cll, *green*) and immunoglobulin (8fabA, *red*) probes. The map is subdivided into regions as discussed in the text: expected similarity (**S**); unexpected similarity (**S**$^?$); expected dissimilarity (**D**); and unexpected dissimilarity (**D**$^?$). Superpositions of selected protein pairs in different regions are marked and displayed. The thick black line corresponds to the empirical exponential function of Eq. 2.

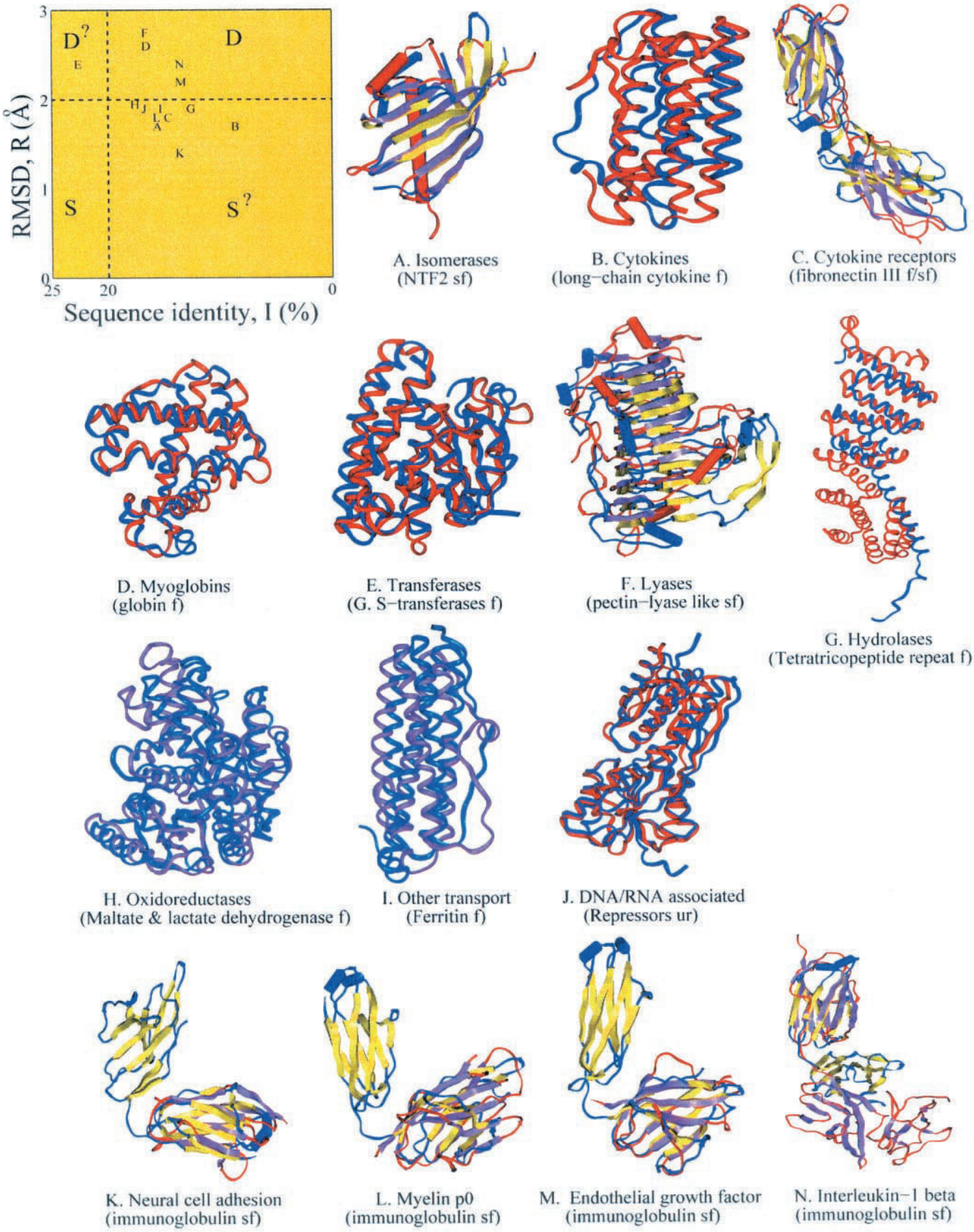

FIGURE 2 (caption on next page)

et al., 1990): $F_{sol}/N = 15.30/N - 1.13$ (kcal/mol). Thus, assuming that $\delta E'/N_1 \approx 0$, we have

$$R = [S(f)/A]^{1/(2\alpha)}. \quad (6)$$

This relation for native proteins states that structural change R depends on sequence change f and protein stiffness parameter α. Next, we estimate the energy cost, $S(f)$, associated with sequence change.

The function $S(f) \geq 0$ because it is a measure of the energy cost associated with sequence variation from the native sequence ($f = 0$) of the probe protein; for the native sequence, $S(0) = 0$. In the simplest model, we may assume that $S(f)$ increases linearly with f, thereby yielding

$$R = (f/A')^{1/(2\alpha)} \quad (7)$$

where A' is a constant. For $\alpha = 1$, R is a slowly increasing function of f, as seen in region **S** of Fig. 1. In contrast, the exponential fitting function (2) is nonzero when $f = 0$ because native proteins have some degree of conformational flexibility. Because conformational flexibility is not accounted for in our simple model, we predict that $R = 0$ when $f = 0$.

The exponent α is a measure of protein stiffness. It may be derived from an analysis of external perturbations on molecular interactions, which is beyond the scope of the present work. Experimental analysis of protein conformational flexibility shows that both harmonic ($\alpha = 1$) and anharmonic effects are present (Zaccai, 2000). We determine the acceptable range of α values by fitting our derived formula (7) to protein alignment data. Fig. 3 compares formula (7) for different α values with aligned protein data and empirical formula (2). The curves for $\alpha = 0.75$, 1, 1.5 fit the data well for $I < 40\%$, but they deviate from the data at lower I values. The lower I value region tends to be better described by smaller α values between 0.5 and 1. These results indicate that protein pairs in region $I > 40\%$ are "stiffer" than those in $I < 40\%$ because they are characterized by larger α values. Proteins in the $I < 40\%$ region are relatively "flexible" as measured by their ability to accommodate significant sequence and structural deviations.

Above, we formulated the consequences of perturbing the protein structure followed by sequence optimization or design. Alternatively, we can reverse the process with sequence variation followed by structural perturbation. In this process, the sequence change (f) for a probe protein S_P from its native sequence to the native sequence of the target protein S_T will increase the free energy by $S'(f)$. To recover the optimal energy, the structure of the probe protein is perturbed to yield the structure of the target protein S_T. This structural change is expected to decrease the energy by $R^{2\alpha}$. Although the absolute values of the energetic terms in the reverse process are not identical with the process described before, the functional forms of the terms should remain the same, hence Eq. 7 holds. The reverse process appears to follow protein evolutionary changes where sequence variation (e.g., mutations) lead to structural changes.

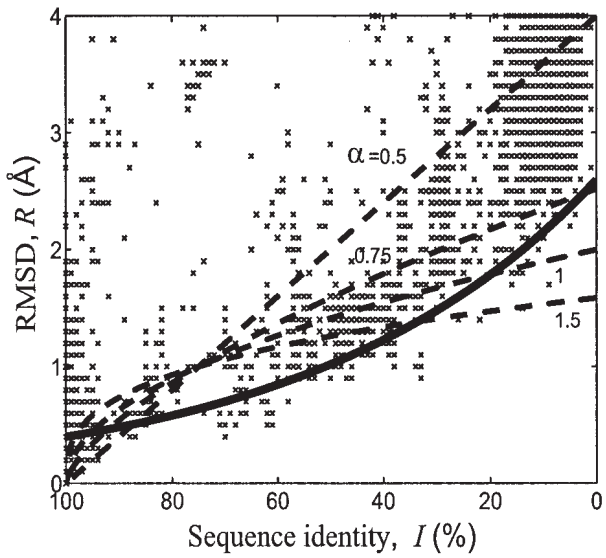

FIGURE 3 Sequence/structure relations from derived formula (7) (*dashed curves*) and empirical formula (Eq. 2; *thick line*) are compared with protein alignment data. Formula (7) is evaluated at different values of α, or the protein stiffness parameter; we used a normalization factor $A' = 25$.

Protein energy function and sequence/structure relations

A more general consideration of protein energy functions and their relation to sequence/structure relations can be

FIGURE 2 Superimposed protein pairs in the twilight zone covering **S**, **S**$^?$, **D**, and **D**$^?$ subregions called the boundary region. Marked protein pairs correspond to: (*A*) 1opy (*blue, yellow*)/1aounA (*red, magenta*) with (I, R) = (9%, 1.7 Å) from the (isomerase) nuclear transport factor 2 (NTF2) superfamily; (*B*) 1ax8 (*blue*)/1bgc (*red*) with (14%, 1.8 Å) from the long-chain cytokine family; (*C*) 1boy (*blue*, *yellow*)/1qg3A (*red, magenta*) with (16%, 1.7 Å) from the fibronectin type III family/superfamily; (*D*) myoglobins 104m (*blue*)/2gdm (*red*) with (17%, 2.6 Å); (*E*) transferases 1pgtA (*blue*)/1a0fA (*red*) with (17%, 2.7 Å); (*F*) lyases 1qcxA (*blue, yellow*)/1air (*red, magenta*) with (23%, 2.5 Å); (*G*) hydrolases 1a17 (*blue*)/1qqeA (*red*) with (13%, 1.9 Å); (*H*) oxidoreductases 1ceqA (*blue*)/1bmdA (*magenta*) with (18%, 2 Å); (*I*) other transport proteins 1bcfA (*blue*)/1qghA (*magenta*) with (16%, 1.9 Å); (*J*) DNA/RNA associated proteins 1qpzA (*blue*)/1bykA (*red*) with (17%, 1.9 Å); (*K*) immunoglobulin 8fabA (*yellow*)/2ncm (neural cell adhesion, *magenta*) with (16%, 1.8 Å); (*L*) immunoglobulin 8fabA (*yellow*)/1neu (myelin p0, *magenta*) with (14%, 1.4 Å); (*M*) immunoglobulin 8fabA (*yellow*)/1fltX (endothelial growth factor, *magenta*) with (14%, 2.3 Å); and (*N*) immunoglobulin 8fabA (*yellow*)/1itbB (interleukin-1 β, *magenta*) with (14%, 2.2 Å). Pairs *K* and *N* are examples of structural cousins of immunoglobulins. Superfamily, family, and unrelated proteins according to SCOP are abbreviated as sf, f, and ur, respectively.

sought by rationalizing the empirical formula (2) and the approximate relation (6) using proposed energy functions.

Let E_0 be the free energy per residue of a protein in the native state. Sequence and structural changes brought about by mutations and external forces, for example, will perturb the energy function. We assume that the perturbed free energy has the form

$$E = E_0[1 + c_1 J_{seq}(f) + c_2 J_{str}(R)] \tag{8}$$

where f and R are as defined before. The functions $J_{seq}(f)$ and $J_{str}(R)$ describe deviations from the native energy due to sequence and structural changes, respectively, with corresponding coefficients c_1 and c_2, which are related to the response of the energy function to the changes. Although the coefficients c_1 and c_2 may be determined experimentally or theoretically, their precise values are not needed in our derivations below. Because the native state is stable with respect to small sequence/structure perturbations, we expect the total free energy to be higher for nonnative states.

We now consider a family of sequence and structural changes that maintains the native energy per residue, i.e., $E = E_0$. From Eq. 8, we have the result

$$c_1 J_{seq}(f) + c_2 J_{str}(R) = 0. \tag{9}$$

Because the sequence/structure changes maintain constant native energy per residue, Eq. 9 expresses the general sequence/structure relation between native proteins.

To derive Eq. 6 and empirical formula (2), we must propose suitable forms for the functions J_{seq} and J_{str}. We assume that the J_{seq} function is a power series in f, $J_{seq}(f) = \Sigma_n a_n f^n$, where a_n are constant coefficients. In the first-order, linear approximation, we have

$$J_{seq} \sim f. \tag{10}$$

It is clear that the sequence/structure relation (6) derived earlier and the empirical formula are obtained from the following J_{str} functions, respectively,

$$J_{str} \sim (R/R_0)^{2\alpha}, \tag{11}$$

$$J_{str} \sim \ln(R/a) \quad \text{for} \quad R \geq a \tag{12}$$

where R_0 is a constant and $a = 0.4$ Å is the smallest RMSD allowed in empirical formula 2. There is a reasonable agreement between the two J_{str} functions for $R_0 = 1$ Å and $\alpha = 1$ in the range of $R \leq 1$ Å. Future improved analysis of energetic estimates for both J_{seq} and J_{str} functions should avoid large discrepancies between the empirical formula and the expression derived using physical analysis.

Energetic considerations for region S?

Protein pairs in region $\mathbf{S}^?$ are highly unlikely because sequence/structure combinations must preserve stable conformational states. We also call region $\mathbf{S}^?$ the "exclusion region." We argue that maintaining the native free energy (F) may be inconsistent with the sequence/structure constraint parameters of the exclusion region. The existence of aligned protein pairs in this region would suggest that R is independent of sequence diversity f. From our analysis of sequence and structural changes, this is only possible when $S(f)$ is a constant, i.e., there is no variation in energetic cost associated with sequence changes away from the native sequence. Such a requirement is highly unlikely to be fulfilled. In theory, our argument may be tested by using a threading technique, assuming that the two proteins being compared have very similar structures (small R). All protein sequences with a given (low) sequence identity could then be threaded through the probe template structure, and the corresponding energies would be calculated to determine the proportion of sequences that leads to favorable or unfavorable energetics. In practice, because the sequence space is very large, only a partial sampling of sequences is possible. Our arguments suggest that native proteins only populate certain regions in (I, R, F) space.

Structural and functional characteristics of proteins in region D?

In contrast to the recent studies that only considered protein domains (Wilson et al., 2000; Levitt and Gerstein, 1998), our alignments of domains and whole proteins highlight unexpected sequence/structure relationships found in region $\mathbf{D}^?$ that arise from proteins whose domains can adopt multiple conformations. We identify some of these protein families to better understand the diversity of sequence/structure/function relationships exhibited by complex multidomain proteins. Important examples of protein pairs in region $\mathbf{D}^?$, shown in Fig. 1, include (*a*) human DNA polymerase β, (*b*) diphtheria toxins, (*c*) calmodulins (*green symbols*), and (*d*) immunoglobulins (*red symbols*). The sequence identities for both the DNA polymerase β and diphtheria toxin pairs are close to 100%, yet their R values are larger than 5 Å. Calmodulins (calcium-binding proteins) also have high sequence similarity (mostly $I > 80\%$), but the immunoglobulin-related pairs have lower sequence similarity (mostly $I < 60\%$). Other noteworthy examples are homeodomain proteins, lactoferrins, SH3-domain proteins, and translation factors.

The biological and physical causes for large structural deviations are interactions with drugs (calmodulins), binding to iron (lactoferrins), binding to GTP (translation factors), binding to DNA (homeodomain proteins), specific interactions between domains (FAD/NAD(P) binding and SH3-domain), and open/close movements between domains inherent to biological function (DNA polymerase β and diphtheria toxins). In most cases, large conformational changes are inherent to the biological functions of the

TABLE 2 Summary of the relation between functional and structural properties for aligned proteins in region $D^?$ with RMSD values of >5 Å

Example	Protein	Function	Reasons for large RMSD
1.	DNA polymerase (1bpyA/9icxA), Fig. 1 *a*	DNA repair	Open/close movements of domains, inherent to function
2.	Toxin (1mdtA/1ddt), Fig. 1 *b*	Intoxication	Open/close movements of domains, inherent to function
3.	Calmodulin (1lin/1cll), Fig. 1 *c*	Calcium binding	Interaction with drugs, flexible helix linker
4.	FAD/NAD(P)-binding (1trb/1f6mE)	Reduction of disulfides	Interactions between domains
5.	Homeodomain (1au7A/1octC)	DNA regulators	Domains bind different parts of DNA
6.	Immunoglobulin (8fabA/1dclB), Fig. 1 *d*	Immune response	Flexible linker between domains
7.	Lactoferrin (1cb6A/1ce2A)	Iron-binding	Open/close movements of domains, iron binding; inherent to function
8.	SH3-domain (1lckA/1fmk)	Signal transduction	Interactions between SH2/SH3 domains
9.	Translation factors (1efcA/1tttA)	Elongation of protein synthesis	Interactions between domains, GTP binding

A representative protein pair (PDB codes) for each functional group is provided. See illustrations in Fig. 1 for structures of selected pairs. The Protein Motions Database (www.ag.uiuc.edu/~fs401/Protein_Motions_Database.html) classifies examples 1–3, 6, 7, and 9 as having hinge motions; motions of the other examples are not classified.

proteins. For diphtheria toxins and immunoglobulins, the existence of flexibly linked regions facilitate large relative (rigid-body) motions of different domains involving rotations of several angles of the linker residues (see below). In the Protein Motions Database (http://www.ag.uiuc.edu/~fs401/Protein_Motions_Database.html), motions between domains are classified as either hinge or shear motions; some domain motions are more difficult to characterize. In Table 2, we summarize examples that illustrate the relationship between functional and structural properties for aligned proteins in region $\mathbf{D}^?$ with RMSD values of >5 Å. Below, we briefly discuss the relationship between function and the sequence/structure alignment for DNA polymerase β, diphtheria toxin, calmodulin, and immunoglobulin.

Biological functions involving large conformational changes

DNA polymerase β. Polymerases catalyze many important reactions essential to life, including DNA repair. Crystal structures of the human DNA polymerase β complexed with DNA, for example, show that the protein exists in closed and open conformations (Pelletier et al., 1996; Sawaya et al., 1997) and possibly in a stable intermediate structure. Mechanisms for catalytic polymerization require rotation (of α-helix N) of the "thumb" subdomain of the protein into the closed conformation. The nicked DNA product then dissociates upon the return of the thumb to the open conformation. Fig. 1 *a* displays the open and closed conformations of human DNA polymerase β with superimposed structures having $R = 5.3$ Å.

Diphtheria toxin. Similarly, the 535-residue diphtheria toxin exists in closed and open forms (Fig. 1 *b*). In this case, conformational changes of eight linker residues in the main chain act as a hinge to open and close the receptor-binding domain of the toxin with respect to the transmembrane domain (Bennett et al., 1994), resulting in $R > 10$ Å (Fig. 1 *b*). Bennett and Eisenberg (Bennett et al., 1994; Bennett and Eisenberg, 1994) proposed that this large-scale motion is an essential part of the intoxication mechanism.

Calmodulin. Most calmodulin pairs in Fig. 1 (*green pairs*) have $I > 80\%$ but $4 < R < 8$ Å. The diversity of calmodulin folds arises from mutations and binding to drug molecules and peptides. Calmodulins are made of two calcium-binding domains (two EF hands in each domain) joined by a flexible linker helix; they represent a major calcium-dependent regulator of important intracellular processes in eukaryotes. Mutations in the linker helix sequence can distort the linker geometry (Persechini et al., 1991) and overall fold (Mirzoeva et al., 1999), changing the relative orientation and/or distance between the two calcium-binding domains dramatically. Inactivation of calmodulin can also result from binding to the drug trifluoperazine, which causes it to fold into a compact, globular protein (1lin) (Vandonselaar et al., 1994).

Immunoglobulins. The scattered (*red*) points generated by human immunoglobulin 8fabA covering region $\mathbf{D}^?$ and other regions are indicative of the large conformational flexibility of the immunoglobulin superfamily. Immunoglobulin domains are joined by a linker segment. The

rotations of a few linker residues can account for the large relative, rigid-body rotations between immunoglobulin domains (Fig. 1 *d*). Members of this large family include receptors from the immune, hemopoietic, and nervous systems, and other membrane proteins. They perform closely related functions of binding either antigens or molecules to initiate signal processing. In Fig. 2, *K–N*, the structures of immunoglobulin-like proteins (neural cell adhesion, myelin p0, endothelial growth factor, and interleukin-1 β) are superimposed with the probe 8fabA.

SUMMARY AND CONCLUSIONS

Our broad survey of protein sequence/structure relationships suggests four regions of similarity relationships: expected similarity, unexpected similarity, expected dissimilarity, and unexpected dissimilarity. This general scheme offers a tentative but instructive and useful categorization of known protein relationships exhibited by various protein functional classes. Previous surveys and studies have only focused on mapping expected similarity region **S** for protein domains, where domain pairs are structurally and functionally related (Wilson et al., 2000; Chothia and Lesk, 1986; Russell et al., 1997; Wood and Pearson, 1999; Sauder et al., 2000).

We provide energetic analyses of the observed patterns of relationships in expected similarity (**S**) and unexpected similarity ($\mathbf{S}^?$) regions. Our analyses show that sequence and structural changes during protein evolution are modulated by the necessity to maintain stability or minimize free energy. This has been demonstrated by establishing the connections among structural similarity, sequence identity, and free energy using energetic estimates of sequence and structural changes in proteins. Our sequence/structure formula is also characterized by the protein stiffness parameter α. Given the available protein alignment data, the formula agrees reasonable well with the data. More precise energetic analyses are needed in future studies.

Our survey also emphasizes that unexpected sequence/structure relationships in region $\mathbf{D}^?$ are not uncommon. We briefly illustrate and describe important protein pairs in this region that exhibit large structural deviations despite high sequence similarity. These complex multidomain proteins exhibit conformational plasticity inherent to biological activity, critical mutations (in linker/loop regions, for example), structural changes induced by ligands, and diversity of conformational requirements for functional activities.

We thank Drs. Andrej Šali and Fred E. Cohen for insightful comments on the manuscript, and Dr. Linjing Yang for helpful discussions. J. E. Noah (now a graduate student at Stanford) and J. Vafai (now an NYU medical student) were undergraduate students at the time this work was performed. We are grateful to the SCOP authors for sending us their files for our analysis.

This work was supported in part by National Institutes of Health Grant GM55164, National Science Foundation Grants BIR-94-23827EQ and ASC-9704681, a John Simon Guggenheim fellowship (to T. Schlick), and by a New York University Department of Chemistry Summer Fellowship (to J. E. Noah).

REFERENCES

Abagyan, R. A., and S. Batalov. 1997. Do aligned sequences share the same fold? *J. Mol. Biol.* 273:355–368.

Andrade, M. A., N. P. Brown, C. Leroy, S. Hoersch, A. de Daruvar, C. Reich, A. Franchini, J. Tamames, A. Valencia, C. Ouzounis, and C. Sander. 1999. Automated genome sequence analysis and annotation. *Bioinformatics.* 15:391–412.

Bennett, M. J., S. Choe, and D. Eisenberg. 1994. Refined structure of dimeric diphtheria toxin at 2.0 Å resolution. *Protein Sci.* 9:1444–1463.

Bennett, M. J., and D. Eisenberg. 1994. Refined structure of monomeric diphtheria toxin at 2.3 Å resolution. *Protein Sci.* 9:1464–1475.

Brenner, S. E., C. Chothia, and T. J. P. Hubbard. 1998. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc. Natl. Acad. Sci. U.S.A.* 95:6073–6078.

Chiche, L., L. Gregoret, F. E. Cohen, and P. Kollman. 1990. Protein model structure evaluation using the solvation free energy of folding. *Proc. Natl. Acad. Sci. U.S.A.* 87:3240–3243.

Chothia, C., and L. M. Lesk. 1986. The relation between the divergence of sequence and structure in proteins. *EMBO J.* 5:823–826.

Hofmann, K., P. Bucher, L. Falquet, and A. Bairoch. 1999. The PROSITE database, its status in 1999. *Nucleic Acids Res.* 27:215–219.

Holm, L., and C. Sander. 1996. Mapping the protein universe. *Science.* 273:595–603.

Kuhlman, B., and D. Baker. 2000. Native protein sequences are close to optimal for their structures. Proc. Natl. Acad. Sci. U.S.A. 97: 10383–10388.

Levitt, M., and M. Gerstein. 1998. A unified statistical framework for sequence comparison and structure comparison. *Proc. Natl. Acad. Sci. U.S.A.* 95:5913–5920.

Mirzoeva, S., S. Weigand, T. J. Lukas, L. Shuvalova, W. F. Anderson, and D. M. Watterson. 1999. Analysis of the functional coupling between calmodulin's calcium binding and peptide recognition properties. *Biochemistry.* 38:3936–3947.

Murzin, A. G., S. E. Brenner, T. Hubbard, and C. Chothia. 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* 247:536–540.

Pearl, F., A. E. Todd, J. E. Bray, A. C. Martin, A. A. Salamov, M. Suwa, M. B. Swindells, J. M. Thornton, and C. A. Orengo. 2000. Using the CATH domain database to assign structures and functions to the genome sequences. *Biochem Soc. Trans.* 28:269–275.

Pelletier, H., M. R. Sawaya, W. Wolfle, S. H. Wilson, and J. Kraut. 1996. Crystal structures of human DNA polymerase β complexed with DNA: implications for catalytic mechanism, processivity, and fidelity. *Biochemistry.* 35:12742–12761.

Persechini, A., R. H. Kretsinger, and T. N. Davis. 1991. Calmodulins with deletions in the central helix functionally replace the native protein in yeast cells. *Proc. Natl. Acad. Sci. U.S.A.* 88:449–452.

Russell, R. B., M. A. S. Saqi, R. A. Sayle, P. A. Bates, and M. J. E. Sternberg. 1997. Recognition of analogous and homologous protein folds: analysis of sequence and structure conservation. *J. Mol. Biol.* 269:423–439.

Sauder, J. M., J. W. Arthur, and R. L. Dunbrack, Jr. 2000. Large-scale comparison of protein sequence alignment algorithms with structure alignments. *Proteins: Struct., Funct., Genet.* 40:6–22.

Sawaya, M. R., R. Prasad, S. H. Wilson, J. Kraut, and H. Pelletier. 1997. Crystal structures of human DNA polymerase β complexed with gapped

and nicked DNA: evidence for induced fit mechanism. *Biochemistry.* 36:11205–11215.

Schlick, T. 2002. Molecular Modeling and Simulation: An Interdisciplinary Guide. Springer-Verlag, New York.

Shindyalov, I. N., and P. E. Bourne. 1998. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.* 11:739–747.

Todd, A. E., C. A. Orengo, and J. M. Thornton. 2001. Evolution of function in protein superfamilies, from a structural perspective. *J. Mol. Biol.* 307:1113–1143.

Vandonselaar, M., R. A. Hickie, J. M. Quail, and L. T. Delbaere. 1994. Trifluoperazine-induced conformational change in Ca(2+)-calmodulin. *Nat. Struct. Biol.* 1:795–801.

Wilson, C. A., J. Kreychman, and M. Gerstein. 2000. Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *J. Mol. Biol.* 297:233–249.

Wood, T. C., and W. R. Pearson. 1999. Evolution of protein sequences and structures. *J. Mol. Biol.* 291:977–995.

Zaccai, G. 2000. How soft is a protein? A protein dynamics force constant measured by neutron scattering. *Science.* 288:1604–1607.