Article No. jmbi.1999.2911 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (1999) 00, 1–19

JMB

Universally Conserved Positions in Protein Folds: Reading Evolutionary Signals about Stability, Folding Kinetics and Function

Leonid A. Mirny and Eugene I. Shakhnovich*

Department of Chemistry and Chemical Biology, Harvard University, 12 Oxford Street Cambridge, MA 02138, USA

**Corresponding author*

Here, we provide an analysis of molecular evolution of five of the most populated protein folds: immunoglobulin fold, oligonucleotide-binding fold, Rossman fold, alpha/beta plait, and TIM barrels. In order to distinguish between ''historic'', functional and structural reasons for amino acid conservations, we consider proteins that acquire the same fold and have no evident sequence homology. For each fold we identify positions that are conserved within each individual family and coincide when non-homologous proteins are structurally superimposed. As a baseline for statistical assessment we use the conservatism expected based on the solvent accessibility. The analysis is based on a new concept of ''conservatism-of-conservatism''. This approach allows us to identify the structural features that are stabilized in all proteins having a given fold, despite the fact that actual interactions that provide such stabilization may vary from protein to protein. Comparison with experimental data on thermodynamics, folding kinetics and function of the proteins reveals that such universally conserved clusters correspond to either: (i) super-sites (common location of active site in proteins having common tertiary structures but not function) or (ii) folding nuclei whose stability is an important determinant of folding rate, or both (in the case of Rossman fold). The analysis also helps to clarify the relation between folding and function that is apparent for some folds.

Keywords: protein evolution; protein folding; conservatism; super-site kinetics; protein stability

Introduction

The amount of data on protein structure, folding and kinetics are exploding. Progress in genomics (gene sequences) and proteomics (structure, function and expression) studies created a new realm for bioinformatics in which a qualitatively different amount of biological information needs to be properly rationalized and used. Success in achieving this goal depends entirely on our understanding of the principles that govern protein stability, folding and function.

Such understanding progressed over last several years to the point that basic principles of folding begin to emerge from theoretical and experimental studies. Of particular importance is the foldability principle in thermodynamics and the discovery of nucleation in folding kinetics. The foldability principle states that protein-like sequences should have their native conformation as pronounced energy minimum (separated by a large energy gap from the bulk of structurally unrelated misfolded conformations) (Goldstein *et al*., 1992; Shakhnovich & Gutin, 1993a; Sali *et al*., 1994; Govindarajan & Goldstein, 1995; Hao & Scheraga, 1994b). Sequences that satisfy this requirement are able to fold fast and have cooperative folding transition (Shakhnovich & Gutin, 1993a; Shakhnovich, 1994; Hao & Scheraga, 1994a,b). Their native structures are stable against mutations (Tiana *et al*., 1998) as well as against variation in solvent conditions and temperature (Pande *et al*., 1995).

The modern concept of nucleation in protein folding emerged from several theoretical and experimental studies (Bryngelson & Wolynes, 1990;

Abbreviations used: CoC, conservatism-of-conservatism; PDB, Protein Data Bank; OB, oligonucleotide-binding; EBN, endo-beta-*N*-acetylglucosaminidase; CLT, central limit theorem.

E-mail address of the corresponding author: eugene@belok.harvard.edu

0022-2836/99/0000001–19 $30.00/0

Abkevich *et al*., 1994; Shakhnovich *et al*., 1996; Mirny *et al*., 1998; Guo & Thirumalai, 1995; Pande *et al*., 1998; Itzhaki *et al*., 1995; Martinez *et al*., 1998) as a paradigm to describe transition state ensemble of protein folding, especially for proteins that fold *via* simple two-state kinetics (Jackson, 1998). Of particular importance is the discovery of specific folding nucleus in some proteins. The specific nucleus scenario of folding suggests that a number of obligatory contacts (specific nucleus) should be formed in order for a protein chain to reach the transition state. The specific nucleus constitutes a spatially contiguous cluster in structure, but not necessarily in sequence: non-local contacts are always present in specific nuclei. After the specific nucleus is formed, subsequent transition occurs downhill in free energy and is fast (Abkevich *et al*., 1994; Shakhnovich, 1998a). Further, it was noted (Abkevich *et al*., 1994; Shakhnovich *et al*., 1996; Mirny *et al*., 1998; Shakhnovich, 1998a; Martinez *et al*., 1998) that location of a specific nucleus depends on the structure to a greater extent than it does on sequence. The major implication of this finding is prediction that different (even non-homologous) sequences that fold into the same structure may have similar folding nuclei. In other words, the location of a folding nucleus in a structure may serve as a fingerprint of a protein fold. The specific nucleus model of folding kinetics has direct implication for experimental results, predicting a substantial variance of kinetic effects of mutations at various locations in protein structure. Another important prediction is robustness of specific nucleus with respect to variation in solvent conditions, temperature and other mutations. These predictions are consistent with experiment (Itzhaki *et al*., 1995; Viguera *et al*., 1997).

Molecular evolution represents an invaluable natural laboratory to test and further develop our understanding of protein folding. Conversely, our understanding of protein folding and function is a key to rational analysis of signals sent by protein evolution. The fusion of theoretical understanding of protein folding with analysis of evolutionary information is the main aim of this study.

Molecular evolution sends us signals in the form of conservation patterns in multiple sequence alignments. However, those signals are hard to decipher because there may be many reasons for conservation: function, stability or maybe ''historical'' reasons (insufficient evolutionary time to diverge). Finally, there may be some evolutionary pressure towards fast folding (perhaps to exceed some rate threshold beyond which aggregation and/or proteolysis of party folded species may present a problem). The kinetic factor may give rise to additional conservation in the kinetically important locations related to folding nucleus.

How can one distinguish between different reasons for amino acid conservation? A possible approach is to use as much evolutionary information as possible. In particular, it is known that besides protein homologs, i.e. proteins that have a clear evolutionary connection and are often (but not always) functionally related, there exist analogus, i.e. structurally similar proteins that have non-homologous sequences, unrelated functions and no evident evolutionary relation (Branden & Tooze, 1998). Since in most cases analogs share a common fold but not function, a proper sequence comparison between them may emphasize positions where conservatism is related to structural stability and folding kinetics rather than function (except in the cases when folds contain functional super-sites (Russell *et al*., 1998a), see below).

However, comparison of sequences of protein analogs should be made with care: a simple sequence alignment between analogs may not always work due to the possibility of multi-amino acid correlated mutations. The easiest way to understand this is to consider a basic example where a certain element of structure needs to be stabilized. However, there are several ways to form strong attractive interactions (i.e. by forming hydrophobic contacts or disulfide bridges or in some cases salt bridges). Therefore, if the same element of structure is stabilized in analogs by different forces, the amino acid residues that deliver such stabilization may be of quite different types. This suggests that a simple sequence alignment between analogs may in some cases yield no indication of conservatism. In other words, energetics may be more conserved than amino acid types that deliver it. On the other hand, within families of homologous proteins one can expect conserved amino acid residues to form stable substructures: the change of amino acid residues in these positions requires compensating mutations in several related positions. Such multi-amino acid correlated mutations are very rare. They can be found only in highly diverged or unrelated proteins rather than within protein families.

This analysis suggests that a factor that may point to a common structure-related property in all analogs may be the intrafamily conservation itself rather than actual amino acid residues at the positions in question. This leads to an important new concept of ''conservatism-of-conservatism'' (CoC) to analyze evolutionary signals that are specific to a given fold (Mirny *et al*., 1998). The principle of CoC calls for alignment of intrafamily conservatism profiles between analogs as a method to find and analyze evolutionary signals that reflect features that are characteristic of a particular fold: structural stability, folding kinetics or in some cases common function or common location of active sites between analogs (super-site).

In the following report, we first explain how CoC is computed and what controls and statistical tests we perform. Next, we consider the case of the immunoglobulin fold in detail, and show how CoC analysis helps to identify the evolutionary pressure towards fast folding and distinguish it from evolutionary pressure aimed at protein stabilization and functional pressure. This will help to identify a possible location of folding nucleus for the immu-

noglobulin fold which allows direct comparison with protein engineering experiments.

Next we carry out similar analysis for all other folds for which sufficient structural and evolutionary data are available oligonucleotide-binding (OB) fold, Rossman fold, alpha/beta plait and TIM barrel). Similar to the case of immunoglobulin fold, the analysis of the observed CoC signal will allow us to identify (in some cases) common nucleation sites characteristic of a given fold and (in some cases) super-sites, i.e. a common location of the active site in proteins with similar structure but possibly different function.

The results of our analysis will be compared with experimental information about the function, thermodynamics and kinetics of studied proteins in cases when such information is available.

Results

Conservatism-of-conservatism (CoC)

As was stated earlier, the analysis of CoC aims to identify positions in a protein structure which are conserved within each family of homologous proteins that acquire this structure. To pursue this goal we need: (i) a large set of analogs - non-homologous proteins sharing the same fold (representative proteins); and (ii) for each representative protein a number of proteins homologous to it (a family).

When these data are available, the evaluation of CoC proceeds as follows: (i) make multiple sequence alignments of proteins homologous to each representative protein; (ii) identify positions which are conserved within each multiple alignment; (iii) structurally align families to each other; and (iv) identify sites where conserved positions coincide between the families.

Figure 1 outlines the major steps of this procedure.

We use the FSSP database (Holm & Sander, 1993) as a source of structural alignments of representative proteins and the HSSP database (Dodge *et al.*, 1998) as a source of sequence alignments among homologous proteins. Some FSSP structural alignments were corrected using our Monte Carlo structural alignment algorithm (see Methods and Mirny & Shakhnovich, 1998).

The degree of evolutionary conservation within a family of homologous sequences is measured by sequence entropy:

$$s(l) = -\sum_{i=1}^{6} p_i(l) \log p_i(l)$$

where $p_i(l)$ is the frequency of each of the six classes i of residues at position l in the multiple sequence alignment. The six classes of residues are: aliphatic {*AVLIMC*}, aromatic {*FWYH*}, polar {*STNQ*}, positive {*KR*}, negative {*DE*}, and special (reflecting their special conformational properties) {*GP*}. A low value of the intrafamily conservatism $s(l)$ indicates that this position was under an evolutionary pressure to keep a particular type of residue.

After representative proteins and their respective families are structurally superimposed we compute conservatism-of-conservatism (CoC):

$$S(l) = \sum_{m=1}^{M} s^m(l)/M \qquad (1)$$

where l is now position in the structural alignment, and $s^m(l)$ is intrafamily conservatism in family m. A low value of $S(l)$ indicates that position l was conserved in most of the protein families acquiring this fold. Note that identities of these residues could be different in different families, what really matters is their conservatism within each family (see Figure 2).

CoC *versus* solvent accessibility: a stability factor

As with any observed quantity, the statistical significance of the obtained $S(l)$ value has to be evaluated. Residues that are less exposed to solvent are known to be more evolutionary conserved (Koshi & Goldstein, 1997; Branden & Tooze, 1998). The main reason for that appears to be due to the selection of thermodynamically stable sequences (see Discussion for more details). ''Buriedness'' in this context is an indicator of the degree to which an amino acid participated in intraprotein interactions. Obviously, amino acids that are more involved in such interactions are more important for stability and should be conserved under pressure towards some (not necessarily highest possible) thermodynamic stability.

Hence, when structures of the same fold are superimposed, buried residues of one protein match buried residues of the other. These buried residues tend to be more conserved in each family, giving rise to low values of $S(l)$ for buried positions, i.e. a straightforward apparent CoC signal. Thus an important control has to be made: Can higher conservatism of buried positions explain the observed values of $S(l)$?

In order to address this question we formulate the following statistical hypothesis: H0: Sequence entropy $s^m(l)$ at a position l in a family m depends solely on the solvent accessibility a_l of this position, and observed CoC $S(l)$ is fully accounted for by the dependence of amino acid conservation on solvent accessibility.

In a formal mathematical language, the H0 means that the intrafamily conservatisms at each position can be treated as independent random values with probability density $f(s|a)$ that depends solely on solvent accessibility a of a position. Since the CoC is given by equation (1) as an average over conservatism within each family, its probability distribution expected under H0 can be derived as convolution of individual probability densities $f(s|a)$ over all representative proteins. Fur-

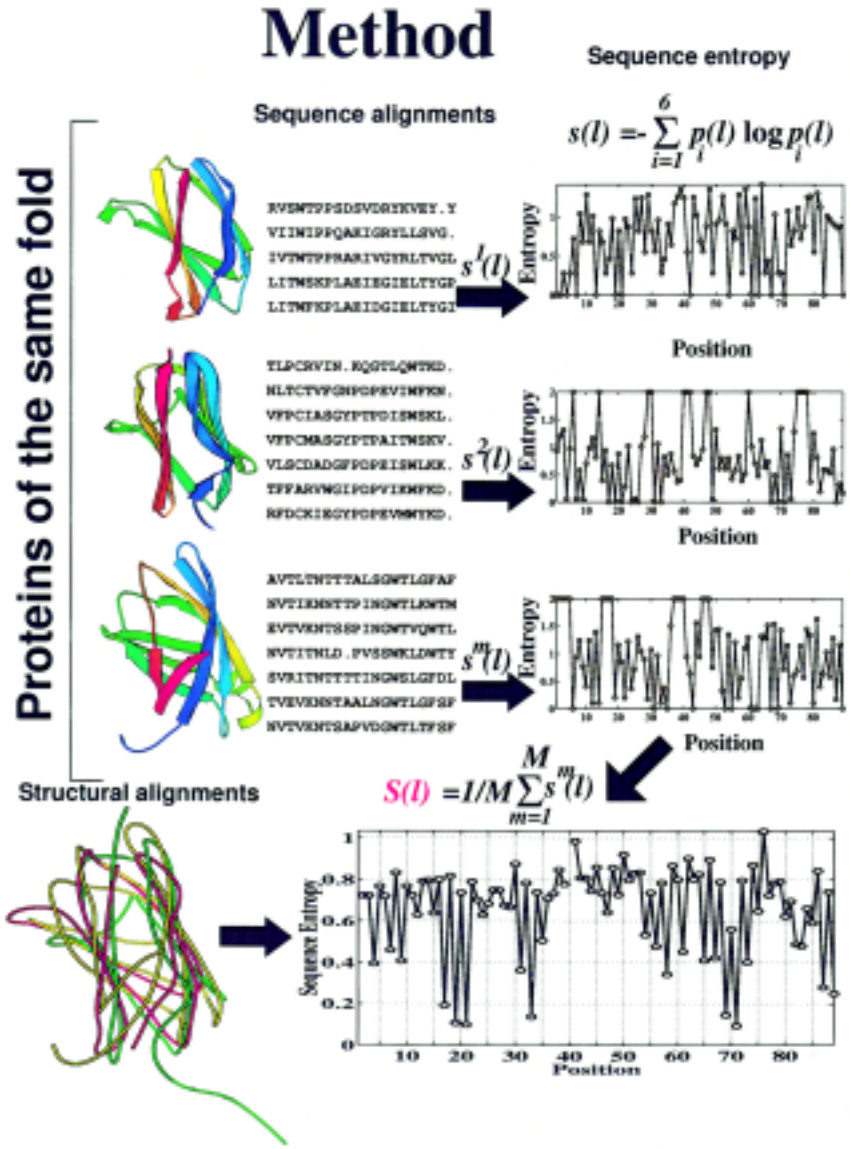

Figure 1. Schematic representation of procedure used to compute conservatism-of-conservatism $S(l)$.

thermore, if the number of analogus is large, the probability density for CoC under H0 will be Gaussian according to Central Limit Theorem. Thus the statistical significance of non-trivial CoC may be estimated by comparing the observed CoC with the predicted value according to the H0, and estimating the deviation between the two in terms of the number of standard deviations in the probability density for CoC obtained according to H0. This can be directly translated into the probability that observed CoC is a trivial consequence of residue accessibility to solvent, as suggested by the zero hypothesis H0.

To carry out this program we first compute the probability $f(s|a)$ of having sequence entropy s at a position with solvent accessibility a. These statistics are taken over all representative structures from the Protein Data Bank (PDB). Next, using $f(s|a)$ and the central limit theorem we compute $S^{exp}(l)$, i.e. the value of CoC expected according to H0. And finally, we compute the probability $P(S)$ of observing $S(l)$ according to hypothesis H0 (see Methods for details). This probability $P(l) \equiv P(S(l))$ together with $S(l)$ are used to assess each position l in studied folds.

Figure 3 presents expected $S^{exp}(l)$ and observed (from actual alignments according to the scheme outlined above) $S^{obs}(l)$ for the immunoglobulin fold. Strikingly, $S^{exp}(l)$ and $S^{obs}(l)$ are in a very good agreement. This fact suggests that most of the variation in $S(l)$ is indeed explained by solvent accessibility (by H0 hypothesis). Correlation between $S^{obs}(l)$ and $S^{exp}(l)$ in this example is $\rho = 0.89$. However, there are few positions that exhibit CoC $S(l)$ well below $S^{exp}(l)$. This gives rise to low values of $P(l)$ that are indicative that at those positions factors other than just solvent accessibility have contributed significantly to conservatism.

From the physical point of view, higher conservatism of buried residues reflects their role in the

stabilization of the fold (buried residues are usually more hydrophobic and form more residue-residue interactions in a protein structure) (Koshi & Goldstein, 1997; Bahar & Jernigan, 1997; Gilis & Rooman, 1997). Remarkable values of correlation between $S^{exp}(l)$ and $S^{obs}(l)$, which vary around 0.9 for studied folds, demonstrate that the dominant contribution to $S^{obs}(l)$ arises from the requirement for thermodynamics stability of a protein structure. The signal which we are interested in, i.e. the deviation of $S^{obs}(l)$ from $S^{exp}(l)$, cannot be explained by the stability requirement alone. Hence, a low value of $P(l)$ indicates some additional evolutionary pressure on a position l in the fold.

Conservatism across the families

Conservatism across the families ($S^{across}(l)$) addresses the following question: are there any positions in the proteins of the same fold that are frequently occupied by the same type of residues in different families. Similar to the CoC computations, we first align sequences within each family using sequence alignment, and then align families against each other using structural alignment. One should b careful to weigh large and small families equally. First, we compute $p_i^m(l)$, the frequence of residue type i at position l within each family m. Next we compute the across-family frequency:

$$P_i(l) = \frac{1}{M} \sum_{m=1}^{M} p_i^m(l)$$

and the across-family entropy:

$$S^{\text{across}}(l) = -\sum_{i=1}^{6} P_i(l) \log P_i(l)$$

A related quantity was analyzed by Ptitsyn (1998) for the cytochrome *C* family. Note that always $S(l) < S^{across}(l)$. To understand the difference between $S(l)$ and $S^{across}(l)$ consider the following example. If position l is conserved in each family, then $S(l)$ is low. If residues of the same type are conserved (e.g. this position is hydrophobic aliphatic in each family), then $S^{across}(l)$ is also low. However, if a position is conserved within each family, but different families have different types of residues at this position, then $S(l)$ is low, but $S^{across}(l)$ is high. Such a situation occurs when one family has a conserved hydrophobic group and another has a conserved cysteine residue forming a disulfide bond or a conserved charged residue participating in a salt bridge. Another class of positions which has low $S(l)$ and high $S^{across}(l)$ correspond to functional super-sites in protein folds. In this case each family has a conserved active site esidue at position l, but since function is different, the types of these residues are different in the different families. In this case a pronounced difference between $S(l)$ and $S^{across}(l)$ would indicate a presence of a super-site. We should note however, that unlike CoC which is a well-defined statistical quantity, S^{across} is less well-defined statistically since it is more dependent on the evolutionary history within individual families. However, we may consider a qualitative difference between S^{across} and $S(l)$ at some position as an heuristic indicator that amino acid residues at such positions have been under an ''unusual'' evolutionary pressure (often related to a super-site or in some cases to folding nucleus, see below). Since $S(l)$ and $S^{across}(l)$ correspond to different patterns of evolutionary pressure we consider signals reflected in both these quantities.

Importantly, both low $S(l)$ and low $P(l)$ are used as indicator of strong conservation in our analysis. In fact, low $S(l)$ and high $P(l)$ mean that although position l is conserved in several families, the degree of conservation does not exceed that expected from consideration of solvent accessibility factor. Conversely, when $S(l)$ is moderate and $P(l)$ is low, position l may not be very conserved in the families, although even this weak conservatism is unusual for positions of such (usually, high) solvent accessibility. This kind of weak but significant CoC could be a signature of a solvent exposed common binding/active site. Examples of such situations will be discussed in more detail (see oligonucleotide-binding fold below).

Immunoglobulin fold

Immunoglobulin fold is the most populated one among known beta-proteins. Tenascin (1ten) is used as a representative protein for this fold. Figure 4 presents $S(l)$, $S^{across}(l)$ and $P(l)$. Positions with low $S(l)$ and low $P(l)$ are the ones that exhibit high CoC. Positions with $P \leqslant 10^3$ in the immunoglobulin fold are marked with stars on Figure 4. There are six positions with $P(l) \leqslant P_c = 10^3$ and $S(l) \leqslant 0.2$ in tenascin: Ala17(819), Ile19(821), Trp21(823), Leu33(835), Val69(871), and Leu71(873). (Here and below residues are counted from the beginning of the PDB file, residue numbers used in the PDB file are shown in parentheses). Figure 5 shows those six residues in the structure of tenascin. Importantly, residues with high CoC form a dense cluster in the core of the protein, connecting strands B, C and F (strand notations are according to Branden & Tooze (1998)).

What is the origin of the high CoC at some positions of the immunoglobulin fold? As stated in the Introduction, three factors can account for high CoC: (i) a super-site; (ii) key positions responsible for stabilization of the fold; and (iii) a folding nucleus, whose stability is required for fast folding. We consider all three possibilities for the immunoglobulin fold.

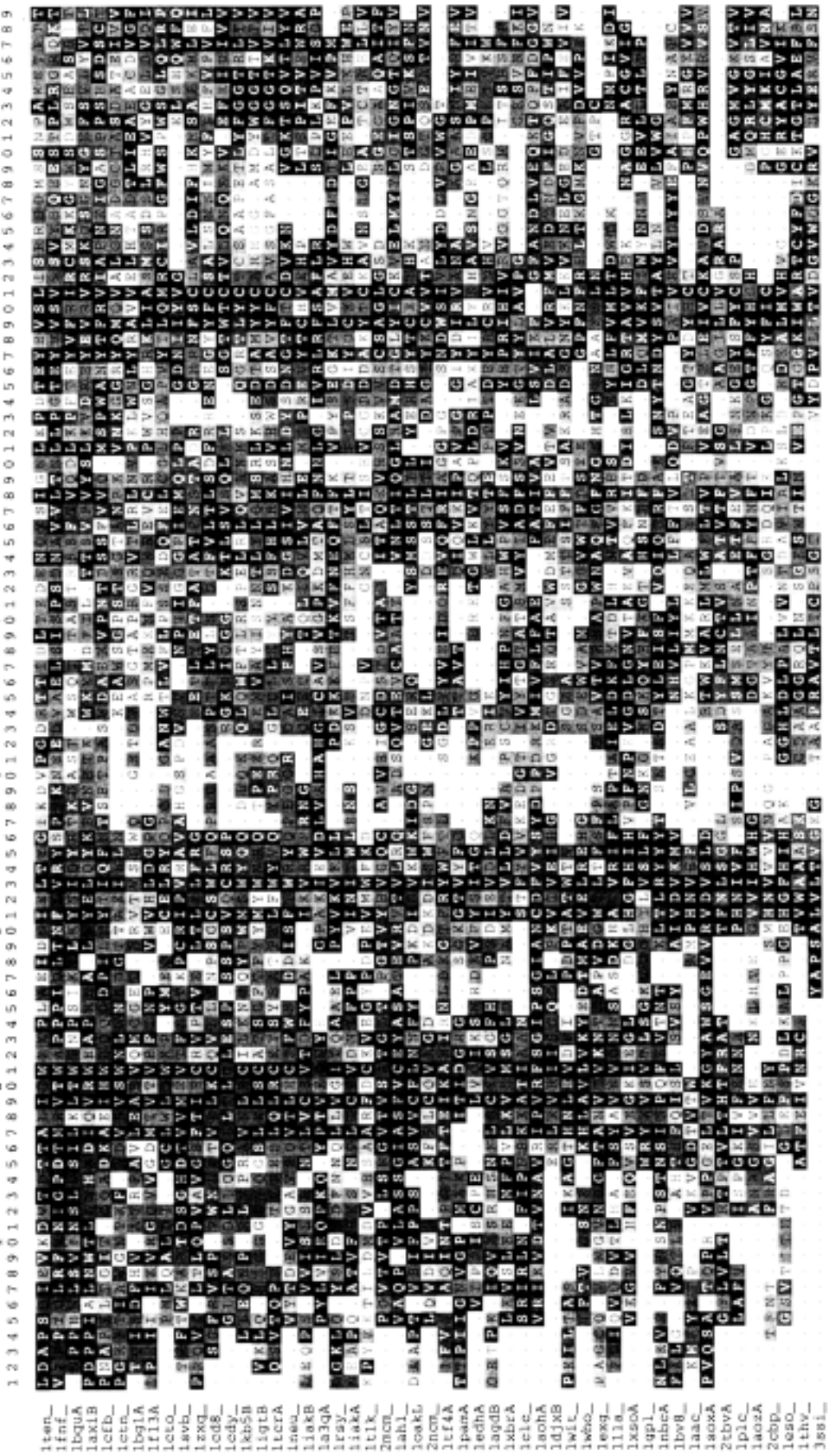

Figure 2 (*Legend shown opposite*)

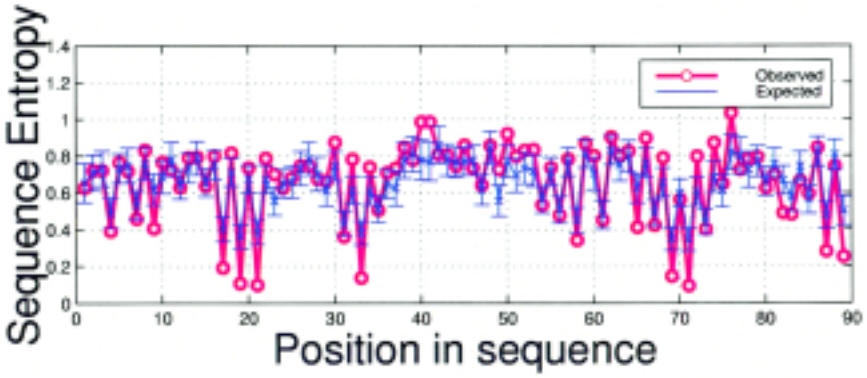

Figure 3. Observed $S^{obs}(l)$ (red) and expected $S^{exp}(l)$ (blue) CoC in the immunoglobulin fold. $S^{exp}(l)$ is calculated based on the solvent accessibility. The remarkable correlation of 0.9 shows that solvent accessibility explains most of the conservatism in protein families Error bars on $S^{exp}(l)$ shows one standard deviation $\sigma(S)$.

Function

Most of the proteins of immunoglobulin fold are extracellular domains responsible for specific binding and/or recognition of small ligands (Fab), peptides (MHC), DNA (p53 DNA-binding domain) or other proteins (hormones, proteins of extracellular matrix, other receptors). Different proteins use very different parts of the fold for specific binding and the binding site is often located on the junction between the two immunoglobulin domains. For example, growth hormone receptor (1cfb) has two domains with the immunoglobulin fold which bind the hormone by their loops, but one domain uses loops at one end of the fold and the other uses the loops located on the opposite side of the fold. In general, there is no specific part of the immunoglobulin fold which is used for binding/active site placement. Hence, high CoC in this fold cannot be explained by conservation of functional residues.

Stability

The low values of $P(l)$ shows that high CoC cannot be explained by solvent accessibility. Hence, high CoC is hardly a result of conservation driven by a requirement for thermodynamic stability. These positions are under some additional evolutionary pressure. Some proteins of the immunoglobulin fold have disulfide bridges at the high CoC positions (see below). This observation points to some special role of the high CoC positions in the fold stabilization and/or initiation.

Kinetics

Importantly, tenascin and some other proteins of the immunoglobulin fold (twitchin, FNIII-9, FNIII-10) are known to fold fast and by a two-state mechanism (Plaxco *et al.*, 1996; Hamill *et al.*, 1998). They are expected to have a stable folding nucleus (see above), i.e. a set of residues that interact with each other in the transition state. Stability of the nucleus provides rapid folding to the native state and, hence, residues contributing to the nucleus are conserved (Mirny *et al.*, 1998). If location of the folding nucleus in the protein structure depends primarily on the fold and not on the sequence, then positions belonging to the nucleus should exhibit high CoC.

Strong evidences in support of this view is provided in a recent experimental study by Lorch *et al.* (1999). These authors studied the N-terminal domain of rat CD2 which has no disulfide bonds. Making nine mutations in the core of this protein and measuring stability and folding kinetics associated with each mutation, they identified residues belonging to the folding nucleus as Ile18, Val30, Trp32 and Val78. When the structure of the CD2 is superimposed with tenascin, the folding nucleus of CD2 maps onto positions Trp21(823), Ile31(833), Leu33(835) and Leu71(873) in tenascin, all of which exhibit statistically significant CoC (see Figure 4).

Different mechanisms of stabilization

As one can see from Figure 4(b), the positions that exhibit significant CoC have low values of across-conservatism $S^{across}(l)$, indicating that these positions carry residues of the same type in most of the families. A substantial difference between

Figure 2. Structural alignment of families in immunoglobulin fold. Each line corresponds to a single family and representative proteins for each family are indicated. The grayscale level shows conservatism within each family: dark, conserved; light, variable. For each family we present the PDB code and the sequence of the representative protein. Note dark vertical strips corresponding to positions with high CoC (positions 17, 19, 21, 33, 69 and 71).

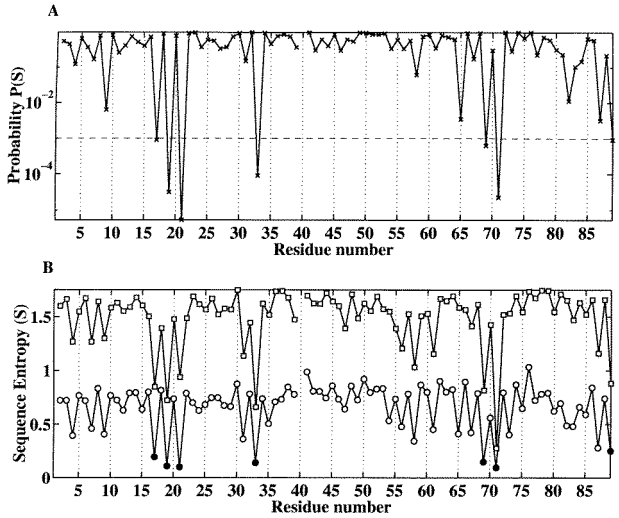

Figure 4. Conservatism in immunoglobulin fold. (a) Probability $P(S < S^{obs})$ of observing $S(l)$ by chance. (b) Observed $S(l)$ (circles) and $S^{across}(l)$ (squares). Positions with $P < 10^{-3}$ are shown by filled circles.

$S^{across}(l)$ and $S(l)$ also shows that not all families have the same types of residues in the outlined positions. Figure 2 presents structural alignments of the proteins of the immunoglobulin fold. Grayscale level indicates the degree of conservation within each family ($s^m(l)$). Positions with high CoC can be seen as dark vertical strips on the diagram.

From the diagram (Figure 2) one can see that different families may have different residues at the high CoC positions. Importantly, pairs of interacting residues from the high CoC set of tenascin correspond to disulfide bridges in some immunoglobulins. In particular, Trp21 and Leu71 in tenascin correspond to a disulfide bridge Cys23-Cus92 in the beta chain of the 14.3.D T-cell antigen receptor (1bec) and to Cys23-Cys94 in CD8 (1cd8). The pair Leu33 and Leu71 corresponds to a disulphide bridge in the second domain of CD4 (3cdy) and to one of the bridges in the first domain of growth hormone receptor (1axi). The presence of the disulfide bridges indicates evolutionary pressure to stabilize interactions between these positions.

Interestingly, in many families of the immunoglobulin fold a strong conserved tryptophan residue is found at one of the six outlined positions. For example, tenascin has Trp21, whereas CD8, (1cd8), Kb5-c20 t-cell antigen receptor (1kb5B), Igg2a intact antibody (1igtB) and myelin p0 protein fragment (1neu) have a tryptophan residue at position corresponding to Leu33 in tenascin. CD4 (1cdy) has a tryptophan residue corresponding to Val69 in tenascin. This ''circular permutation'' of tryptophan illustrates a possibility of correlated mutations in the putative folding nucleus of very far diverged proteins.

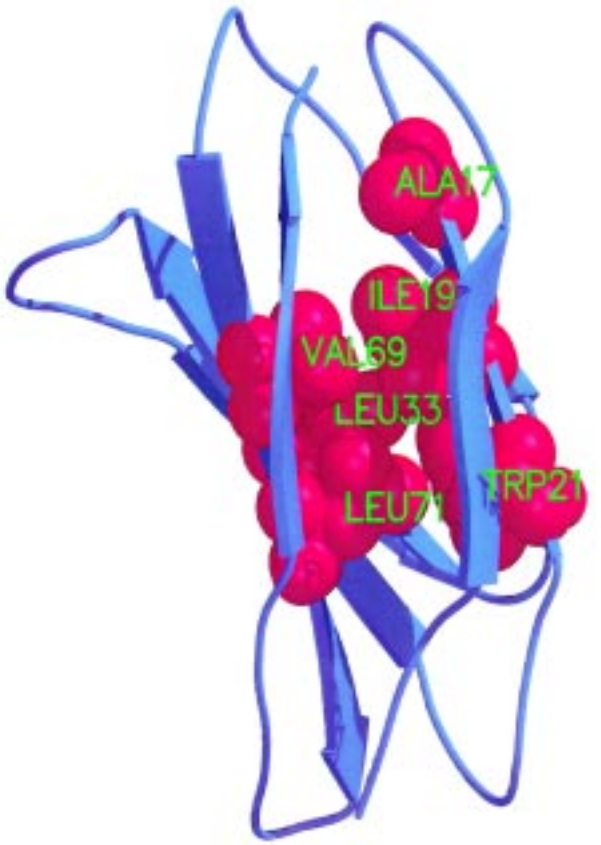

Figure 5. Structure of tenascin (the immunoglobulin fold). Residues with high and significant CoC are shown by space-filling models. All protein structure cartoons are produced using Molscript (Kraulis, 1991) and Raster3D (Merritt & Bacon, 1997).

Both examples demonstrate the distinction between $S(l)$ and $S^{across}(l)$. Proteins of the immunoglobulin fold use different types of interaction (e.g. disulfide bonds, hydrophobic and aromatic stacking) to stabilize the same positions in the structure. Hence, different types of residues are conserved in different families. Such a diversity of conserved residues makes $S^{across}(l)$ high, while keeping $S(l)$ low. However, in many positions of immunoglobulin fold dominant factor in stabilization of the folding nucleus are hydrophobic aliphatic interactions and hence $S^{across}(l)$ has rather low values at positions where $S(l)$ is low.

A previous study of sequence and structure conservation in the immunoglobulins (Bork *et al.*, 1994) did not identify those key positions in the fold. Instead, the authors came to the conclusion that ''no single interaction (or localized set of interactions) can be uniquely identified as a principal determinant of the Ig-like fold''. Although conserved residues in each family were identified, the fact that different types of conserved residues were found at corresponding positions in different proteins obscured the analysis made by Bork *et al.* (1994).

Oligonucleotide-binding fold

The oligonucleotide-binding (OB) fold is a β-fold with a barrel topology. Proteins belonging to this fold have diverse sequences and functions. About 20 non-homologous protein families share this fold. Figure 6 presents $S(l)$ and $P(l)$ for the major cold shock (CspA) protein (1mjc in PDB), a typical OB fold protein. Positions with high CoC ($S(l) \leqslant 0.2$ and $P(l) < 1\,\%$) in CspA are: Val8(9), Ile20(21), Val50(51) and Val66(67). These residues form a dense cluster located inside the β-barrel, closer to one open end of the barrel (see Figure 7). In contrast to immunoglobulin fold, where the high CoC cluster is located in the center of the protein, residues with high CoC in the OB fold are asymmetrically grouped near ''the bottom'' of the barrel. Strands 1, 2, 4 and 5 are involved in the cluster. Hence, when the four outlined residues come together, the overall topology of the chain becomes well defined.

The across-conservatism S^{across} has deep minima at the outlined positions demonstrating that most of the analogs carry the same type of residues there.

What is the origin of the CoC in the OB fold: function, stability or kinetics? Proteins of the OB fold are (i) nucleotide binding (transcription signals, RNA binding, tRNA synthetase, staphylococcal nucleases); (ii) inorganic pyrophosphates; (iii) tissue inhibitor of metalloproteinases; or (vi) toxins. The binding site of single-stranded DNA and RNA is localized mostly on the face of the molecule in strands 2, 3 and on the surface loops (Newkirk *et al*., 1994). Residues contributing to this site are different from the four residues belonging to the high CoC cluster shown in Figure 7. Therefore, functional conservation cannot account for the observed CoC.

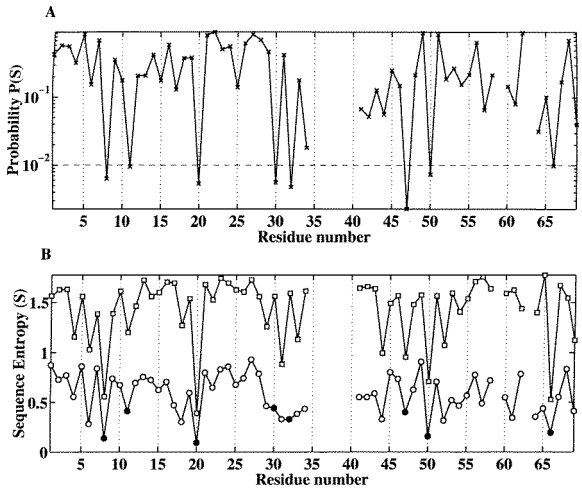

Figure 6. Conservatism in OB fold. (a) Probability $P(S < S^{obs})$ of observing $S(l)$ by chance. (b) Observed $S(l)$ (circles) and $S^{across}(l)$ (squares). Positions with $P < 10^{-2}$ are shown by filled circles.

Figure 7. Structure of major cold shock protein CspB (OB fold). Residues with high and significant CoC are shown by space-filling models.

Thermodynamic properties of the cold shock proteins are very well studied (Perl *et al*., 1998; Reid *et al*., 1998; Schindler *et al*., 1998, 1999). This family of proteins serves as a clear example where thermodynamics are separated from the kinetics of folding: different proteins in the family have very different stability, but all fold very fast and by a two-state mechanism (Perl *et al*., 1998; Reid *et al*., 1998; Schindler *et al*., 1999). A great variation in stability with almost no changes in folding rates demonstrates that amino acid residues responsible for stability and fast folding are located in different regions of the structure of these proteins. Unfortunately, there is no study where stability and folding kinetics of a variety of CspA/CspB mutants are measured. Schindler *et al*. (1998) mutated surface-exposed phenylalanine residues in CspB and showed a substantial destabilization upon mutation of Phe15, Phe17 and Phe27 (corresponding to Ph17(18), Phe19(20) and Phe30(31) in CspA). Stabilization of the CspB structure by exposed phenylalanine residues is another evidence that different regions of the protein structure are responsible for stability and for kinetics.

Low values of $P(l)$ for the four outlined positions in the OB fold indicate that CoC in those positions can hardly be explained by solvent accessibility alone. Thus we predict that Val8(9), Ile20(21), Val50(51) and Val66(67) constitute a folding nucleus in the OB fold proteins, and their conservation gives rise to the conservation of rapid folding.

Note that several positions in the OB fold (Phe11(12), Phe30(31), His32(33) and Gly47(48))

exhibit CoC that is moderate in absolute value ($0.2 < S(l) \leqslant 0.5$) but is of high statistical significance ($P < 1\,\%$). Interestingly, these positions constitute a nucleotide/phosphate binding super-site. We examined several proteins having the OB fold and found that in all nucleotide binding proteins (nucleases, DNA and RNa binding, etc.) and inorganic pyrophosphatases the active/site is localized at the same face of the barrel and involves these exposed aromatic residues (Newkirk *et al.*, 1994; Schindelin *et al.*, 1994). For example, Arg35 is central to the active site of Staphylococcal nuclease (1snc). This position corresponds to His32(33) in CspB; ferredoxin-NADP+ reductase (1fnc) places its FAD binding site in the same location. On the other hand, toxins that have the OB fold do not use this face of the barrel for specific binding. Instead, they form large complexes where a helix located on the top of the barrel is involved in ATP binding (e.g. pertussis toxin S2/S3 subunits). Hence, these residues are conserved in nucleotide/phosphate binding proteins but not in the toxins forming a ''weak super-site''. As a result, $S(l)$ is not very low, but still very significant (low $P(l)$) for such exposed positions. It is an important feature of the CoC analysis that allows us to identify a consensus between functionally conserved positions in non-homologous proteins of the same fold. Later we discuss possible biological implications of this peculiar interplay of function and stability in the cold shock proteins.

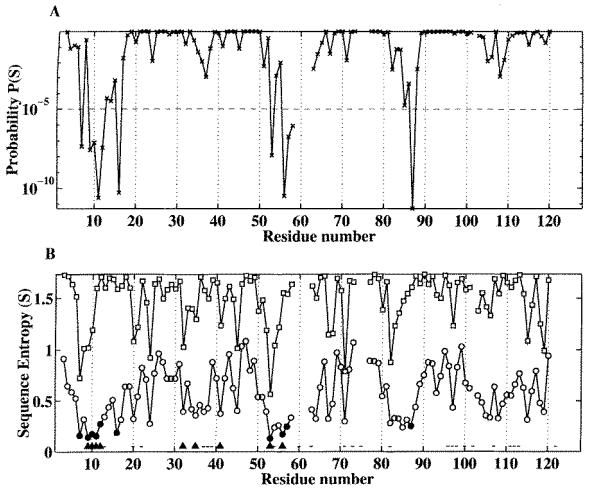

Figure 8. Conservatism in Rossman Fold. (a) Probability $P(S < 10^{obs})$ of observing $S(l)$ by chance. (b) Observed $S(l)$ (circles) and $S^{across}(l)$ (squares). Positions with $P < 10^{-5}$ are shown by filled circles.

Rossman fold

The Rossman fold is the most populated fold among α/β-folds. We use chemotactic protein CheY (3chy) as a representative of this fold. Figure 8 presents CoC and $S^{across}(l)$ for CheY. Residues with $S(l) < 0.3$ and $P(l) < 10^{-5}$ are Phe7(8), Val9(10), Val10(11), Asp11(12), Asp12(13), Met16(17), Val53(54), Asp56(57), Trp57(58) and Ala87(88). The numbers are from the PDB file, and the numbers in parenthesis are as reported by Lopez-Hernandez & Serrano (1996). In the protein structure (see Figure 9) residues Asp11(12), Asp12(13), Met16(17), Asp56(57), Trp57(58) and Ala87(88) form a dense, solvent-exposed cluster at the C termini of strands 1, 3 and 4 and N termin of helix 1. Residues Phe7(8), Val9(10), Val10(11) and Val53(54) are lined on strands 2 and 3. The exposed cluster of Asp11(12), Asp12(13) and Asp56(57) is stabilized by a bound Mg^{2+}.

The folding nucleus of CheY was identified by Lopez-Hernandez & Serrano (1996). They mutated several positions scattered through the whole protein and measured changes in stability and the (un)folding rate of the mutant proteins. Importantly, for most of the mutated positions ϕ-values are either close to 0 or above 0.5. Comparison of the CoC data with the measured ϕ-values give evidence in support of kinetic origin of the CoC in Rossman fold. Positions where mutations had been made are shown on the bottom of Figure 8, with triangles marking those with $\phi > 0.5$. Agreement between high CoC positions and those with $\phi > 0.5$ is very good (both high CoC and $\phi > 0.5$: Val9(10), Val10(11), Asp11(12), Asp12(13), Val53(54), and Asp56(57); there are no positions with high CoC and $\phi \leqslant 0.5$; residues Phe7(8), Met16(17), Ala87(88), no measurements; low CoC and $\phi > 0.5$ Val32(33), Ala35(36), Ala41(42)). No mutations were made for Phe7(8). Residue Ala87(88), which has $S(l) = 2.29$ and $P < 10^{-10}$, may also be important for kinetics as an A87G mutation makes the

Figure 9. Structure of CheY protein (Rossman fold). Residues with high and significant CoC are shown by space-filling models.

protein fold much slower without affecting its stability (L. Serrano, personal communication). Being at the end of strand 5, Ala87 can be responsible for terminating this strand. Another explanation is its functional role in the proteins of Rossman fold (see below). Positions Val32(33), Ala35(36), Ala41(42), which have $\phi > 0.5$ but no CoC, probably belong to ''an extended folding nucleus'', a part of the nucleus which vary from family and hence exhibit no CoC signal. This scope of data from protein engineering experiments strongly supports a link between CoC and folding kinetics.

The Rossman fold is known to have a super-site with functional residues located at the C termini of the β-strands (Branden & Tooze, 1998; Russell *et al*., 1998b). In CheY these positions are 11-12, 16, 56 and 87. Remarkably, the same positions in the fold are used to provide fast folding (as shown by protein engineering experiments) and to build the active/binding site. Residues Asp11(12), Asp12(13), Met16(17) and Asp56(57) form Mg^{2+} binding sites (Lopez-Hernandez & Serrano, 1996); residue Ala87(88) is in contact with Lys109 and is probably involved in the mechanism of allosteric transition in CheY (Welch *et al*., 1994; Bellsolell *et al*., 1996). There is an important relation between function of CheY and its stability: cation binding substantially stabilizes the structure of CheY and coordinates the charged side-chains of Asp11(12), Asp12(13), and Asp56(57) (Wilcock *et al*., 1998). In the Discussion we examine possible biological implications for linking protein function with kinetics and stability. Importance of the high CoC positions in the Rossman fold for both function and folding kinetics imposed a strong evolutionary pressure at those positions.

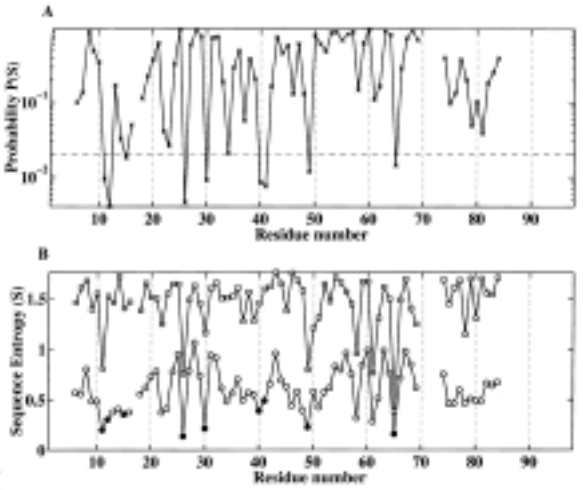

Figure 10. Conservatism in alpha/beta plait. (a) Probability $P(S < S^{obs})$ of observing $S(l)$ by chance. (b) Observed $S(l)$ (circles) and $S^{across}(l)$ (squares). Positions with $P < 2\,\%$ are shown by filled circles.

The alpha/beta plait

The alpha/beta plait has an antiparallel $\alpha + \beta$-topology and consists of a few helixes and four-stranded β-sheet. This fold is the third most populated after TIM barrels (see below) and Rossman folds. Proteins of this fold have very diverse functions, thermodynamic and kinetic properties (Villegas *et al*., 1989; van Nuland *et al*., 1998a,b; T. Ternstro *et al*., unpublished results).

Results for this fold are presented in Figure 10. We chose acylphosphatase (pdb:2acy) as a representative protein. In acylphosphatase; positions with high and significant CoC ($S(l) < 0.25$ and $P(l) < 2.5\,\%$) are: Tyr11, Thr26, Gly30, Gly49 and Leu65. Note that statistical significance of these results is lower than for all other folds described above. This lower statistical significance comes from fewer non-homologous families known to have this fold. Only 29 families were used in our analysis in contrast to 51 families for immunoglobulin and 166 families for Rossman folds.

Similar to other cases, residues with high and significant CoC are all located close to each other in space (see Figure 11). This CoC cannot be attributed to functional conservatism, since there is no super-site for this fold (Branden & Tooze, 1998; Russell *et al*., 1998b). Functional/binding residues are typically located at either a solvent-exposed face of the beta-sheet (active site of the glutamine synthetase (Liaw *et al*., 1994); catalytic site of BIRA protein (Wilson *et al*., 1992); active site of the

Figure 11. Structure of ADA2 h protein (alpha/beta plait). Residues with high and significant CoC are shown by space-filling models.

human DNA polymerase beta (Pelletier *et al*., 1996)) or at the loops (ligand binding site of the D-3-phosphoglycerate dehydrogenase (Schuller *et al*., 1995)).

Comparison of our results with those from protein engineering experiments (Villegas *et al*., 1998) indicate that CoC at some positions is related to the folding nucleus. For the four proteins of the alpha/beta plait: human procarboxypeptidase A2 (ADA2h), spliceosomal protein U1A, acylphosphatase (AcP), and histidine-containing phosphocarrier protein (HPr), folding kinetics and thermodynamics have been studied (van Nuland *et al*., 1998a,b; T. Ternstro *et al*., unpublished results; Villegas *et al*., 1998). For two of these proteins (ADA2h and U1A), the transition state (and folding nucleus) has been characterized. Importantly, all four proteins exhibit two-state folding transition. However, AcP and HPr fold slowly ($k_f^{H_2O} = 0.23\ s^{-1}$ and $k_f^{H_2O} = 14.9\ s^{-1}$, respectively), while ADA2 h and U1A fold very fast ($k_f^{H_2O} = 897\ s^{-1}$ and $k_f^{H_2O} = 316\ s^{-1}$).

When the structures of ADA2h, U1A and AcP are superimposed, some nucleation residues coincide with each other and with the high CoC residues. Particularly, there is a clear consensus between two experimentally identified and one putative nucleus residue in positions Tyr11 and Thr26 in AcP (Ile14 and Leu30 in U1A; Ile15 and Leu26 in ADA2 h). Another position that can also belong to the folding nucleus in Leu65 in AcP (Phe65 in ADA2 h; Phe34 in U1A, or perhaps Met72 or Met82, which were not studied in experiment). Other nucleation residues in ADA2 h and U1A do not coincide with each other with positions of the high CoC. These residues either constitute ''an extended folding nucleus'', which varies from family to family, or are under some other sort of evolutionary pressure. The difference between folding nuclei in U1A and ADA2h may be due to a very different twist of the U1A structure and substantial angle between the first helixes (M. Oliveberg, personal communication).

TIM barrel

TIM barrel is the third most populated α/β-fold. Sequences and functions of the proteins sharing this fold are very diverse. Very little is known about stability of the TIM barrel proteins and no data are available regarding their folding kinetics. On the other hand, functions of the majority of TIM barrel proteins are well known. This fold has a distinctive super-site at the loops on the top of the barrel (Russell & Ponting, 1998; Branden & Tooze, 1998).

Figure 12 presents results of our analysis of the TIM barrel fold mapped onto the structure of endo-beta-*N*-acetylglucosaminidase (EBN, pdb:2ebn). Remarkably, solvent accessibility is a very good predictor of the CoC signal yielding a correlation of 0.9 between $S^{obs}(l)$ and $S^{exp}(l)$ over the whole structure of about 300 residues. The only positions in EBN with high significant

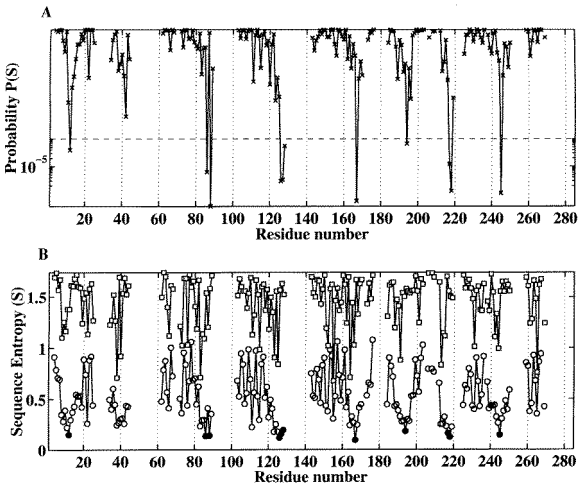

Figure 12. Conservatism in TIM barrels. (a) Probability $P(S < S^{obs})$ of observing $S(l)$ by chance. (b) Observed $S(l)$ (circles) and $S^{across}(l)$ (squares). Positions with $P < 10^{-4}$ are shown by filled circles.

CoC ($P(l) < 10^{-3}$ and $S(l) < 0.3$) are: Thr12(16), Ser42(46), Ser86(90), Leu88(92), Asp126(130), Asp127(131), Glu128(132), Tyr167(171), Asp194(198), Ser217(221), Gln218(222), and Phe245(249). All these positions belong to the super-site.

In 19 TIM barrel representative structures, the active site is reported in the PDB file (record ''SITE''). A total of 67 positions are reported. The distribution of these positions along the fold is the following. The vicinity (±three residues) of Thr12(16) contain four active site positions, Ser42(46)-4; Ser86(90)/Leu88(92)-6; Asp126(130)/Asp127(131)/Glu128(132)-16; Tyr167(171)-16; Asp194(198)-6; Ser217(221)/Gln218(222)-4. In total, 58 out of 67 reported active site residues are localized at the high CoC positions. All of them are located at the top loops connecting the β-barrel with helixes (see Figure 13).

Importantly, the ''super-site'' induces no signal on S^{across}, since different amino acid residues are used in the active sites of different families. The CoC, in contrast, very clearly identifies the super-site, since amino acid residues in the active site are very conserved within each family.

Discussion

Here, we report a detailed study of molecular evolution of five of the most populated protein folds. Out of ≈2200 domains in known structures without evident sequence homology, 564 belong to five dominant folds that were analyzed in this study. High data-demanding nature of the method of analysis limits it to the folds that contain at least 20-30 non-homologous families. As the number of solved protein structures increases, this analysis can be extended to other folds.

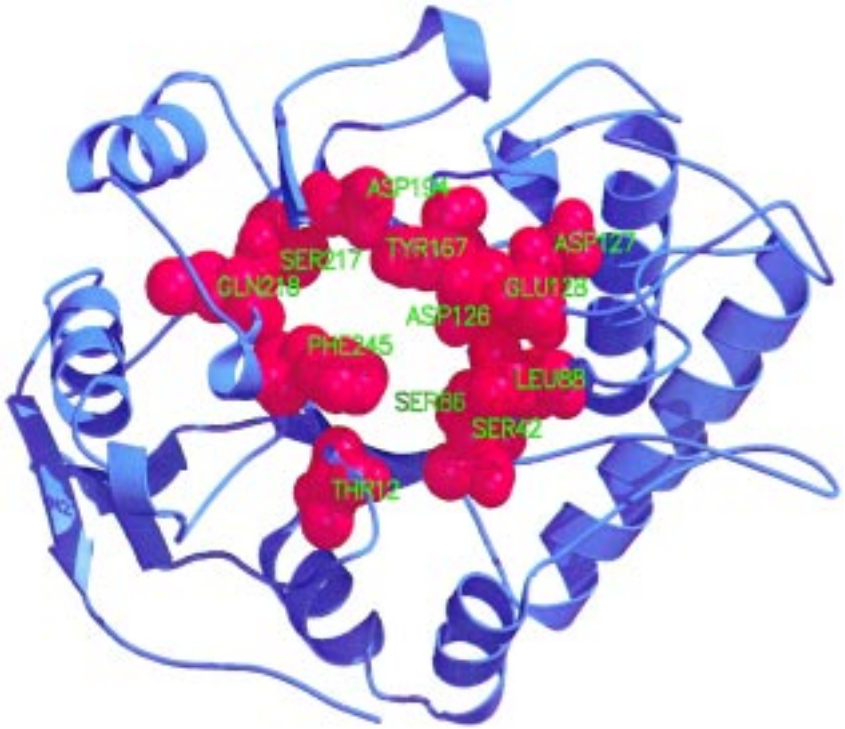

Figure 13. Structure of endo-beta-*N*-acetylglucosaminidase (TIM barrel). Residues with high and significant CoC are shown by space-filling models. All those residues correspond to the active site positions in TIM barrel proteins.

For each of folds shaded, we identified positions that are conserved in most of the protein families sharing this fold. In all studied folds residues that show high CoC form a dense cluster in the native structure. Location of this cluster and the nature of interactions stabilizing it are, however, different in different folds and even in different families of the same fold. For example, in immunoglobulin domain residues with high CoC form a cluster deeply buried into the fold. Some families of this fold stabilize this cluster by hydrophobic interactions, some by the disulfide bonds. On the contrary, proteins of the Rossman fold have high CoC residues mostly on the solvent-exposed helix-sheet loops. In different families having Rossman fold this cluster is stabilized by either Coulomb interactions between aspartic acids and a bound metal ion, or by hydrophobic interactions, or, perhaps, by interactions with the ligand in flavodoxins. Similarly in the OB fold the position of the CoC cluster is shifted to the bottom of the β-barrel

Correlated mutations

Different interactions between high CoC residues in different families lead to the emergence of ''correlated mutations'' (Altschuh *et al*., 1988; Thomas *et al*., 1996). In fact, substitution of the two hydrophobic residues by two cysteine residues forming a disulfide bridge can be considered as a clear example of a correlated mutation. Several substitutions of these kind can be observed when families of a particular fold are aligned with each other (by a sequence alignment within a family and by a structural alignment between the families). Deeper analysis of these cases leads to a very different picture for correlated mutations.

In the immunoglobulin folds, substitution of a hydrophobic pair in high CoC positions by a pair of cysteine residues is typical (see above). However different pairs of residues are substituted by cysteine residues, e.g. position 71 in tenascin can be occupied by a cysteine forming a disulfide bridge with either position 21 or position 33. Hence, an analysis of any of these pairs (71-21 or 71-33) reveal very little or no correlation. A more striking example is the ''circular permutation'' of the tryptophan residue among the high CoC positions of the same fold (see above). This cluster of residues usually contains a single tryptophan residue which can be at either position 19, 21, 33, 69 or 71. Clearly, no pair of these positions exhibit correlated mutations. It is the whole cluster, not any single pair, that exhibits correlated mutations. This analysis explains why observed correlated mutations are so rare: there are several ways to stabilize even a small cluster of residues and no pair of positions in this cluster is superior to the others. Another lesson is that correlated mutations do exist, but they involve more residues than two. Identification of such cases requires analysis of far diverged proteins and hence can only be done over alignments of several families sharing the same fold. These correlated mutations in distantly diverged homologous or analogous proteins is a manifestation of the fact that interactions are more conserved than residues in protein evolution.

Origins of CoC

The main goal of our study is to find, in each specific case the physical and evolutionary reason for the observed conservatism. Apparently the most common selection pressure is thermodynamic stabilization of sequences which should serve as a ''noise baseline'' for our analysis. One can expect (see below) that pressure towards thermodynamic stabilization will be stronger on amino acid residues that participate in a larger number of intra-protein interactions, i.e. the ones that are more buried in structure. To this end one would predict a strong correlation between conservation and any measure of buriedness of an amino acid residue, such as solvent accessibility. Our analysis indeed reveals a (surprisingly) strong correlation between CoC and solvent accessibility.

A more detailed quantitative explanation of the correlation between CoC and solvent accessibility comes from statistical-mechanical theory (L.A.M. & E.I.S., unpublished results). The formal analysis is based on the detailed analogy between protein sequence selection and certain statistical-mechanical spin models (Shakhnovich & Gutin, 1993b; Shaknovich, 1998b). Within this analogy the degree of evolutionary pressure towards stabilization is analogous with temperature in statistical mechanics; it can be shown that more buried amino acid residues are at ''lower'' effective temperature, i.e. they are indeed under stronger evolutionary pressure towards stabilization. We would like to stress that solvent accessibility in our analysis serves a measure of amino acid involvement in a protein structure. The statistical-mechanical explanation of the correlation between CoC and solvent accessibility does not imply or assume that interaction with solvent is an only or a dominant force in protein stabilization. Our analysis rather points out to the integral effect of all interactions that lead to protein stabilization as a major reason for correlation between solvent accessibility and CoC. Thus we conclude that the ''baseline noise'' level of our analysis, i.e. the correlation between CoC and solvent accessibility, actually accounts for the evolutionary selection of stable sequences (not necessarily the most stable ones but just at some acceptable level).

However, we found a number of universal positions in each fold which CoC is much stronger than that expected from the stability pressure alone. These positions are obviously at some additional selective pressure. Since individual evolutionary histories and functions of analogs are very different, the only common features that they share is their native structure (fold). Hence the origin of the stronger than expected CoC can be attributed primarily to the evolutionary pressure to preserve some structural features. Among these features two factors dominate in determining the CoC: the function super-site and the folding nucleus. We also cannot exclude the possibility that some of the high CoC positions may play a somewhat special role in stabilizing the fold (serving as ''anchor'' positions), and hence, are under stronger evolutionary pressure than expected from the solvent accessibility only (L.A.M & E.I.S., unpublished results). A possible example of this kind may be helix or β-initiation and termination signals.

High CoC in most of identified positions corresponds to either super-site or to the folding nucleus. Experimental evidence exists that for three out of five studied folds (immunoglobulin, Rossman fold and alpha/beta plaits) the high CoC is indeed related to folding nucleus. However, analysis of the high CoC positions in different proteins brought us to a surprising conclusion that some positions in proteins contribute to both the active site and the folding nucleus, or to the binding/active site and are essential for stabilizing the native structure. Consider this interplay between function, stability and of folding kinetics in more detail.

In the immunoglobulin domain functional and structural load is clearly separated: loops are responsible for binding and recognition while interactions between several residues of the buried core provide stability and fast folding. Importantly, stability and kinetics seam to use the same set of interactions. As we noted above the high CoC positions are located at strands B, C and F; those amino acid residues form the folding nucleus. NMR and HD labeling experiments, however, indicate that these strands have the highest protection index in different proteins of the immunoglobulin fold (Parker *et al*., 1998; Meekhof & Freund, 1999). Therefore, for proteins having an immunoglobulin fold one can expect a noticeable correlation between folding rate and stability.

In the major cold shock protein different structural elements are responsible for stability and function and folding rate. Proteins from the family of cold shock proteins have various stability (ΔG ranges from 2.7 to 6.3 kcal/m) exhibiting, however, in all cases very fast folding ($k_f^{H_2O} \approx 1000\ s^{-1}$) (Perl *et al*., 1998; Reid *et al*., 1998; Schindler *et al*., 1999). This indicates that thermodynamic stability and folding kinetics for this fold are provided by different interactions (and different residues). Stability and binding (most of these proteins bind DNA or RNA) are, in turn, provided by the same set of solvent exposed aromatic residues (Newkirk *et al*., 1994; Schindelin *et al*., 1994). This link between stability and binding has an important biological implications. Marginally stable cold shock proteins get stabilized when bound to the DNA. The excess of these proteins which are not bound to the DNA are rapidly eliminated by proteolysis (Schinlder *et al*., 1999). Clearly this kind of regulation favors selection of fast folding and marginally stable proteins. (Slow folding proteins will be eliminated by proteolysis before they bind the DNA; too stable proteins will not be removed by proteolysis and hence destroy the regulation). It is possible that such evolutionary pressure may

have been applied to other regulatory proteins. Very interestingly, the observed significant CoC in the putative folding nucleus positions of the OB fold provides a direct evidence, for this fold, of special evolutionary pressure towards fast folding, the biological reason for such pressure has been explained before.

This is in contrast to Rossman fold proteins where the super-site and nucleus are close to each other. The reason for this is that most proteins having this fold are enzymes. Chemical catalysis performed by enzymes requires precise (up to fraction of *A*) localization and spatial coordination of electrophylic and nucleophilic groups. Therefore, the most rigid part of a structure may be most suitable as a location of an enzymatic active site. The most conformational rigidity is in the nucleus: This is the part of the structure that forms first in the assembly of the native conformation and is least distorted by local unfolding fluctuations (Abkevich *et al*., 1994). This is due to the fact that nucleus contacts are formed at the folding transition state barrier. Therefore all local unfolding fluctuations that do not reach the top of the folding-unfolding barrier preserve the nucleus intact making the nucleus the most protected from thermal fluctuation part of structure.

To summarize this part of the discussion we note that relation between nucleus, stabilization and function depends very much on the dominant function that in several cases can be associated with a fold. Proteins having immunoglobulin fold participate in specific macromolecular recognition as receptors, cell adhesion proteins, immunoglobulins, DNA-binding domains etc. A multitude of relatively weak non-bonded interactions results in strong and specific interactions between proteins. It is impossible to simultaneously orient and constraint a large number of interacting groups. Hence an induced fit principle may be operational in protein-protein recognition in immunoglobulin fold proteins. Their recognition sites should be located in flexible parts of the molecule (loops) far from the nucleus. Further, since different analogs have different specificity in protein-protein recognition, functional sites in such proteins may vary from protein to protein; hence proteins with such range of functions (having Immunoglobulin or OB fold) may not have a clearly detectable super-sites. In contrast, protein folds that are heavily used primarily by enzymes (Rossman fold and TIM barrel) participate functionally in a very small number of strong chemical, covalent interactions and the required spatial precision of the active site is much higher for such function. In this case it becomes advantageous to place active sites near folding nucleus; this ensures sufficient stability of the active site against thermal fluctuations.

Comparison with protein engineering analysis

The experimental approach to determine folding nucleus has been pioneered by Fersht and co-workers (Itzhaki *et al*., 1995). It is based on the idea to use site-specific mutagenesis to determine which positions are most important for folding kinetics. The results are usually expressed in terms of $\phi = \Delta\Delta G_{kin}/\Delta\Delta G_{eq}$ where $\Delta\Delta G_{kin}$ is the change of activation free energy upon mutation (derived from transition-state assumption) and $\Delta\Delta G_{eq}$ is change on stability upon mutation. The parameter ϕ reflects the degree of participation of a mutated residue in the transition state: when a residue participates in the transition state $\phi = 1$ and $\phi = 0$ otherwise.

The protein engineering analysis of the transition state for folding has been made for cd2 (immunoglobulin fold) (Lorch *et al*., 1999), CheY (Rossman fold) (Lopez-Hernandez & Serrano, 1996), and ADA2H (alpha/beta plait) (Villegas *et al*., 1998). In all cases the results are consistent with the CoC analysis, pointing out that residues that exhibit strongest CoC belong to the folding nucleus as judged by high ϕ-values. In some cases (specifically ADA2H) a slight discrepancy is observed: a shift of one β-register (two residues) of highest-ϕ position from the highest CoC (for the third β-strand) position. This discrepancy may be due to possible uncertainties in structural alignment with respect to register shifts. Another, perhaps more important reason, for such small discrepancies is that nucleus always represents a cluster of residues. In other words, if a pair of residues forms a contact if often induces also a contact between their neighbors along the chain. This makes the nucleus boundaries somewhat fuzzy so that in any particular protein family the evolutionary pressure towards fast folding could have been applied to slightly different amino acids from the nucleating cluster (e.g. for ''historical'' reasons) that may give rise to a slight variation of the exact nucleus location (amino acids with highest ϕ-values) between analogs. In this sense it may be more meaningful to compare the location of the nucleus positions between analogs (and with CoC signal) with respect to their location on different elements of secondary structure. To this end the CoC analysis give very accurate predictions in all cases where experimental information is available.

The ϕ-value analysis often returns fractional values (Itzhaki *et al*., 1995; Martinez *et al*., 1998). This may mean that contacts in the transition state are weaker than in the native state or that each contact that an amino acid makes in the transition state is as strong as in the native state but only a fraction of contacts that an amino acid forms in the native state are actually formed in the transition state. The latter explanation is more plausible in view of recent results of Serrano and co-workers who showed that ϕ-values are robust with respect to change in solvent conditions (e.g. pH; Martinez *et al*., 1998). In this case even for nucleus residues that show strong CoC one can expect only fractional ϕ-values. We expect this to be especially the case for deeply buried nucleus clusters like in Ig fold. In this case each high CoC amino acid makes

many contacts and not all of them are nucleus ones. An experimental way to address this issue is through multiple mutant cycles that address specific interactions. A potential difficulty in carrying out this analysis is that some mutations would have low effect on stability $\Delta\Delta G_{eq}$ resulting in big uncertainty in ϕ-values (Gutin *et al*., 1998).

Evaluating the statistical errors

Finally we would like to comment on some mathematical details of our analysis. The decision of what to consider residues with high CoC depends on a choice of cutoff probability P_c so that when $P(l) < P_c$, the position l is attributed statistically significant CoC. Clearly some freedom in choice of P_c can potentially introduce some arbitrariness in identification of high CoC positions. A possible way to quantitatively estimate possible errors originating from the choice of P_c is to evaluate the probability of ''false positives''. We consider as a false positive the positions that do not have any special nucleation-related or other significant CoC (i.e. whose level of conservatism can be entirely explained by their solvent assessibility) but apparently showing some ''signal'' $P(l) < P_c$, due to a statistical fluctuation. Since $P_c \ll 1$ the probability of false positives follows the rare event statistics that is described by Poisson distribution (Feller, 1970):

$$p_n(\xi) = e^{-\xi}\frac{\xi^n}{n!}$$

where p_n is the probability that n false positives are reported, and $\xi = P_c L$ where L is the length of a sequence. Specifically, probability that no false positives are reported is $p_0 = e^{-\xi}$. P_c cannot also be chosen too low, since in that case no position will be identified as having significant CoC. The choice of P_c is outlined by shaded lines in Figure 4, 6, 8, 10 and 12. It can be seen that probability of a false positive $1 - p_0$ is very low for immunoglobulin, Rossman and TIM barrel folds and is substantial $(0.50 - 0.7)$ for OB fold and alpha/beta plaits. In the latter two cases two or three positions identified as having significant CoC in OB fold proteins and alpha/beta plaits are likely to be false positive, i.e. they may not belong to folding nucleus or be functionally relevant.

Conclusion

Here, we provided a detailed statistical analysis of molecular evolution of most common protein folds. Our results clearly point out that physical factors related to protein folding such as stability and folding rate have undergone considerable evolutionary optimization. In particular we presented a direct evidence for evolutionary pressure towards fast (but not necessarily the fastest) folding for several proteins.

One of the most striking discoveries that emerge from growing data on protein folding kinetics is that proteins that have similar structures and comparable stabilities may fold *via* a two-state mechanism with rates that differ as much as four orders of magnitude (Jackson, 1998). The only model of transition state that is consistent with this observation is of specific nucleus (Abkevich *et al*., 1994; Itzhaki *et al*., 1995; Shakhnovich, 1997; Pande *et al*., 1998; Martinez *et al*., 1998) that points out that there exist particular nucleus positions in the structure that serve as ''accelerator pedals'' for folding. Stronger or weaker evolutionary pressure on those ''accelerator pedals'' (i.e. variation in the nucleus stability) in different proteins gives rise to substantial variation in folding rates. While a first glance comparison of sequences of slow and fast folders does not reveal any striking differences between them, a deeper analysis that compares interactions between amino acid residues at nucleus positions in different analogs provides a possible physical and evolutionary rationale for the surprisingly broad range in which folding rate of analogs may vary. This also suggest an exciting experimental way to control folding rate by ''transplanting'' nucleus of a fast folding protein into its slow folding analogs. Alpha/beta plait and OB fold proteins seem to be the best candidates for such protein surgery.

Finally, we identified two cases where conserved properties of a fold are linked to functionally important locations-super-sites (Rossman fold and TIM barrel). For proteins having these folds prediction of function from structure (and ultimately from sequence) may be a feasible goal.

Methods

Control for solvent accessibility

If $f(s|a)$ is the probability density function (pdf) of entropy s given accessibility a, (normalized $\int_0^{\log(6)} f(s|a)\mathrm{d}s = 1$ for $\forall a$) we can compute the pdf of $S(l)$ based on the H0. Assuming families are independent (see a note below) we can apply central limit theorem (CLT) to compute the pdf of $\bar{S}(l)$. Since $S(l)$ is a sum of large number of independent random variables $s^m(l)$. Hence, according to the CLT $S(l)$ has Gaussian distribution with the mean and the variance:

$$\bar{S}(l) = \frac{1}{M}\sum_{m=1}^{M}\bar{s}(a^m(l));$$

$$\sigma_S^2(l) = \frac{1}{M^2}\sum_{m=1}^{M}\sigma_s^2(a^m(l))$$

where $\bar{s}(a) = \int f(s|a)s\mathrm{d}s$ and $\sigma_s^2(a = \overline{s^2} - \bar{s}^2)$ are the mean and the variance of the in-family entropy as a function of accessibility.

The probability to observe $S^{obs}(l) < S$ by chance:

$$P(S^{\mathrm{obs}}(l) < S) = \int_{-\infty}^{S^{\mathrm{obs}}(l)-\bar{S}(l)/\sigma_S(l)} \exp\left(-\frac{x^2}{2}\right)\mathrm{d}x$$

Positions which exhibit $P(S < S(l)) < P_c$ are said to have significant CoC. The threshold value P_c depends on

the amount of data available for a given fold. In our analysis it varies between 10^{-2} (OB fold) to 10^{-5} (Rossman fold).

Probability $P(S)$ can also be computed using a convolution. CLT however makes computations easier and faster. Importantly we assumed that families are independent, i.e. conservatism in one family does not change the probability to observe the same position conserved in the other family. This assumption is motivated by the choice of families that are distant enough in sequences ($ID < 25\,\%$).

Solvent accessibility was taken from HSSP files (Dodge *et al.*, 1998) where it is computed as the solvated residues surface area in Å^2 (number of contacting water molecules $\times 10$). To compute $P(s|a)$ we quantified accessibility into intervals of 1 Å^2. Then smoothed $P(s|a)$ with a window of width 31 for all intervals where less then town counts where observed in the PDB dataset.

Selection of representative proteins

All our results were obtained using sequence alignments from HSSP (Dodge *et al.*, 1998) and structural alignments (see below) from FSSP (Holm & Sander, 1993). Representative set of proteins from FSSP, Sep98 release was used as a basic set. From this set we removed all proteins which have sequence identity $ID > 25\,\%$ with any other protein in the set. This allowed us to eliminate some obvious homologs. Next we excluded from our analysis all the families where all positions are conserved, i.e., where $1/L\Sigma_{l=1}^{L} s^m(l) > 0.4$.

In this study we used sequence entropy $s(l)$ as a measure of evolutionary conservation in a protein family. However, multiple sequence alignments used to derive the sequence entropy can be biased in various ways and may not represent the divergent evolution of the family. For example, closely homologous sequences can be over-represented in a multiple sequence alignment and dominate over a few distant homologs with high variability. This effect was partially accounted for in our calculations by exclusion of families where all amino acid residues are conserved as such families clearly represent insufficiently divergent sequences. By weighting sequences in a multiple alignment one can compensate for this bias (for a review, see Henikoff & Henikoff, 1994). Lack of weighting, however, does not affect our results. The control by solvent accessibility shows that a simple sequence entropy can be predicted from solvent accessibility with a great accuracy when both are averaged over several families of the same fold. Hence, biases from individual families are ''averaged out'' over very large number of families that we used. We plan to introduce sequence weighting in the future, which can make our method more sensitive and less data demanding.

Structural alignments

Some structural alignments in the FSSP were corrected using Monte Carlo alignment algorithm (Mirny & Shakhnovich, 1998). In brief, each of the two aligned structures are represented by a distance matrix. We use the same similarity measure as Holm & Sander (1993) in their Dali program. Structural alignment is obtained by optimization of this function. In contrast to Holm and Sander, we optimized this function using Monte Carlo alignment which gives systematically higher scores than optimization protocol implemented in Dali (Mirny & Shakhnovich, 1998). Final alignments, however, are not very different from those obtained by Dali. This refinement of structural alignments was applied primarily to alpha/beta plait where low degree of structural similarity made structural alignments a complicated problem. However, one should bare in mind that ambiguities in structural alignments are inevitable since the results depend mostly on the choice of the similarity score. Two structures were considered similar if they have FSSP Z-score $Z_{FSSP} > 2.5$ and $DRMS_{C\beta} < 6$ Å.

Treating gaps in alignments

Positions with gaps in structural alignments were neglected in computations of $S(l)$, $S^{across}(l)$ and $P(S)$. Therefore summation in equation (1) is over those families m which do not have a gap in structural alignment in position l. Taking this into account equation (1) turns into:

$$S(l) = \frac{\Sigma_{m=1}^{M} s^m(l)\delta_l^m}{\Sigma_{m=1}^{M} \delta_l^m}$$

where $\delta_l^m = 0$ if position l in family m has a gap in structural alignment.

This treatment of gaps however leads to a problem when all except a few families have gaps at position l. Then $S(l)$ and $P(l)$ fluctuate and are unreliable. To avoid this problem we deleted from our analysis fragments of a protein fold where more than 50 % of structural alignments have gaps.

Acknowledgments

This work is supported by NIH grant RO1 GM52126. We are grateful to Fabrizio Chiti, Jane Clarke, Chris Dobson, Alan Fersht, Stephen Hamil, Mikael Oliveberg and Luis Serrano for illuminating discussions of experimental results and making many of them available to us prior to publication.

After this paper had been completed we heard sad news that Oleg Ptitsyn passed away. Oleg had been very excited about emerging understanding of deep relation between protein folding and evolution, which is the main topic of the present paper. In fact Oleg's last paper is entirely devoted to this subject (Ptitsyn, 1998). E.I.S. enormously benefited from his insights over many years of our close collaboration and friendship.

References

Abkevich, V., Gutin, A. & Shakhnovich, E. (1994). Specific nucleus as the transition state for protein folding: evidence from the lattice model. *Biochemistry*, **33**, 10026-10036.

Altschuh, D., Vernet, T., Berti, P., Moras, D. & Nagai, K. (1988). Coordinated amino acid changes in homologous protein families. *Protein Eng.* **2**, 193-199.

Bahar, I. & Jernigan, R. (1997). Inter-residue potentials in globular proteins and the dominance of highly specific hydrophilic interactions at close separation. *J. Mol. Biol.* **266**, 195-214.

Bellsolell, L., Cronet, P., Majolero, M., Serrano, L. & Coll, M. (1996). The three-dimensional structure of two mutants of the signal transduction protein chey

suggest its molecular activation mechanism. *J. Mol. Biol.* **257**, 116-128.
Bork, P., Holm, L. & Sander, C. (1994). The immunoglobulin fold. Structural classification, sequence patterns and common core. *J. Mol. Biol.* **242**, 309-320.
Branden, C. & Tooze, J. (1998). *Introduction to Protein Structure*, Garland Publishing, Inc., New York.
Bryngelson, J. & Wolynes, P. (1990). A simple statistical field theory of heteropolymer collapse with application to protein folding. *Biopolymers,* **30**, 177-188.
Dodge, C., Schneider, R. & Sander, C. (1998). The hssp database of protein structure-sequence alignments and family profiles. *Nucl. Acids Res.* **26**, 313-315.
Feller, W. (1970). *An Introduction to Probability Theory and its Applications*, Wiley, New York.
Gilis, D. & Rooman, M. (1997). Predicting protein stability changes upon mutation using database-derived potentials: solvent accessibility determines the importance of local *versus* non-local interactions along the sequence. *J. Mol. Biol.* **272**, 276-290.
Goldstein, R., Luthey-Schulten, Z. & Wolynes, P. (1992). Optimal protein-folding codes from spin-glass theory. *Proc. Natl Acad. Sci. USA,* **89**, 4918-4922.
Govindarajan, S. & Goldstein, R. (1995). Why are some protein structures so common?. *Proc. Natl Acad. Sci. USA,* **93**, 3341-3345.
Guo, Z. & Thirumalai, D. (1995). Nucleation mechanism for protein folding and theoretical predictions for hydrogen-exchange labelling experiments. *Biopolymers,* **35**, 137-139.
Gutin, A., Abkevich, V. & Shakhnovich, E. (1998). A protein engineering analysis of the transition state for protein folding: simulation in the lattice model. *Fold. Design*, **3**, 183-194.
Hamill, S., Meekhof, A. & Clarke, J. (1998). The effect of boundary selection on the stability and folding of the third fibronectin type iii domain from human tenascin. *Biochemistry,* **37**, 8071-8079.
Hao, M.-H. & Scheraga, H. (1994a). Monte-Carlo simulation of a first order transition for protein folding. *J. Phys. Chem.* **98**, 4940-4945.
Hao, M.-H. & Scheraga, H. (1994b). Statistical thermodynamics of protein folding: sequence dependence. *J. Phys. Chem.* **98**, 9882-9886.
Henikoff, S. & Henikoff, J. (1994). Position-based sequence weights. *J. Mol. Biol.* **243**, 574-578.
Holm, L. & Sander, C. (1993). Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* **233**, 123-138.
Itzhaki, L., Otzen, D. & Fersht, A. (1995). The structure of the transition state for folding of chymotrypsin inhibitor 2 analyzed by protein engineering methods: evidence for a nucleation-condensation mechanism for protein folding. *J. Mol. Biol.* **254**, 260-288.
Jackson, S. (1998). How do small single-domain proteins fold? *Fold. Design,* **3**, R81-R91.
Koshi, J. & Goldstein, R. (1997). Mutation matrices and physical-chemical properties: correlations and implications. *Proteins: Struct. Funct. Genet.* **27**, 336-344.
Kraulis, P. (1991). Molscript: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallog.* **24**, 946-950.
Liaw, S., Jun, G. & Eisenberg, D. (1994). Interactions of nucleotides with fully unadenylylated glutamine synthetase from salmonella typhimurium. *Biochemistry,* **33**, 11184-11188.
Lopez-Hernandez, E. & Serrano, L. (1996). Structure of the transition state for folding of the 129 aa protein chey resembles that of a smaller protein, ci2. *Fold. Design,* **1**, 43-55.
Lorch, M., Mason, J., Clarke, A. & Parker, M. (1999). Effects of core mutations on the folding of a beta-sheet protein: implications for backbone organization in the i-state. *Biochemistry,* **38**, 1377-1385.
Martinez, J., Pissabarro, T. & Serrano, L. (1998). Obligatory steps in protein folding and the conformational diversity of the transition state. *Nature Struct. Biol.* **5**, 721-729.
Meekhof, A. & Freund, S. (1999). Probing residual structure and backbone dynamics on the milli- to picosecond timescale in a urea-denatured fibronectin type III domain. *J. Mol. Biol.* **286**, 579-592.
Merritt, E. & Bacon, D. (1997). Raster3D: phosorealistic molecular graphics. *Methods Enzymol.* **277**, 505-524.
Mirny, L. & Shakhnovich, E. (1998). Protein structure prediction by threading. Why it works and why it does not. *J. Mol. Biol.* **283**, 507-526.
Mirny, L., Abkevich, V. & Shakhnovich, E. (1998). How evolution makes proteins fold quickly. *Proc. Natl Acad. Sci. USA,* **95**, 4976-4981.
Newkirk, K., Feng, W., Jiang, W., Tejero, R., Emerson, S., Inouye, M. & Montelione, G. (1994). Solution nmr structure of the major cold shock protein (cspa) from *Escherichia coli*: identification of a binding epitope for DNA. *Proc. Natl Acad. Sci. USA,* **91**, 5114-5118.
Pande, V., Grosberg, A. & Tanaka, T. (1995). How accurate must potentials be for successful modeling of protein folding?. *J. Chem. Phys.* **103**, 1-10.
Pande, V., Grosberg, A., Rokshar, D. & Tanaka, T. (1998). Pathways for protein folding: is a ''new view'' needed? *Curr. Opin. Struct. Biol.* **8**, 68-79.
Parker, M., Dempsey, C., Hosszu, L., Waltho, J. & Clarke, A. (1998). Topology, sequence evolution and folding dynamics of an immunoglobulin domain. *Nature Struct. Biol.* **5**, 194-198.
Pelletier, H., Sawaya, M., Wolfle, W., Wilson, S. & Kraut, J. (1996). Crystal structures of human DNA polymerase beta complexed with DNA: implications for catalytic mechanism, processivity, and fidelity. *Biochemistry,* **35**, 12742-12761.
Perl, D., Welker, C., Schindler, T., Schroder, K., Marahiel, M., Jaenicke, R. & Schmid, F. (1998). Conservation of rapid two-state folding in mesophilic, thermophilic and hyperthermophilic cold shock proteins. *Nature Struct. Biol.* **5**, 229-235.
Plaxco, K., Soitzfaden, C., Campbell, I. & Dobson, C. (1996). Rapid refolding of a proline-rich all β-sheet fibronectin type III module. *Proc. Natl Acad. Sci. USA,* **93**, 10703-10706.
Ptitsyn, O. (1998). Protein folding and protein evolution: common folding nucleus in different subfamilies of *c*-type cytochromes?. *J. Mol. Biol.* **278**, 655-666.
Reid, K., Rodriguez, H., Hillier, B. & Gregoret, L. (1998). Stability and folding properties of a model beta-sheet protein, *Escherichia coli* CSPA [published erratum appears in protein Sci 1998]. *Protein Sci.* **7**, 470-479.
Russell, R. & Ponting, C. (1998). Protein fold irregularities that hinder sequence analysis. *Curr. Opin. Struct. Biol.* **8**, 364-371.
Russell, R., Sasieni, P. & Sternberg, M. (1998a). Supersites within superfolds. Binding site similarity in the absence of homology. *J. Mol. Biol.* **282**, 903-918.
Russell, R., Sasieni, P. & Sternberg, M. (1998b). Supersites within superfolds. Binding site similarity in the absence of homology. *J. Mol. Biol.* **282**, 903-918.

Sali, A., Shakhnovich, E. & Karplus, M. (1994). Kinetics of protein folding. A lattice model study for the requirements for folding to the native state. *J. Mol. Biol.* **235**, 1614-1636.

Schindelin, H., Jiang, W., Inouye, M. & Heinemann, U. (1994). Crystal structure of cspa, the major cold shock protein of Escherichia coli. *Proc. Natl Acad. Sci. USA,* **91**, 5119-5123.

Schindler, T., Perl, D., Graumann, P., Sieber, V., Marahiel, M. & Schmid, F. (1998). Surface-exposed phenylalanines in the rnp1/rnp2 motif stabilize the cold-shock protein cspb from *Bacillus subtilis*. *Proteins: Struct. Funct. Genet.* **30**, 401-406.

Schindler, T., Graumann, P., Perl, D., Ma, S., Schmid, F. & Marahiel, M. (1999). The family of cold shock proteins of *Bacillus subtilis*. Stability and dynamics *in vitro* and *in vivo*. *J. Biol. Chem.* **274**, 3407-3413.

Schuller, D., Grant, G. & Banaszak, L. (1995). The allosteric ligand site in the vmax-type cooperative enzyme phosphoglycerate dehydrogenase. *Nature Struct. Biol.* **2**, 69-76.

Shakhnovich, E. (1994). Proteins with selected sequences fold to their unique native conformation. *Phys. Rev. Letters,* **72**, 3907-3910.

Shakhnovich, E. (1997). Theoretical studies of protein-folding thermodynamics and kinetics. *Curr. Opin. Struct. Biol.* **7**, 29-40.

Shakhnovich, E. (1998a). Folding nucleus: specific of multiple? Insights from simulations and comparison with experiment. *Fold. Design,* **3**, R108-R111.

Shakhnovich, E. (1998b). Protein design: a perspective from simple tractable models. *Fold. Design,* **3**, R45-R58.

Shakhnovich, E. & Gutin, A. (1993a). Engineering of stable and fast-folding sequences of model proteins. *Proc. Natl Acad. Sci. USA,* **90**, 7195-7199.

Shakhnovich, E. & Gutin, A. (1993b). A novel approach to design of stable proteins. *Protein Eng.* **6**, 793-800.

Shakhnovich, E., Abkevich, V. & Ptitsyn, O. (1996). Conserved residues and the mechanism of protein folding. *Nature,* **379**, 96-98.

Thomas, D., Casari, G. & Sander, C. (1996). The prediction of protein contacts from multiple sequence alignments. *Protein Eng.* **9**, 941-948.

Tiana, G., Broglia, R., Roman, H., Vigezzi, E. & Shakhnovich, E. (1998). Folding and misfolding of designed protein like chains with mutations. *J. Chem. Phys.* **108**, 757-761.

van Nuland, H., Chiti, F., Taddei, N., Raugei, G., Ramponi, G. & Dobson, C. (1998a). Slow folding of muscle acylphosphatase in the absence of intermediates. *J. Mol. Biol.* **283**, 883-891.

van Nuland, H., Meijberg, W., Warner, J., Forge, V., Scheek, R., Robillard, G. & Dobson, C. (1998b). Slow cooperative folding of a small globular protein hpr. *Biochemistry,* **37**, 622-637.

Viguera, A., Serrano, L. & Wilmanns, M. (1997). Different folding transition states may result in the same native structure. *Nature Struct. Biol.* **4**, 939-946.

Villegas, V., Martinez, J., Aviles, F. & Serrano, L. (1998). Structure of the transition state in the folding process of human procarboxypeptidase a2 activation domain. *J. Mol. Biol.* **283**, 1027-1036.

Welch, M., Oosawa, K., Aizawa, S. & Eisenbach, M. (1994). Effects of phosphorylation, Mg^{2+}, and conformation of the chemotaxis protein chey on its binding to the flagellar switch protein flim. *Biochemistry,* **33**, 10470-10476.

Wilcock, D., Pisabarro, M., Lopez-Hernandez, E., Serrano, L. & Coll, M. (1998). Structure analysis of two chey mutants: importance of the hydrogen-bond contribution to protein stability. *Acta Crystallog. sect. D,* **54**, 378-385.

Wilson, K., Shewchuk, L., Brennan, R., Otsuka, A. & Matthews, B. (1992). *Escherichia coli* biotin holoenzyme synthetase/bio repressor crystal structure delineates the biotin- and DNA-binding domains. *Proc. Natl Acad. Sci. USA,* **89**, 9257-9261.

Edited by A. R. Fersht

(Received 26 March 1999; received in revised form 21 May 1999; accepted 27 May 1999)