A strategy for detecting the conservation of folding-nucleus residues in protein superfamilies

Stephen W Michnick[1] and Eugene Shakhnovich[2]

Background: Nucleation-growth theory predicts that fast-folding peptide sequences fold to their native structure via structures in a transition-state ensemble that share a small number of native contacts (the folding nucleus). Experimental and theoretical studies of proteins suggest that residues participating in folding nuclei are conserved among homologs. We attempted to determine if this is true in proteins with highly diverged sequences but identical folds (superfamilies).

Results: We describe a strategy based on comparisons of residue conservation in natural superfamily sequences with simulated sequences (generated with a Monte-Carlo sequence design strategy) for the same proteins. The basic assumptions of the strategy were that natural sequences will conserve residues needed for folding and stability plus function, the simulated sequences contain no functional conservation, and nucleus residues make native contacts with each other. Based on these assumptions, we identified seven potential nucleus residues in ubiquitin superfamily members. Non-nucleus conserved residues were also identified; these are proposed to be involved in stabilizing native interactions. We found that all superfamily members conserved the same potential nucleus residue positions, except those for which the structural topology is significantly different.

Conclusions: Our results suggest that the conservation of the nucleus of a specific fold can be predicted by comparing designed simulated sequences with natural highly diverged sequences that fold to the same structure. We suggest that such a strategy could be used to help plan protein folding and design experiments, to identify new superfamily members, and to subdivide superfamilies further into classes having a similar folding mechanism.

Addresses: [1]Département de biochime, Université de Montréal, C.P. 6128, Succursalle Centre-ville, Montréal, Quebec, Canada H3C 3J7. [2]Department of Chemistry, Harvard University, 12 Oxford Street, Cambridge, MA 61328, USA.

Correspondence: Stephen W Michnick
E-mail: michnick@bch.umontreal.ca

Key words: folding nucleus, protein folding, protein superfamilies, sequence design

Received: **01 December 1997**
Revisions requested: **14 January 1998**
Revisions received: **30 March 1998**
Accepted: **03 April 1998**

Published: **26 May 1998**
http://biomednet.com/elecref/1359027800300239

Folding & Design 26 May 1998, **3:**239–251

Introduction

Two major challenges in structural biology are the development and the testing of theories towards explaining kinetic and equilibrium aspects of protein folding and the prediction of three-dimensional structure from primary peptide sequences. A particularly perplexing problem is that of how highly divergent sequences fold to identical structures (superfamilies). Such cases occur at the point at which amino acid identity between sequences that fold to the same structure falls to or below 15% [1,2]. Various strategies that use secondary structure prediction and 'threading' of these predicted regions of structure to known structures can be successful for predicting three-dimensional structures in limited cases, particularly when combined with sequence alignments and other structural or functional information [3–6]. Such methods can obviously fail, however, if the secondary structure prediction is inaccurate or when a sequence is assigned to a structure having identical stretches of secondary structure elements and even having similar topology, but an incorrect overall fold. Also, it is not clear whether or not these methods can provide any insight into what elements of the primary sequences of two divergent proteins are important for specifying a unique fold.

Recent theoretical studies and protein engineering experiments have shown that small proteins that fold via simple two-state kinetics may form a transition-state 'nucleus' in which specific residues form native contacts; folding to the native structure then follows in a cooperative manner from this nucleus [7–16]. This 'nucleation-growth' or 'nucleation-condensation' model for protein folding can predict both equilibrium and kinetic aspects of folding. It can also explain how it is that the integrity of certain positions in the sequence of a protein are crucial for fast folding of sequences (i.e. nucleus residues) whereas others are not [17–19]. A particularly important implication of nucleation-growth theory is that residues that participate in the transition-state nucleus are not necessarily associated with specific secondary structure elements nor are they necessarily of a particular type, such as hydrophobic residues that contribute to hydrophobic cores (which must be distinguished from the transition-state nucleus). Nucleation-growth theory would

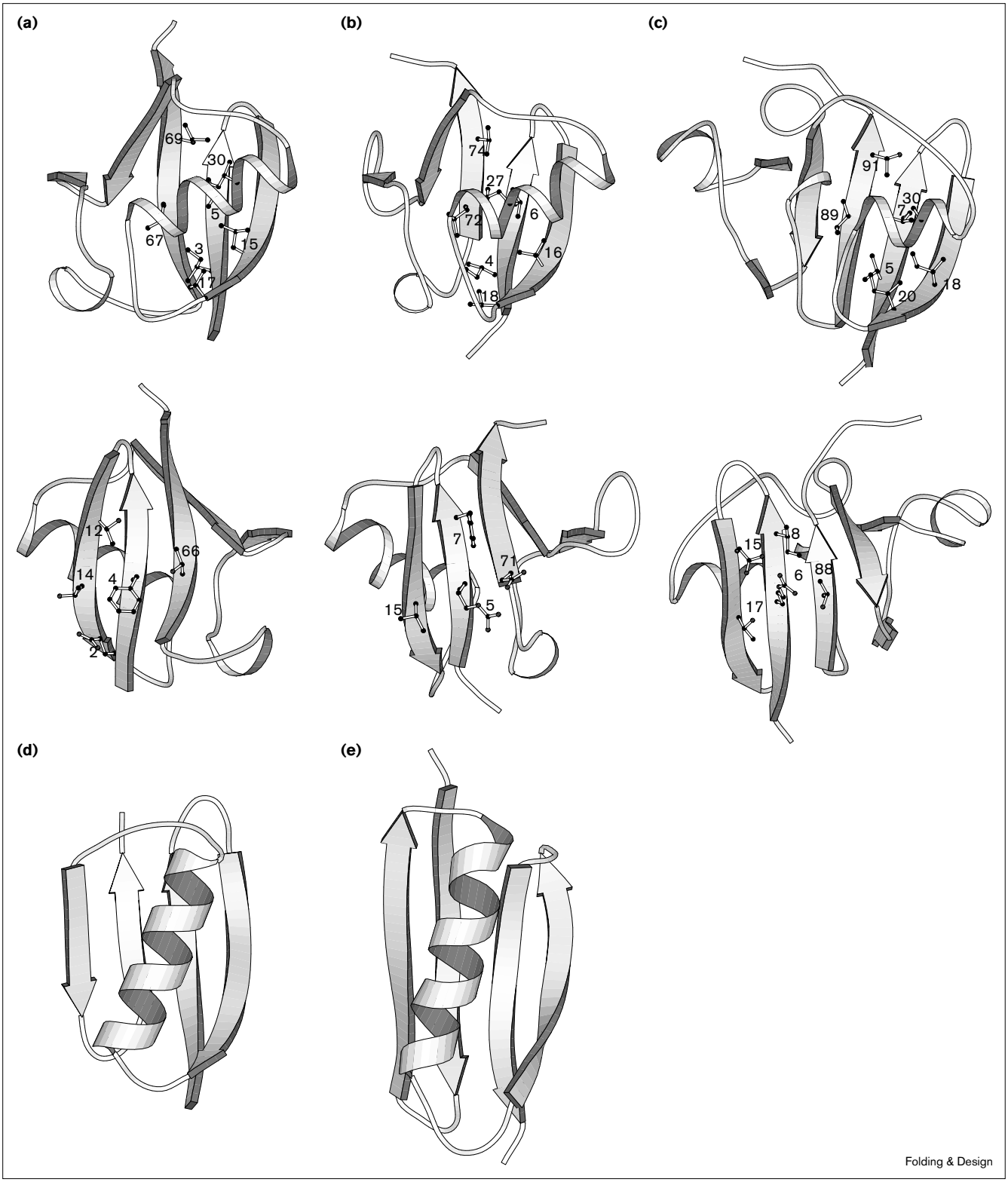
predict that the conservation of folds through evolution requires the maintenance of strong contacts between nucleus-forming residues, which may result in the conservation of these residues within each family; indeed this has turned out to be the case for a well-studied example of a protein that folds through a transition-state nucleus, chymotrypsin inhibitor 2 (CI2; [12,20–23]). The transition state of CI2 has been studied extensively using protein engineering experiments and lattice and off-lattice computer simulations of designed sequences [12]. The protein engineering studies identified key residues involved in the transition-state nucleus and the simulations resulted in predictions of the key nucleus-forming residues identified in the experiments. Furthermore, it was shown that the conservation of nucleus residues in designed sequences also correlated with the conservation of residue identity or class in 23 sequences of naturally occurring CI2s. Other examples showing similar correlations have been observed for CheY (E. Shahknovich, unpublished observations; [18]) and acyl-coenzyme binding protein (F. Poulsen, personal communication). If it could be demonstrated that this correlation can be extended to other proteins and specifically to protein superfamilies, it may be possible to develop a strategy to identify new superfamily members, to define other superfamilies and perhaps aid the design of proteins with specific folds. As a step in this direction, we have chosen to study the sequence design of a superfamily of small monomeric proteins that include ubiquitin and the p21 ras binding domain (RA) of the serine/threonine kinase raf. We chose these proteins because they have identical folds (Figure 1a,b; 1.46 Å root mean square deviation (rmsd) for backbone atoms of aligned residues), yet have only 10% sequence identity (Figure 2). The structures of both proteins have been determined by both X-ray crystallography [24,25] and NMR spectroscopy [26–28], whereas the folding of ubiquitin has been studied extensively [29–36]. Ubiquitin is a small (76 amino acids) very hydrophobic protein. It was shown that the formation of an intermediate in the folding of ubiquitin is temperature dependent [32]. Specifically, they showed that at 8°C ubiquitin folds via a two-state mechanism whereas at a higher temperature (25°C) it folds with a detectable early intermediate. Mutations were found, however, that resulted in apparent two-state kinetics of folding [33]. Although the existence of intermediates may potentially complicate the analysis of the transition state for folding, arguments have been presented [37] that show that in this case nucleation contacts may be partially formed in the early intermediate and hence the analysis of nucleation sites may reveal some structural features of burst intermediates [38].

The functions of the raf RA domain and ubiquitin are completely different, except that they both provide recognition sites for association with other proteins. In the case of ubiquitin, its function is to become covalently attached via its C terminus to proteins destined for proteolysis by proteosomes [39–41], whereas the raf RA domain binds to the GTP-loaded form of the oncogene GTPase ras in a process involving translocation of raf to the plasma membrane of cells and activation of the kinase activity by phosphorylation [42,43]. There is extensive sequence information about ubiquitin and RA domains of other proteins. 33 sequences of ubiquitins from various species have been identified, as have putative RA domains of 23 proteins [5]. Of particular note in these studies is that the sequence identity among different RA domains is very low (17%). These results demonstrate that even among small proteins that are likely to have identical function and structure, there can be considerable divergence in sequences. Nevertheless, we immediately noticed a strong coincidence in residue positions that are highly conserved among these RA domains, raf RA, ubiquitin, and another superfamily member, ferredoxin 1. We propose that among these conserved sites, some may be involved in the formation of a folding nucleus. Here, we discuss a strategy to identify conserved residues that may contribute to the folding nucleus based on comparisons of protein superfamily sequences with simulated designed sequences.

Results

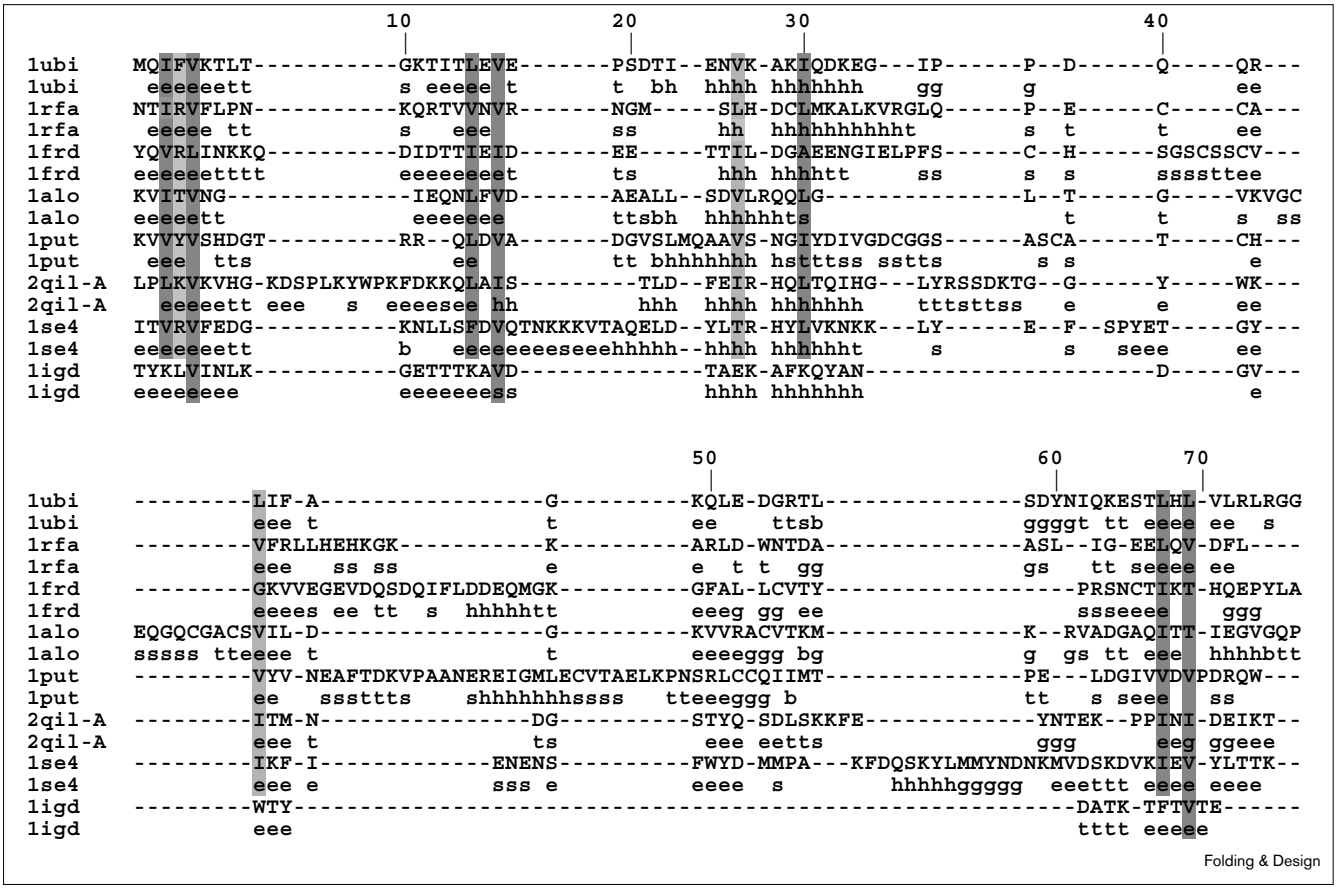
We began with the hypothesis that some fast-folding proteins fold via a transition-state nucleus consisting of residues that make native contacts in the folding transition state and form a unique structure around which the rest of the peptide folds to the unique native structure. We reasoned that proteins with identical folds, but highly diverged sequences (superfamilies) must nevertheless retain identity or near identity at sequence positions that participate in the nucleus. The goal then was to identify such positions by aligning the sequences of known members of a protein superfamily with related, but highly divergent, sequences of other proteins (those showing average sequence identities in the range 5–20%). It might seem more reasonable to use sequences of superfamily members that are closer in homology (> 30%) to the test sequences. The problem with this approach, as we show below, is that sequences that are too homologous will not reveal the limiting number of highly conserved positions that could represent those that participate in the nucleus. Even if the set of sequences we use to compare with the superfamily members are very different, however, it is also possible that positions would be conserved because of their involvement in stabilizing the native structure (we call these 'design' sites because of their importance in efforts to design proteins) or a common function of the proteins, such as binding or catalysis. Design sites are positions that could be important for, for example, the packing of the hydrophobic core, the formation of surface clusters that might stabilize the native structure, and residues involved in long-range interactions, such as salt bridges. The distinction between the design and the nucleus sites is that mutations in design sites should lead to an increase in the rate of unfolding because the native state is destabilized. Mutations in nucleation sites would simultaneously

Figure 1



Ribbon structures of proteins rendered in MOLSCRIPT [67]. **(a)** Ubiquitin, **(b)** raf and **(c)** ferredoxin. Upper panels show the helix face with potential nucleus residues rendered as sticks; the lower panels show the β sheet face with possible surface cluster residues. **(d)** Streptococcal protein G B1 IgG-binding domain and **(e)** protein L IgG-binding domain.

Figure 2



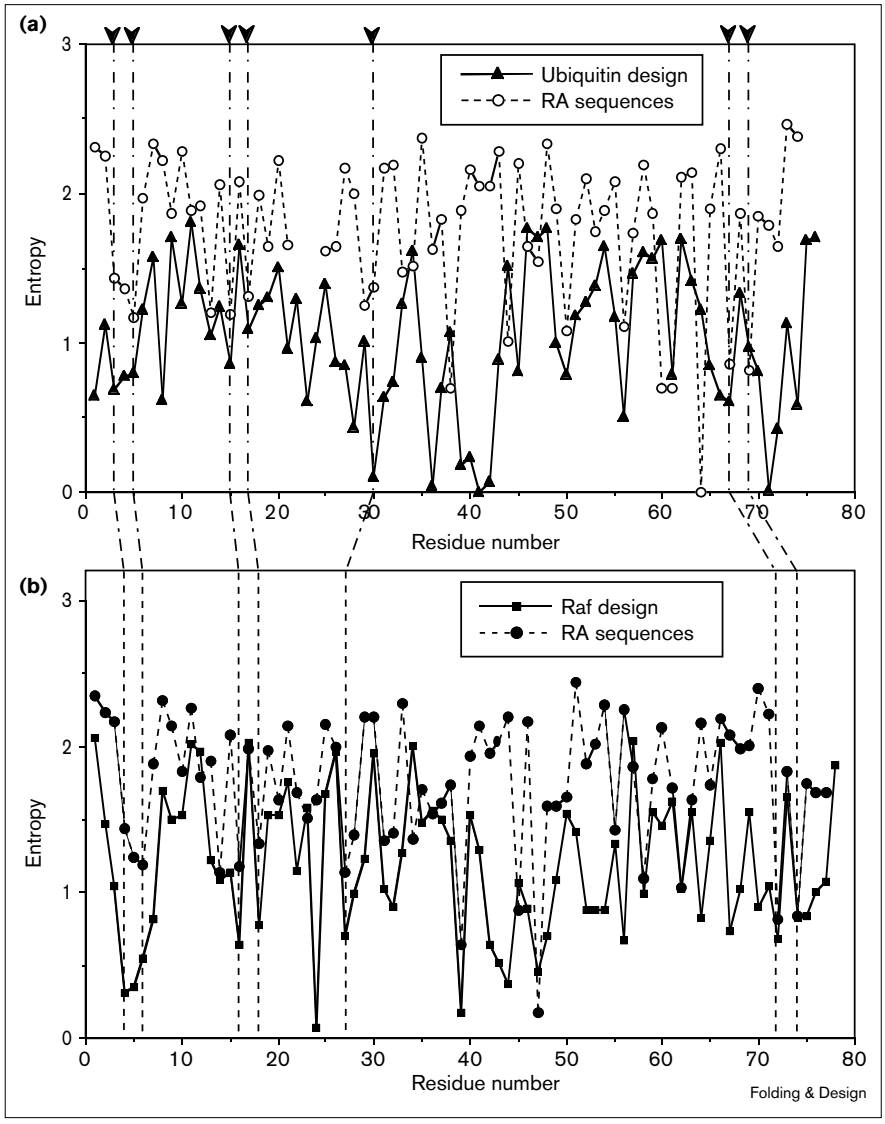
Sequence and secondary structure alignments for ubiquitin superfamily proteins (see the Materials and methods section). Alignments were based on DSSP assignments with minor visual modification as described (see the Materials and methods section). PDB codes for the corresponding proteins are indicated: 1ubi, ubiquitin; 1rfa, raf RA domain; 1frd, ferredoxin 1; 1alo, aldehyde oxidoreductase; 1put, putidaredoxin; 2qil-A, superantigen enterotoxin C2 from *Staphylococcus aureus*; 1se4, enterotoxin B superantigen from *S. aureus*; 1igd, immunoglobulin-binding domain of streptococcal protein G. Dark shaded positions, those with the potential to be residues involved in the folding nucleus; lightly shaded positions, sites where design and natural sequences show low entropy but are not potential nucleation sites. Secondary structure assignment is as in Kabsch and Sander [66]: h, α helix; e, β sheet; t, β turn; s, non-hydrogen-bonded bend structure; g, 3^{10} helix.

decrease both folding and unfolding rates as well as effecting populations of intermediates and destabilizing the native state. This is because mutations of nucleation sites will destabilize the transition state regardless of whether or not the protein is folding or unfolding.

How can we distinguish the conserved design or functional residues from folding-nucleus residues? What is needed is a set of comparison sequences that are neutral to functional conservation. It is not possible to imagine such sequences in nature, but it is possible to simulate them. We did this using the sequence-design strategy [44,45], a stochastic (Monte-Carlo) optimization routine in sequence space in which residues are assigned to positions on the known three-dimensional structure of a protein with a probability determined by the effect of such a substitution on a total residue-contact potential for the entire protein. Comparison of designed and natural sequences of a protein superfamily would allow us to distinguish the conservation of nucleus residues and design residues from those conserved for function. Another condition for nucleation growth is that residues that form the folding nucleus also contact each other in the native structure. Thus, it would be possible to distinguish nucleus residues from design residues based on this criterion.

The specific strategy we used to compare natural and designed sequences was as follows. First, we obtained a set of sequences related to the raf RA domain, but with a low sequence identity with raf RA. We were fortunate that a set of 23 sequences had already been identified [5]. Residue position entropies were calculated for ubiquitin and raf aligned with these sequences, as described in the Materials and methods section. Second, we performed sequence design simulations on ubiquitin and raf and calculated residue entropies at each position according to Equation 2 (see the Materials and methods section). Third, we compared design

Figure 3



A comparison of simulated and natural RA sequence entropies for **(a)** ubiquitin and **(b)** raf. Arrowheads indicate sequence positions that show low entropy simultaneously in simulated and ubiquitin-aligned or raf-aligned RA sequences; stippled lines indicate the corresponding positions in ubiquitin and raf. Broken lines are drawn to indicate differences in alignments for identical positions in the ubiquitin and raf sequences.

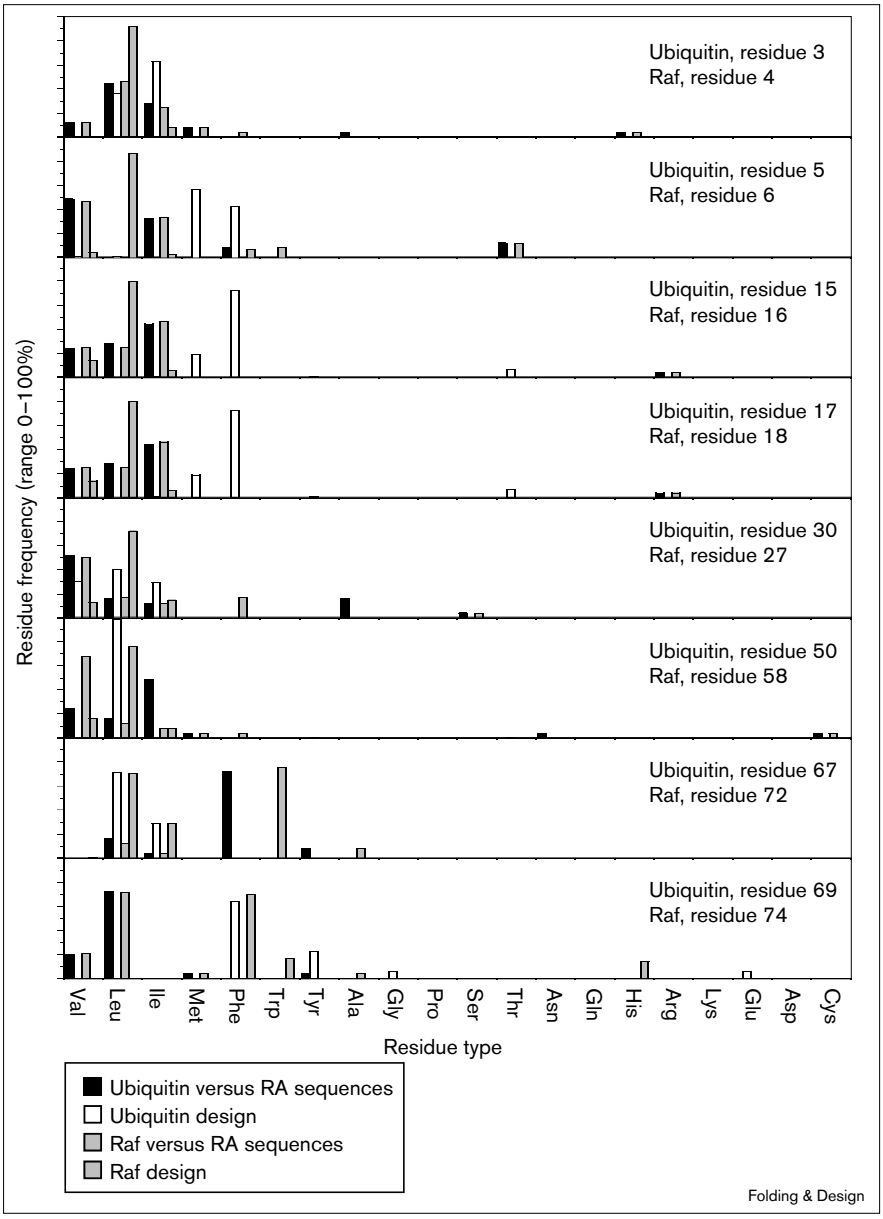
entropy with sequence entropy for ubiquitin sequences to determine the sequence positions that showed low entropies in both cases. A position was defined as having a low entropy if the observed entropy fell below the average entropy observed for all sequences (assuming a Gaussian distribution). We observed that the majority of low-entropy sites lay below one standard deviation unit from the average. Finally, homologous sequence positions in ubiquitin and raf that showed a low entropy in design and RA sequence alignments and could be shown to make contact in the native structures of both proteins were flagged as potential folding-nucleation residues.

Correlation of conservation of residues in designed ubiquitin and raf sequences and RA domain sequences

The sequence alignment of ubiquitin and raf RA (Figure 2) were based on a least-squares fit of the crystal structure of ubiquitin with the NMR-determined structure of the raf RA domain [25,27]. For the NMR structure of RA we chose the first of 30 models deposited in the PDB (PDB code 1rfa) because all these structures already had low backbone rmsds within the secondary structure elements chosen for alignment (the rmsd for backbone Cα atoms was 0.61 Å and for all heavy atoms was 1.1 Å). A fit of the raf RA and ubiquitin structures based on an alignment of residues with identical secondary structure and a visual alignment resulted in an rmsd for Cα atoms of 1.46 Å. This alignment is also consistent with the alignment comparing the crystal structure of Raf RA and ubiquitin [24,25].

The alignment of sequence entropy profiles revealed positions of low entropy in both the designed and the natural RA sequences (Figure 3). Qualitatively, there seemed to be a good correspondence between the observed sequence

Figure 4

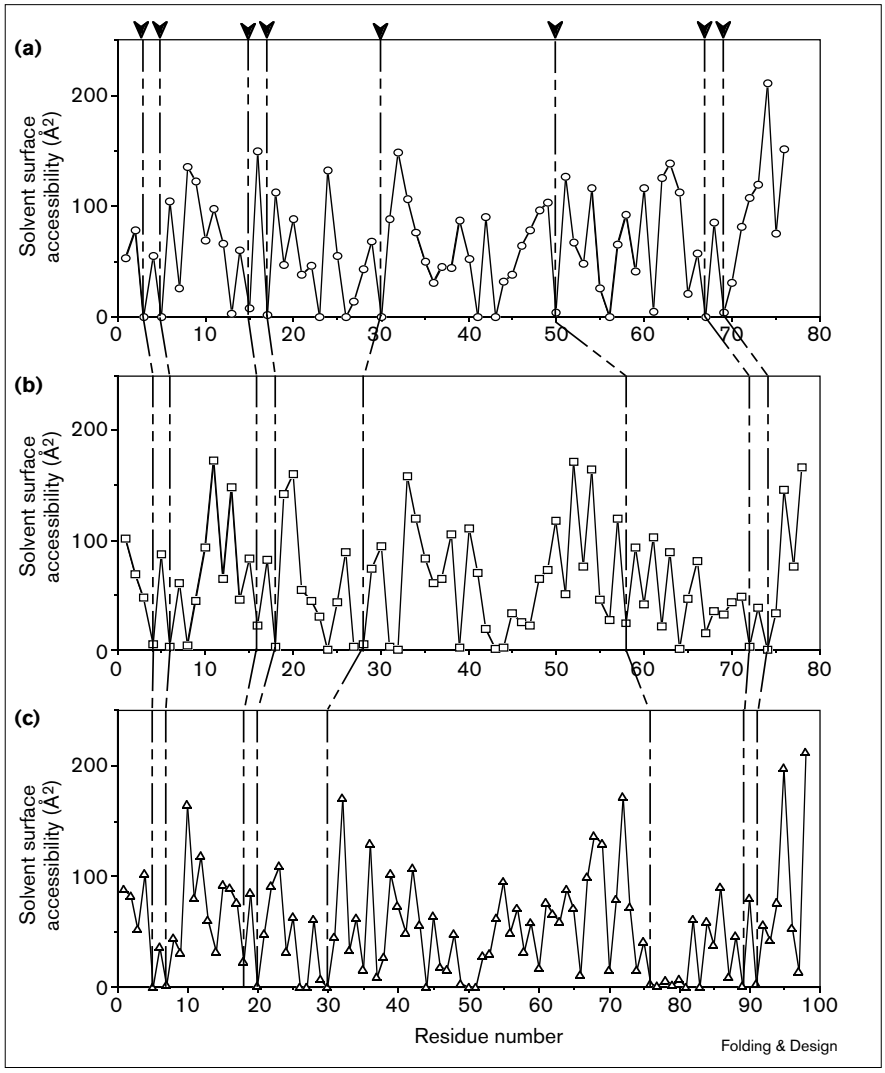


A breakdown of residue-type frequencies in designed, ubiquitin-aligned and raf-aligned sequences.

position entropy in the designed ubiquitin or raf sequences and the natural RA sequences, particularly in the N and C termini. It is important to note, however, that the goal of our analysis was to identify only those sites for which there is low entropy in both designed and natural sequences; it is not the goal to demonstrate a correlation between natural and design sequence entropies at each residue position. At sites for which there was a correlation of low entropies we tested whether or not these met the other criteria to be potential nucleation sites; that is, are they conserved in both raf and ubiquitin and do they make native contacts (i.e. contacts to other residues within 4.5 Å) with at least one of the other conserved sites? Based on these analyses we could identify seven potential sites. These included residues 3, 5, 15, 17, 30, 67 and 69. An analysis of these sites by type and structural environment of the proteins revealed two things. First, all these residues participate in the hydrophobic core of both proteins. Second, the residues are represented by either aliphatic or aromatic residues in the designed or RA sequences, but not necessarily of the same class in both design and RA sequences (Figure 4). Have we simply demonstrated that the hydrophobic core of these proteins is conserved? We think not; if this were true, then we should be able to see a correlation between design and RA sequences for all core residues. Assuming that core residues will be well packed and will therefore have a low surface

Figure 5

Residue solvent accessibilities of **(a)** ubiquitin, **(b)** raf and **(c)** ferredoxin.



accessibility, we examined the correlation for the three proteins (Figure 5). We see that all the potential nucleation residues have among the lowest surface accessibility and are therefore part of the hydrophobic core; these represent less than half of the residues in the core, however (for example, seven out of 16 sites in raf and ubiquitin).

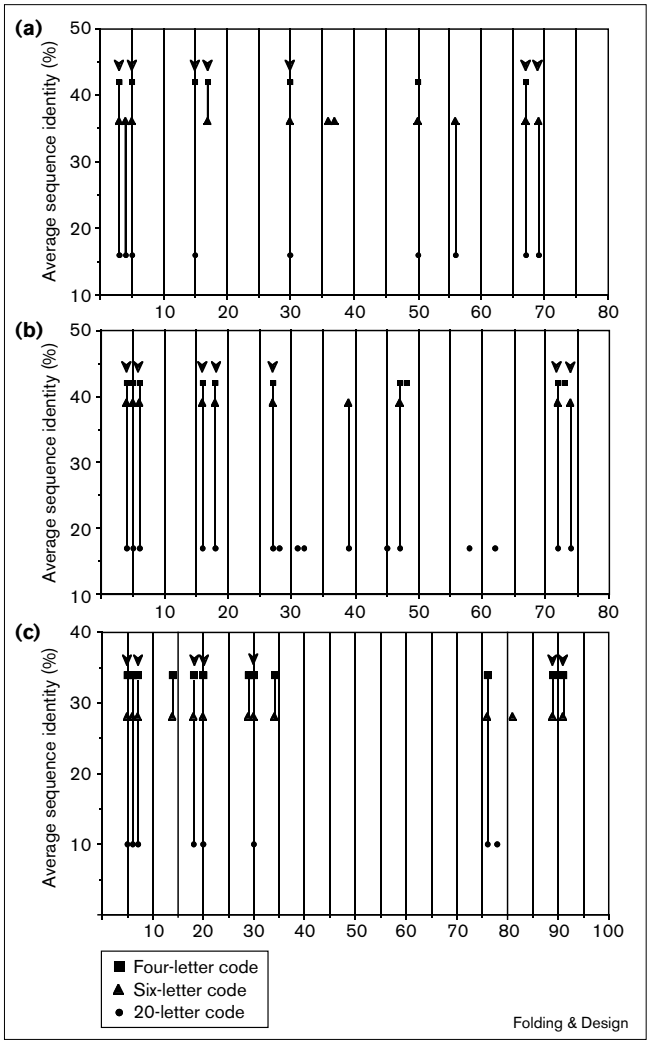
With regards to the issue of the classes of residues observed in RA sequences versus those predicted by the design simulations, the common substitution we see, aliphatic for aromatic (and vice versa), is a conservative one (Figure 4). To test how robust our procedure is for detecting specific features, we also analyzed sequence entropies at the level of four-letter and six-letter codes. This accomplishes two things. First, if the correlations we make are merely coincidental, then when we reduce the complexity of the sequence space so that the sequence homology increases we should see more and more correlations between sequence and design entropies. We chose two levels of residue code simplification: a six-letter code and a four-letter code. The six-letter code divided the residues into six groups: aliphatic (Leu, Ile, Val and Met), aromatic (Phe, Tyr and Trp), small and polar (Ala, Gly, Ser, Thr, Asn and Gln), basic (His, Lys and Arg), acidic (Asp and Glu) and cysteine. Cysteine was treated as a separate group because of its possible involvement in disulfide bonding; this may not be relevant here, however, because raf and ubiquitin are both intracellular proteins and because neither is involved in oxidation-reduction reactions, it is unlikely that the conservation of cysteines would reflect any functional involvement of such residues. The four-letter code is a simplification of the six-letter code in which aliphatic and aromatic classes are grouped together and cysteine is added to the small and polar group. The results of our analysis with the four-letter and six-letter codes were surprising and opposite to expected. The number of sites with low entropy correlation

in designed and aligned RA sequences actually decreased in some cases for both ubiquitin and raf (Figure 6). The reason for this paradoxical result is that although there is an overall decrease in entropy going from the 20- to six- to four-letter code, the contrast between low and high entropy sites becomes more defined. This is because the distribution of entropy values is much broader at the level of the 20-letter code than it is at the level of the six-letter code. For example, imagine that at a particular sequence position, all the substitutions are hydrophobic residues, but the numbers of each type of residue are equally represented. It might not be possible to distinguish this case from one in which there are several different types of residue substitutions, such as hydrophobic and small polar types, but with a more narrow distribution than the first case. At the six-letter code level, however, at the position for which substitutions were equally represented, all residue substitutions will have an entropy of zero, but the other site will have a positive entropy.

Comparisons of designed and RA sequences may reveal sites important to the stability of protein superfamilies or unique to a single member

We have identified seven sequence positions common to ubiquitin and raf that meet our criteria as being participants in the folding nucleus. There are clearly other positions that show either a high correlation between designed and RA sequence entropies, however, either in one of the proteins or both (Figures 2, 3 and 6). These are not residues that are likely to participate in the folding nucleus because they do not make native contacts with other nucleus-forming residues. Furthermore, they do not serve any functional role. These residues may be involved in the stability of the native state; that is, for the 'design' of these folds. An example is Phe4 in ubiquitin and the corresponding Arg5 in raf (the same site is also identified in ferredoxin, discussed below). They may, however, be involved in stabilizing the native structure of these proteins through the formation of surface clusters of interacting amino acids [46]. The reason that these positions are conserved may have to do with their positions in the structures. In both ubiquitin and raf this residue is located in the first and central strand of the β sheet (lower part of Figure 1a,b). It has already been noted that such central strand positions are important for the stability of β sheet proteins and attempts to develop indices of β sheet 'propensities' for residue types have concentrated on studying mutations at such sites and the effects of intra-strand and inter-strand substitutions on protein stability [47–50]. Our results suggest that the conservation of this position is crucial to the stability of the ubiquitin/raf fold, but not necessarily to its folding kinetics. It is useful to note that our analysis can reveal such details of fold design and may prove useful in this respect for other folds. There are positions that are conserved in the RA sequences that show a low entropy in design for one protein but not the other. A good example is

Figure 6



A comparison of low entropy sites for four-, six- and 20-letter codes versus the average sequence identities for the aligned sequence compared with all RA sequences. **(a)** Ubiquitin, **(b)** raf and **(c)** ferredoxin. Arrowheads indicate sequence positions that show low entropy simultaneously in simulated and ubiquitin-aligned, raf-aligned and ferredoxin-aligned RA sequences.

Leu47 of raf. This site is also a surface residue associated with a long loop found in raf, but not ubiquitin, and participates in a surface hydrophobic cluster in the raf structure.

Thus, it appears that the analysis we have presented here may be useful not only to interpret sequence data to identify residues that participate in the folding nucleus, but also to identify residues that may contribute to the stability of a specific fold type or even positions uniquely important to a member of a superfamily.

Extension of analysis to other superfamily members

We have extended our analyses to the rest of the proteins represented in the ubiquitin superfamily, including the

protein L IgG-binding domain of *Peptostreptococcus Magnus* ([51]; Figure 1d,e), the streptococcal protein G B1 IgG-binding domain [52,53], ferredoxin 1 from Anabaena 7120 [54], aldehyde oxidoreductase from *Desulfovibrio gigas* [55], toxic-shock syndrome toxin-1 from *Staphylococcus aureus* [7], enterotoxin B superantigen from *S. aureus* [56] and putidaredoxin from *Pseudomonas Putida* [57]. These proteins are topographically related to ubiquitin and raf in that the order of secondary structure elements and positions in their structures are identical; they are all of different polypeptide lengths than ubiquitin and raf, however. We examined the structure and sequence alignment for these proteins. We found no sequence relationship between the two IgG-binding domains and raf and ubiquitin beyond the topological features. These proteins are structurally quite different from raf and ubiquitin in respect to their dimensions, lengths of secondary structural elements and the orientations of these elements relative to each other. We were able to align the sequences of all other members such that a similar alignment of sequence homologies was possible (Figure 2). For example, ferredoxin 1 has topological features and dimensions of structural elements similar to raf and ubiquitin (Figure 1). Ferredoxin 1 also has the same pattern of sequence identities as ubiquitin and raf; that is, an alignment of secondary structural elements between ubiquitin and ferredoxin results in a simultaneous alignment of structural features, as well as alignment of the residues that we identified in our comparison between design sequences and natural sequences (Figure 2). We were able to identify these weak sequence and structural correlations in spite of the fact that these proteins contain several additional structural elements. It should be noted that at some of the potential nucleus sites, the residue types represented in the superfamily members are not necessarily of the same types as observed in ubiquitin and raf. For instance, at residue 69 of ubiquitin, the corresponding substitution in ferredoxin 1 and putidaredoxin is a threonine instead of an aliphatic residue. Although these substitutions represent a different class of amino acid, the threonine sidechain nevertheless contacts other nucleus residues through its γ-methyl. We make this point to reinforce the idea that it is not the type of residue in an individual sequence that identifies it as a potential nucleation site, but rather the overall tendency to observe conservation at this site when many sequences are compared. It is also possible that the position of a nucleus residue in the sequence and the possible interactions that can occur in the folding nucleus, rather than a specific type, are more important in determining whether a residue participates in the folding nucleus. To determine if the same analysis we performed on raf and ubiquitin would result in the same prediction of potential nucleation sites, we performed sequence-design simulations on ferredoxin (PDB code 1frd) as for ubiquitin and raf and performed the same analysis (Figure 6c). As can be seen, the residue positions that correspond to those identified in ubiquitin and raf as potential nucleus sites are also identified in ferredoxin. In addition, the key surface residue noted above, Phe4 in ubiquitin and Arg5 in raf, is also detected: Arg6 in ferredoxin.

These results suggest that the analysis presented here may be useful for detecting superfamily members in sequences of proteins where the structure is unknown. This could be achieved by introducing additional restraints via 'anchoring' conserved residues to the nucleus of the scaffold structure through which a sequence is threaded. The fact that it distinguishes structures that are only superficially related (i.e. the IgG-binding domains) from a truly related protein may make this analysis a useful method for subdividing superfamilies. This begs the question of how proteins with similar topology, in this case the IgG-binding domains, are related to the other members if there is no apparent relationship at the sequence level. It may be that some other relationship does exist and we are investigating this.

Structure and sites of conserved residues

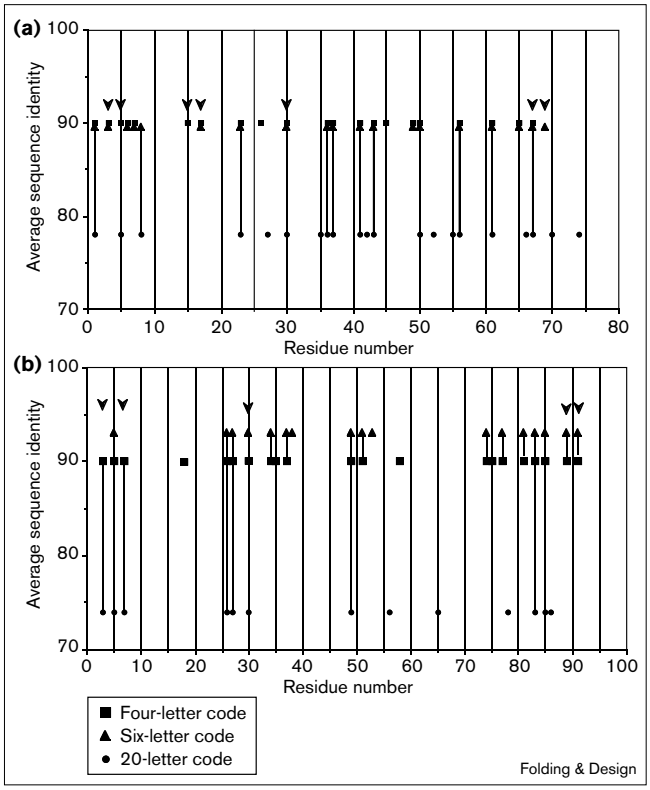
We examined the positions of conserved, potential nucleus residues in ubiquitin, raf and ferredoxin 1 (upper part of Figure 1a,b,c, respectively). All but one of these sites are located in the five-stranded β sheet in the central three strands. All the residues make at least one contact with another site. All the residues are also part of the hydrophobic core of the native structure. If these residues are those that participate in the folding nucleus then one might imagine a transition-state structure in which the central strands of the β sheet along with the helix, formed or not, are tethered to this structure. The nascent sheet may then form a scaffold for the condensation of the rest of the sheet and the helix around it. In addition, the design site (Phe4, Arg5 and Arg6 in ubiquitin, raf, and ferredoxin 1, respectively) could aid in stabilizing the nascent sheet by acting as a central scaffold on the outer surface of the sheet during condensation. Studies of site-specific β-sheet amino acid propensities have concentrated on such central regions of the sheet. As noted above, two groups have studied the effects of central β-sheet amino acid substitutions on the stability of the streptococcal protein G B1 IgG-binding domain [47–50]. It was found that the stability of the proteins to thermal denaturation was dependent on the types of substitutions that were made on the same or adjacent strands. The propensity of an amino acid to be found in a β sheet depends on the residues that will be its immediate neighbors in the final native structure. Also of note, is that there is only one potential nucleus residue in the α helices of ubiquitin, raf and ferredoxin 1, the only other major region of secondary structure in the protein. These results may suggest that the formation of the helix during folding is independent of nucleation and instead, is directed to its native structure by the local secondary structure propensity of the sequence and template-assistance by the β sheet. An interesting study on the protein G B1 domain has demonstrated that the native-helix region can be replaced with a

'chameleon' sequence with mixed propensity for both helix and β-strand formation. Replacing the native helix or a β strand with the chameleon peptide resulted in a native structure identical to the wild-type protein, demonstrating that secondary structure in these regions is context dependent [58]. It would be interesting to see if similar such substitutions could be made in ubiquitin, raf and ferredoxin.

A comparison of designed and highly homologous sequences does not allow for the distinction of uniquely conserved residues in ubiquitin and ferredoxin I

As noted above, for our procedure to work, we must choose natural sequences that are likely to fold to the same structure as our superfamily, but are not so homologous that we cannot distinguish a nucleus composed of residues conserved for functional or design reasons. To illustrate this problem we examined sequence alignments for evolutionarily related ubiquitin and ferredoxin sequences; that is, those having > 50% identity and therefore likely to have identical folds [1,59]; a list of alignments for raf was not available. These sequences were extracted from the HSSP sequence database. The HSSP files also contain calculated position residue entropies, calculated as described below and in [59]. Among these sequences the average sequence identity was very high (76%, averaged over all sequences for the 20-letter code for ubiquitin and 54% for ferredoxin). When we performed the comparison of designed and natural sequences we found a very high correspondence of low-entropy sites (Figure 7). In fact, if we examine all four-, six- and 20-letter code correlations we find that all low-entropy sites in the natural sequences are predicted by the design simulations. These results suggest that the design simulations are effective at predicting the conservation of residues in natural, evolutionarily close sequences, as has been demonstrated before [12]. The high levels of residue identity and similarity among evolutionarily close sequences, however, make it impossible to distinguish the conservation of unique common residue positions from those that are conserved simply because they have not had the time to diverge.

Figure 7



Correlation of low entropy sites for four-, six- and 20-letter codes for simulated and natural ubiquitin and ferredoxin sequences versus average sequence identities for **(a)** ubiquitin and **(b)** ferredoxin.

Discussion

A key question in asking how proteins fold and how folds are conserved despite the evolution of protein sequences, is what information in a sequence must be conserved to ensure that it will fold to a particular structure? A result of nucleation-growth theory is that a minimum requirement is the conservation of the transition-state nucleus itself [12]. Mutation of residues in these nuclei would ultimately result in sequences whose free-energy barrier for folding would be increased to a point at which they are no longer biologically viable; alternatively, a sequence could evolve a different nucleus, resulting in a different fold. In either case, it is central to nucleation-growth that nucleus-forming residues are conserved; that is, folds themselves may evolve upwards or laterally towards other folds. The identification of such sites in a sequence would be a useful indicator of whether a sequence folds to a particular class of three-dimensional structure. Here, we have demonstrated that three proteins with identical folds retain the conservation of key residues despite a very low (10%) sequence identity, and that these sites may be predicted with the sequence-design strategy discussed here. We also showed that the results are specific, in that proteins having superficially similar, but not identical folds, do not show the same patterns of sequence identity. These results have immediate practical use as well as being interesting in themselves. Attempts to assign a particular fold to a novel protein sequence, in the absence of high sequence homology to any known structure, may be possible using a combination of secondary structure prediction and the use of a threading algorithm as is currently done. We propose that an additional step would be to perform a sequence design simulation for a particular fold, and perform a comparison of design and highly diverged natural sequence entropies to identify potential folding-nucleus residues.

Our strategy in this work was to infer the possible nucleation sites for ubiquitin and related protein folds using evolutionary information and sequence design. This approach

could allow us to avoid the formidable task of simulating complete folding of these proteins. A previous application of related ideas [12] helped in the prediction of nucleus residues in CI2. Alternative approaches have been proposed by Shoemaker *et al.* [60], to study the distribution of contacts in the transition-state ensemble, that agree reasonably well with previously published experimental results for CI2. The main idea of the approach presented in [60] is to search for stronger and entropically least costly contacts in constrained simulations. To this end, the results may be sensitive to parameters used and other possible constraints in the simulations. It would be interesting to explore this approach further to see if it is able to generate successful prediction of nucleation sites for other proteins, including ubiquitin, for which the nucleus has not been experimentally characterized.

Another consequence of our results is a suggestion to use the information in this study as an experimental design criterion. Attempts to design sequences that fold to a particular structure from partially random sequences have been shown to be successful, even with very minimal bias, such as a two letter (hydrophobic, hydrophilic) code [61,62]. It would be interesting to see the effect of generating semi-random oligonucleotide libraries biased to code for the nucleus of a particular fold, and examine whether or not such a fold is indeed achieved. Rational engineering efforts have recently resulted in transmutation of one fold into another, in which the original sequence was altered by only 50% [63]. Our results suggest that significantly fewer substitutions could be made, resulting in sequences in which perhaps only 20–30% of the sequence was altered if one started with sequences that were already minimally different and key substitutions of nucleus and design residues were made. Efforts in this direction are in progress and will be reported elsewhere.

We have used the expression 'nucleus-forming' residues throughout this article as being equivalent to sequence positions of low entropy detected in sequence design simulations and in natural sequences. It must be made clear, however, that such observations do not prove that such residue positions do take part in the folding nucleus. Site-directed mutagenesis, combined with kinetic studies of raf, ubiquitin and ferredoxin folding will be necessary to establish whether or not specific residues identified in these studies participate in the folding nucleus. These studies are now in progress.

Materials and methods

The design procedure has been described in detail elsewhere [42,43]. It is a stochastic (Monte-Carlo) optimization routine in sequence space that keeps amino acid composition unchanged. It minimizes the energy of the native conformation, based on calculation of a contact potential defined in Equation 1. The condition of constant amino acid composition makes it equivalent to optimizing the relative energy of the native state, or Z-score. As is characteristic of Monte-Carlo searches, unfavorable mutations can also be accepted, with a small probability, given by a Metropolis criterion with selective temperature T_{sel}.

The sequence alignment of raf with RA domain sequences was taken directly from Ponting and Benjamin [5]. Alignments of ubiquitin and ferredoxin 1 with the RA sequences were first based on superposition of their structures (PDB codes 1ubi and 1frd; [25,54]) with that of raf determined by NMR (PDB code 1rfa; [27]). The superposition was optimized to give the minimal rms atomic deviations (cutoff 2.0 Å) of backbone atom positions using the program DALI [64,65], and the alignments were then checked visually. The sequence alignments of ubiquitin and ferredoxin to raf were then made and alignment was made to the RA sequences based on the superposition of the ubiquitin and ferredoxin alignments to raf and the raf alignment to the RA sequences. These were then visually inspected to see if further adjustments could be made. The only sequences that needed further visual adjustments were those of the ferredoxin family. Here, two alternative structural alignments of the ferredoxin helix are possible. This was because these structures have three turn helices compared with raf and ubiquitin, which have four turn helices. Of the two alignments, we found that the DALI-based alignment placed the first helical turns of ubiquitin/raf and the ferredoxins in the same place, whereas we found that the optimal amino acid sequence alignment was that with the first helical turns of ubiquitin/raf matched to the positions of the second helical turns in the ferredoxins.

Alignments of the IgG-binding domain structures [51–57] with raf failed to show adequate superposition to meet the cutoff condition, nor were alignments of the sequences with that of raf based on secondary structure alignment successful.

Sequence entropies were calculated as for the designed sequences using Equation 2 [59]. The six-letter code was classified by six groups: aliphatic (Val, Ile, Leu and Met), aromatic (Phe, Trp and Tyr), small and polar (Ala, Gly, Pro, Ser, Thr, Gln and Asn), basic (His, Lys and Arg), acidic (Asp and Glu), and cysteine. The four-letter code is the same as the six-letter code, except that the hydrophobic and aliphatic classes were combined and cysteine was grouped with the small polar class.

Alignments and sequence entropies for ubiquitin and ferredoxin sequences were extracted from the HSSP database as described in the Results section. Four-letter and six-letter code entropies were calculated from sequence residue frequencies contained in these files using Equation 2.

Solvent surface accessibility was calculated using the program DSSP [66].

$$E = \sum_{i<j} B(\eta_i, \eta_j)\ \Delta(r_i - r_j) \quad (1)$$

$$S(i) = -\sum_{j=1}^{m} p_j(i) \ln p_j(i) \quad (2)$$

Acknowledgements

This work was supported by the MRC of Canada (grant no. DGN 059 to S.W.M.), The Burroughs-Wellcome Fund (S.W.M.) and The National Institute of Health (grant no. 52126 to E.S.). S.W.M. is an Awardee of a Burroughs-Wellcome Fund New Investigator Award in the Basic Pharmacological Sciences. We are grateful to Joelle Pelletier and Leonid Mirny for helpful discussions and for carefully reading the manuscript.

References

1. Doolittle, R.F. (1995). The multiplicity of domains in proteins. *Annu. Rev. Biochem.* **64**, 287-314.
2. Doolittle, R.F., Feng, D.F., Tsang, S., Cho, G. & Little, E. (1996). Determining divergence times of the major kingdoms of living organisms with a protein clock. *Science* **271**, 470-477.
3. Holm, L. & Sander, C. (1993). Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* **233**, 123-138.

4. Holm, L. & Sander, C. (1996). Alignment of three-dimensional protein structures: network server for database searching. *Methods Enzymol.* **266**, 653-662.
5. Ponting, C.P. & Benjamin, D.R. (1996). A novel family of Ras-binding domains. *Trends Biochem. Sci.* **21**, 422-425.
6. Russell, R.B., Copley, R.R. & Barton, G.J. (1996). Protein fold recognition by mapping predicted secondary structures. *J. Mol. Biol.* **259**, 349-365.
7. Abkevich, V.I., Gutin, A.M. & Shakhnovich, E.I. (1994). Specific nucleus as the transition state for protein folding: evidence from the lattice model. *Biochemistry* **33**, 10026-10036.
8. Guo, Z.Y. & Thirumalai, D. (1995). Kinetics of protein folding - nucleation mechanism, time scales, and pathways. *Biopolymers* **36**, 83-102.
9. Alexander, P., Orban, J. & Bryan, P. (1992). Kinetic analysis of folding and unfolding the 56 amino acid IgG-binding domain of streptococcal protein G. *Biochemistry* **31**, 7243-7248.
10. Johnson, B.H. & Hecht, M.H. (1994). Recombinant proteins can be isolated from *E. coli* cells by repeated cycles of freezing and thawing. *BioTechnology* **12**, 1357-1360.
11. Schindler, T., Herrler, M., Marahiel, M.A. & Schmid, F.X. (1995). Extremely rapid protein folding in the absence of intermediates. *Nat. Struct. Biol.* **2**, 663-673.
12. Shakhnovich, E.I., Abkevich, V. & Ptitsyn, O. (1996). Conserved residues and the mechanism of protein folding. *Nature* **379**, 96-98.
13. Shakhnovich, E.I. (1997). Theoretical studies of protein-folding thermodynamics and kinetics. *Curr. Opin. Struct. Biol.* **7**, 29-40.
14. Sosnick, T.R., Mayne, L., Hiller, R. & Englander, S.W. (1994). The barriers in protein folding. *Nat. Struct. Biol.* **1**, 149-156.
15. Viguera, A.R., Blanco, F.J. & Serrano, L. (1995). The order of secondary structure elements does not determine the structure of a protein but does affect its folding kinetics. *J. Mol. Biol.* **247**, 670-681.
16. Perl, D., *et al*., & Schmid, F.X. (1998). Conservation of rapid two-state folding in mesophilic, thermophilic and hyperthermophilic cold shock proteins. *Nat. Struct. Biol.* **5**, 229-235.
17. Itzhaki, L.S., Otzen, D.E. & Fersht, A.R. (1995). The structure of the transition state for folding of chymotrypsin inhibitor 2 analysed by protein engineering methods: evidence for a nucleation-condensation mechanism for protein folding. *J. Mol. Biol.* **254**, 260-288.
18. 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, CI-2. *Fold. Des.* **1**, 43-55.
19. Jackson, S.E., Main, E.R.G. & Fulton, K.F. (1998). Folding of FKBP12: pathway of folding and thermodynamic analysis of the transition state for folding. *Biochemistry*, in press.
20. de Prat Gay, G., Ruiz-Sanz, J., Davis, B. & Fersht, A.R. (1994). The structure of the transition state for the association of two fragments of the barley chymotrypsin inhibitor 2 to generate native-like protein: implications for mechanisms of protein folding. *Proc. Natl Acad. Sci. USA* **91**, 10943-10946.
21. Fersht, A.R. (1995). Optimization of rates of protein folding - the nucleation-condensation mechanism and its implications. *Proc. Natl Acad. Sci. USA* **92**, 10869-10873.
22. Itzhaki, L.S., Otzen, D.E. & Fersht, A.R. (1995). The structure of the transition state for folding of chymotrypsin inhibitor 2 analysed by protein engineering methods - evidence for a nucleation-condensation mechanism for protein folding. *J. Mol. Biol.* **254**, 260-288.
23. Ruiz-Sanz, J., de Prat Gay, G., Otzen, D.E. & Fersht, A.R. (1995). Protein fragments as models for events in protein folding pathways: protein engineering analysis of the association of two complementary fragments of the barley chymotrypsin inhibitor 2 (CI-2). *Biochemistry* **34**, 1695-1701.
24. Nassar, N., *et al*., & Wittinghofer, A. (1995). The 2.2 Å crystal structure of the Ras-binding domain of the serine/threonine kinase c-Raf1 in complex with Rap1A and a GTP analogue. *Nature* **375**, 554-560.
25. Vijay-Kumar, S., Bugg, C.E. & Cook, W.J. (1987). Structure of ubiquitin refined at 1.8 Å resolution. *J. Mol. Biol.* **194**, 531-544.
26. Emerson, S.D., *et al*., & Fry, D.C. (1994). Chemical shift assignments and folding topology of the Ras-binding domain of human Raf-1 as determined by heteronuclear three-dimensional NMR spectroscopy. *Biochemistry* **33**, 7745-7752.
27. Emerson, S.D., *et al*., & Fry, D.C. (1995). Solution structure of the Ras-binding domain of c-Raf-1 and identification of its Ras interaction surface. *Biochemistry* **34**, 6911-6918.
28. Di Stefano, D.L. & Wand, A.J. (1987). Two-dimensional ^{1}H NMR study of human ubiquitin: a main chain directed assignment and structure analysis. *Biochemistry* **26**, 7272-7281.
29. Briggs, M.S. & Roder, H. (1992). Early hydrogen-bonding events in the folding reaction of ubiquitin. *Proc. Natl Acad. Sci. USA* **89**, 2017-2021.
30. Cox, J.P., Evans, P.A., Packman, L.C., Williams, D.H. & Woolfson, D.N. (1993). Dissecting the structure of a partially folded protein. Circular dichroism and nuclear magnetic resonance studies of peptides from ubiquitin. *J. Mol. Biol.* **234**, 483-492.
31. Harding, M.M., Williams, D.H. & Woolfson, D.N. (1991). Characterization of a partially denatured state of a protein by two-dimensional NMR: reduction of the hydrophobic interactions in ubiquitin. *Biochemistry* **30**, 3120-3128.
32. Khorasanizadeh, S., Peters, I.D., Butt, T.R. & Roder, H. (1993). Folding and stability of a tryptophan-containing mutant of ubiquitin. *Biochemistry* **32**, 7054-7063.
33. Khorasanizadeh, S., Peters, I.D. & Roder, H. (1996). Evidence for a three-state model of protein folding from kinetic analysis of ubiquitin variants with altered core residues. *Nat. Struct. Biol.* **3**, 193-205.
34. Stockman, B.J., Euvrard, A. & Scahill, T.A. (1993). Heteronuclear three-dimensional NMR spectroscopy of a partially denatured protein: the A-state of human ubiquitin. *J. Biomol. NMR* **3**, 285-296.
35. Wintrode, P.L., Makhatadze, G.I. & Privalov, P.L. (1994). Thermodynamics of ubiquitin unfolding. *Proteins* **18**, 246-253.
36. Woolfson, D.N., Cooper, A., Harding, M.M., Williams, D.H. & Evans, P.A. (1993). Protein folding in the absence of the solvent ordering contribution to the hydrophobic interaction. *J. Mol. Biol.* **229**, 502-511.
37. Ptitsyn, O. (1998). Protein folding and protein evolution: common folding nucleus in different subfamilies of c-type cytochromes. *J. Mol. Biol.* **278**, 655-666.
38. Mirny, L.A. & Shakhnovich, E.I. (1996). How to derive a protein folding potential? A new approach to an old problem. *J. Mol. Biol.* **264**, 1164-1179.
39. Varshavsky, A. (1996). The N-end rule: functions, mysteries, uses. *Proc. Natl Acad. Sci. USA* **93**, 12142-12149.
40. Smith, S.E., Koegl, M. & Jentsch, S. (1996). Role of the ubiquitin/proteasome system in regulated protein degradation in *Saccharomyces cerevisiae*. *Biol. Chem.* **377**, 437-446.
41. Hershko, A. (1996). Lessons from the discovery of the ubiquitin system. *Trends Biochem. Sci.* **21**, 445-449.
42. Wittinghofer, A. & Nassar, N. (1996). How Ras-related proteins talk to their effectors. *Trends Biochem. Sci.* **21**, 488-491.
43. Avruch, J., Zhang, X.F. & Kyriakis, J.M. (1994). Raf meets Ras: completing the framework of a signal transduction pathway. *Trends Biochem. Sci.* **19**, 279-283.
44. Shakhnovich, E.I. & Gutin, A.M. (1993). A new approach to the design of stable proteins. *Protein Eng.* **6**, 793-800.
45. Shakhnovich, E.I. & Gutin, A.M. (1993). Engineering of stable and fast-folding sequences of model proteins. *Proc. Natl Acad. Sci. USA* **90**, 7195-7199.
46. Tisi, L.C. & Evans, P.A. (1995). Conserved structural features on protein surfaces: small exterior hydrophobic clusters. *J. Mol. Biol.* **249**, 251-258.
47. Minor, D., Jr. & Kim, P.S. (1994). Measurement of the beta-sheet-forming propensities of amino acids. *Nature* **367**, 660-663.
48. Minor, D., Jr. & Kim, P.S. (1994). Context is a major determinant of beta-sheet propensity. *Nature* **371**, 264-267.
49. Smith, C.K., Withka, J.M. & Regan, L. (1994). A thermodynamic scale for the beta-sheet forming tendencies of the amino acids. *Biochemistry* **33**, 5510-5517.
50. Smith, C.K. & Regan, L. (1995). Guidelines for protein design: the energetics of beta sheet side chain interactions. *Science* **270**, 980-982.
51. Wikstrom, M., Drakenberg, T., Forsen, S., Sjobring, U. & Bjorck, L. (1994). Three-dimensional solution structure of an immunoglobulin light chain-binding domain of protein L. Comparison with the IgG-binding domains of protein G. *Biochemistry* **33**, 14011-14017.
52. Gronenborn, A.A., *et al*., & Clore, G.M. (1991). A novel, highly stable fold of the immunoglobulin binding domain of streptococcal protein G. *Science* **253**, 657-661.
53. Achari, A., *et al*., & Whitlow, M. (1992). 1.67-Å X-ray structure of the B2 immunoglobulin-binding domain of streptococcal protein G and comparison to the NMR structure of the B1 domain. *Biochemistry* **31**, 10449-10457.
54. Jacobson, B.L., Chae, Y.K., Markley, J.L., Rayment, I. & Holden, H.M. (1993). Molecular structure of the oxidized, recombinant, heterocyst [2Fe-2S] ferredoxin from Anabaena 7120 determined to 1.7-Å resolution. *Biochemistry* **32**, 6788-6793.

55. Romao, M.J., *et al*., & Huber, R. (1995). Crystal structure of the xanthine oxidase-related aldehyde oxido-reductase from *D. gigas*. *Science* **270**, 1170-1176.
56. Swaminathan, S., Furey, W., Pletcher, J. & Sax, M. (1992). Crystal structure of staphylococcal enterotoxin B, a superantigen. *Nature* **359**, 801-806.
57. Ye, X.M., Pochapsky, T.C. & Pochapsky, S.S. (1992). ^{1}H NMR sequential assignments and identification of secondary structural elements in oxidized putidaredoxin, an electron-transfer protein from Pseudomonas. *Biochemistry* **31**, 1961-1968.
58. Minor, D., Jr. & Kim, P.S. (1996). Context-dependent secondary structure formation of a designed protein sequence. *Nature* **380**, 730-734.
59. Sander, C. & Schneider, R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins* **9**, 56-68.
60. Shoemaker, B.A., Wang, J. & Wolynes, P.G. (1997). Structural correlations in protein folding funnels. *Proc. Natl Acad. Sci. USA* **94**, 777-782.
61. Beasley, J.R. & Hecht, M.H. (1997). Protein design: the choice of de novo sequences. *J. Biol. Chem.* **272**, 2031-2034.
62. Kamtekar, S., Schiffer, J.M., Xiong, H., Babik, J.M. & Hecht, M.H. (1993). Protein design by binary patterning of polar and nonpolar amino acids. *Science* **262**, 1680-1685.
63. Dalal, S., Balasubramanian, S. & Regan, L. (1997). Protein alchemy - changing beta-sheet into alpha-helix. *Nat. Struct. Biol.* **4**, 548-552.
64. Holm, L. & Sander, C. (1997). Dali/FSSP classification of three-dimensional protein folds. *Nucleic Acids Res.* **25**, 231-234.
65. Holm, L. & Sander, C. (1995). Dali: a network tool for protein structure comparison. *Trends Biochem. Sci.* **20**, 478-480.
66. Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577-2637.
67. Kraulis, P.J. (1991). MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr.* **24**, 946-950.

Because *Folding & Design* operates a 'Continuous Publication System' for Research Papers, this paper has been published on the internet before being printed. The paper can be accessed from http://biomednet.com/cbiology/fad – for further information, see the explanation on the contents pages.