Glimmers in the Midnight Zone: Characterization of Aligned Identical Residues in Sequence-Dissimilar Proteins Sharing a Common Fold

Iddo Friedberg, Tommy Kaplan and Hanah Margalit

Department of Molecular Genetics and Biotechnology
The Hebrew University- Hadassah Medical School
POB 12272, Jerusalem 91120 ISRAEL

Telephone: 972-2-6758647
Fax: 972-2-6784010
e-mail: idoerg@cc.huji.ac.il

Keywords: molecular evolution; sequence conservation; protein structure; protein folding

Abstract

Sequence comparison of proteins that adopt the same fold has revealed a large degree of sequence variation. There are many pairs of structurally similar proteins with only a very low percentage of identical residues at structurally aligned positions. It is not clear whether these few identical residues have been conserved just by coincidence, or due to their structural and/or functional role. The current study focuses on characterization of STructurally Aligned Identical ResidueS (STAIRS) in a data set of protein pairs that are structurally similar but sequentially dissimilar. The conservation pattern of the residues at structurally aligned positions has been characterized within the protein families of the two pair members, and mutually highly and weakly conserved positions of STAIRS could be identified. About 40% of the STAIRS are only moderately conserved, suggesting that their maintenance may have been coincidental. The mutually highly conserved STAIRS show distinct features that are associated with protein structure and function: a relatively high fraction of these STAIRS are buried within their protein structures. Glycine, cysteine, histidine, and tryptophan are significantly over-represented among the mutually conserved STAIRS. A detailed survey of these STAIRS reveals residue-specific roles in the determination of the protein's structure and function.

Introduction

The fold of a protein is determined chiefly by its sequence (Anfinsen, 1973), implying that similar sequences fold into similar three-dimensional structures. However, with the recent explosion in protein sequence and structural data, it has been demonstrated that many protein pairs with sequence identities of only 10% or even less may assume the same fold. This level of similarity, virtually undetectable by sequence alignment methods, and in many cases no better than that expected in random (Rost, 1997), has been termed to reside within the "midnight zone" of protein sequence similarity (Rost, 1999).

The large proportion of variable residues at equivalent positions along the structural alignments of pairs of structurally-similar, sequence-dissimilar (SSSD) proteins suggests that many sequence positions have no critical role in structure determination. This was demonstrated both experimentally and computationally. Experimentally it was shown that many of the mutations introduced along a protein sequence have had no effect on the protein's stability and/or activity (Rennell *et al.*, 1991; Markiewicz *et al.*, 1994; Milla *et al.*, 1994; Suckow *et al.*, 1996). Recent computational studies have reached similar conclusions by comparing the patterns of conservation within families that share the same structure and compose large protein superfamilies (Mirny *et al.*, 1998; Mirny and Shakhnovich, 1999; Ptitsyn and Ting, 1999). These studies identified a small number of conserved spatially close residues within each family. The conserved clusters of residues occupied equivalent positions across the superfamily, suggesting that they may form folding nuclei. The cluster residues were not necessarily similar when compared between families within the superfamily, implying that folding nuclei composed of different sets of residues may lead to similar structures.

The current study focuses on aligned *identical* residues in remote sequences that assume the same structure. We named these residues STAIRS (STructurally Aligned Identical ResidueS). By evaluating the evolutionary conservation of STAIRS within their corresponding protein families we attempt to distinguish between STAIRS whose occurrence may be just arbitrary and the ones that pertain to the structure and/or function of the proteins.

The latter are analyzed to search for common features that may be related to their structural role. We examine the distribution of amino acids in STAIRS, their spatial proximity, and their structural location. The approach taken is a global one, in which we analyze a database of 126 SSSD protein pairs and characterize the identical residues in corresponding positions. These residues are examined in a generic aspect, in order to characterize their overall attributes in the population of known SSSD protein pairs. Hence, the analysis reaches general conclusions regarding these residues, and does not identify features that are specific to a certain protein family.

Results

Database of SSSD protein pairs Our database consists of 133 proteins, none of which exhibits more than 25% sequence identity with the other proteins. These are organized into 126 protein pairs that are similar in structure but are dissimilar in sequence.

Residue type distribution STAIRS were determined based on DALI structural alignments (Holm and Sander, 1996) of the paired sequences (see Methods). In total there are 2324 STAIRS out of 18711 aligned positions in the SSSD protein pairs. On average the STAIRS comprise 12% of the aligned positions in an SSSD protein pair, with a standard deviation of 3.8%. In order to find out whether there are certain types of residues that tend to be kept unchanged in the structural alignments, the distribution of amino acid residues within the STAIRS data set was compared to their random distribution, based on the amino acid frequencies in the entire database of SSSD proteins. Using a χ^2 contingency test we could show that the distribution of amino acid residues among the STAIRS differs significantly from that expected by random ($\chi^2_{(df=19)}$=294.7, p≤0.0001). We further analyzed the individual amino acids to detect the residues that contributed mostly to this significance. This was done by extracting the significance of the χ^2 with df=1 for each individual cell in the 20×2 table. Amino acid residues with significant deviations between observed and expected frequencies ($p_{(df=1)} \leq 0.05$)) are either significantly over-represented (observed > expected) or underrepresented (observed < expected) in the data set of STAIRS. Glycine and the hydrophobic amino acid residues leucine and valine were over-represented in STAIRS, whereas methionine and the polar amino acid residues, asparagine, glutamine, serine and threonine were underrepresented.

In order to ascertain that our results are not dependent on the method used for structural alignment, the residue type distribution analysis has been repeated using SSAP (Taylor and Orengo, 1989), a different structural alignment algorithm (see the Methods). The structures in each SSSD pair were aligned by SSAP, STAIRS were determined, and the analysis of residue type distribution was carried out as described above. That analysis has revealed the same fraction of STAIRS and a very similar (essentially the same) distribution of residue types among the STAIRS.

Spatial proximity of STAIRS STAIRS may be located in distant positions in the structures of the aligned pair of proteins, or they may be spatially close in both structures. STAIRS in the latter class are better candidates for playing a special structural or functional role, as apparent conservation in such a case is not only of residue identity, but of a spatial arrangement of several amino acid residues that are in contact. Such clusters may be found in folding nuclei or in functional sites. To investigate the STAIRS that are organized in spatially close clusters, a new subgroup of STAIRS was defined, dubbed NSTAIRS: Neighboring STAIRS. Two or more STAIRS that are in contact within both aligned structures were classified in the NSTAIRS subgroup (for definition of contacting residues see the Methods section). Overall, 50% of the STAIRS were identified as NSTAIRS (Table 1), but their distribution among the protein pairs differed. There were 18 protein pairs that did not show any spatial cluster of STAIRS, where the others were shown to contain at least two such clusters. The size of the clusters also varied: from cluster size of two for 177 clusters, cluster size of three for 69 clusters, and the number decreases as the cluster size increases.

Conservation of STAIRS and NSTAIRS In any random alignment of two strings there is a baseline probability of identical characters being aligned due to chance. The number of characters expected to be aligned by random is N = L/S, where L is the alignment length and S is the alphabet size. Therefore caution should be taken before drawing conclusions regarding the general characteristics of the identified STAIRS, because there is always the possibility that a fraction of the STAIRS, or maybe even most of them, have been conserved by chance. There is a need therefore to select among the STAIRS the ones for which there is higher confidence in their directed conservation. Our premise is that if a position that contains STAIRS is highly conserved in both families that correspond to the two proteins which form the pair, there is a higher confidence that the residue identity at this position is not arbitrary. The rationale being that the two protein families are so remote that mutual conservation at certain structurally aligned positions is probably meaningful.

The analysis of mutual conservation along the structural alignments was done as follows: initially, each protein was aligned with its family members. Here, the choice of sequences used in the multiple alignment is important. Using close family members of each protein will inundate us with well-conserved positions, thus the high conservation score we might expect for the few (12% in average) STAIRS identified in each chain will be masked by high conservation scores contributed by other positions. Therefore, multiple sequence alignment data which includes distantly related proteins is required. This data was obtained using PSI-BLAST (Altschul *et al.*, 1997). PSI-BLAST identifies remote homologues iteratively, by generating a profile from all the sequences identified in previous iterations, and in each iteration searching the database using the updated profile. This process is repeated until a predetermined number of iterations is reached, or until the searches converge. Using enough iterations, with a large enough database, weak but biologically meaningful similarities may be obtained.

After the multiple sequence alignment was achieved by PSI-BLAST, the degree of amino acid conservation at a position was determined by calculating the position's Information Content, IC (see the Methods). These values were then expressed as relative values: for each sequence a mean IC was determined, and each position was assigned with a z-score, which is the IC score expressed in number of standard deviations from the mean IC of the sequence, hereafter, Z_{ic}. The Z_{ic} scores at corresponding positions were used to evaluate the correlation between the degrees of conservation of aligned positions in the two pair members.

Practically, each of the sequences in the SSSD database was compared using PSI-BLAST to the NR database (a non-redundant compilation of all known protein sequences). PSI-BLAST was run for six iterations or until convergence, and IC and Z_{ic} values were determined for each position. Sequences having an identity percentage of 98% or higher with the query sequence were discarded, to avoid bias by extremely close sequences outvoting the more distant ones. It is interesting to note that PSI-BLAST detected the pair members of 46 sequences, despite their weak sequence similarity (fifteen pairs were identified when the search was carried out with each of the sequences of the pair members as a query). This is consistent with previous evaluations of the power of PSI-BLAST (Park *et al.*, 1998; Salamov *et al.*, 1999).

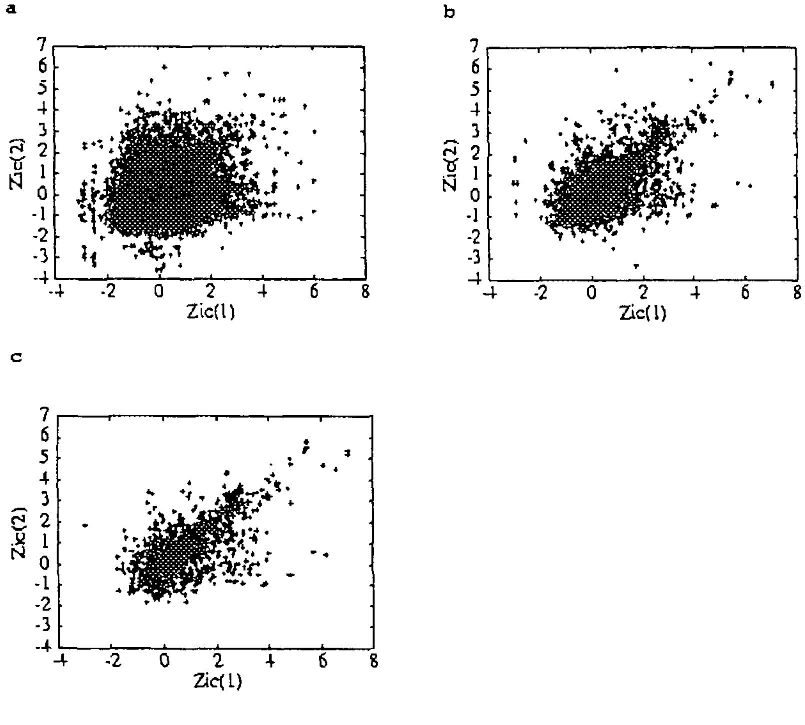

Figure 1: Correlation between the normalized conservation measures (Z_{ic} values) of all aligned positions in the database of SSSD protein pairs. a: non-STAIRS positions (r=0.3); b: STAIRS (r=0.49); c: NSTAIRS (r=0.62).

Figure 1 demonstrates the correlation between the Z_{ic} values of all aligned positions in our database. As can be seen, the correlation between Z_{ic} values of STAIRS (r= 0.49) is higher than that of non-STAIRS positions (r=0.3), and the correlation between NSTAIRS positions is the highest (r= 0.62). All correlation coefficients are statistically significant.

The Z_{ic} scores show that indeed a fraction of 39.5% of STAIRS are not highly conserved and have conservation scores that are lower than the average. These probably should not have been included in the set of residues used to characterize the STAIRS. On the other hand, at the tail of the distribution of Z_{ic} scores the presence of STAIRS and NSTAIRS is remarkable. As shown in Table 1, the STAIRS comprise about 50% of the residues with Z_{ic} values above 1.65 for both pair members ($p \leq 0.03$, see Methods), while their fraction within the whole data set is only 12%. Most remarkably, three quarters of the mutually highly conserved STAIRS are NSTAIRS.

Residue type distribution of well conserved STAIRS One of the goals of the conservation analysis was to exclude "chance" STAIRS and repeat the analyses with the remaining STAIRS. A more stringent approach would be to look only at the STAIRS that are mutually highly conserved. Indeed, repeating the residue type distribution analysis only for residues that constitute the highly conserved STAIRS (positions with $Z_{ic} \geq 1.65$ in both pair members), shows a different distribution than the one observed for the whole database of SSSD protein pairs (Figure 2). In the subset of mutually highly conserved STAIRS cysteine is significantly over-represented, as well as glycine, and the aromatic amino acids histidine and tryptophan. The same results have been obtained when the analysis has been repeated using SSAP for the structural alignments (see the Methods).

Table 1: Abundance of STAIRS and NSTAIRS

	All Aligned Positions	non-STAIRS	STAIRS (without NSTAIRS)	NSTAIRS	STAIRS (Total)
Total[a]	18711 (100)	16387 (87.5)	1170 (6.3)	1154 (6.2)	2324 (12.5)
Zic ≥ 1.65[b]	501 (100)	259 (51.6)	61 (12.2)	181 (36.2)	242 (48.4)

[a] In the whole database of SSSD proteins

[b] Only in the subset of mutually highly conserved positions

Each table entry contains the number of positions in a given category. Percentages for that category are parenthesized.

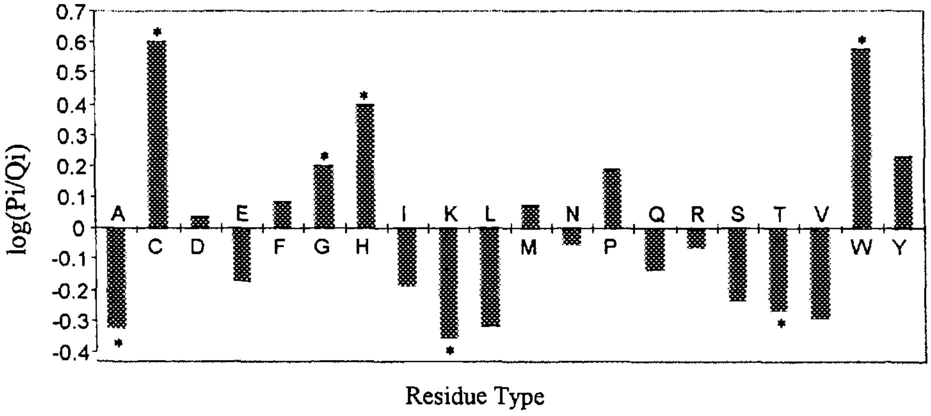

Figure 2: Distribution of residue types in mutually highly conserved STAIRS ($Z_{ic} \geq 1.65$), expressed as log-odds ratio between the frequency of a given residue type in STAIRS (P_i) and its frequency in the entire database of SSSD proteins (Q_i). Cysteine, glycine, histidine and tryptophan are significantly over-represented in STAIRS. Alanine, lysine, and threonine are significantly underrepresented ($p \leq 0.05$ by a χ^2 test).

Solvent accessibility. To examine the location of the STAIRS within the structures, the solvent accessibility was calculated for each residue in the database of SSSD proteins (see the Methods). As expected, there is a positive correlation between the accessibility values of corresponding positions within the pairs of aligned structures (r=0.52). We compared the solvent accessibility between STAIRS, NSTAIRS and the whole set of residue positions in our data (Figure 3). We looked at the fraction of buried vs. exposed residues in these categories. (See the Methods for determination of residue solvent exposure). As seen in Figure 3a, about 51% of residues are buried and about 49% are exposed. When divided into hydrophobic and hydrophilic (as detailed in the legend to Figure 3), a higher fraction of the hydrophobic residues are buried and a higher fraction of hydrophilic residues are exposed, consistent with many previous studies that demonstrated the higher presence of hydrophilic residues on the protein surface and of hydrophobic residues buried in the protein core (e.g. Lesk and Chothia, 1980). The pattern of solvent accessibility of the STAIRS is very similar to that of all residues. The accessibility pattern of NSTAIRS is remarkable, where 78% of the residues are buried, a trend that is emphasized (90% burial) within the hydrophobic residues. Note the buried/exposed ratio reversal within hydrophilic NSTAIRS: 61% of the hydrophilic NSTAIRS are buried, 39% are exposed. The same type of analysis was carried out for the mutually highly conserved positions only. These mutually conserved positions (Figure 3b) are highly populated with buried residues. This is true for all positions, and somewhat more pronounced for the STAIRS and NSTAIRS subsets where the buried residues reach 90%. Remarkably, among the mutually conserved hydrophilic STAIRS and NSTAIRS a very high fraction (85%) are buried.

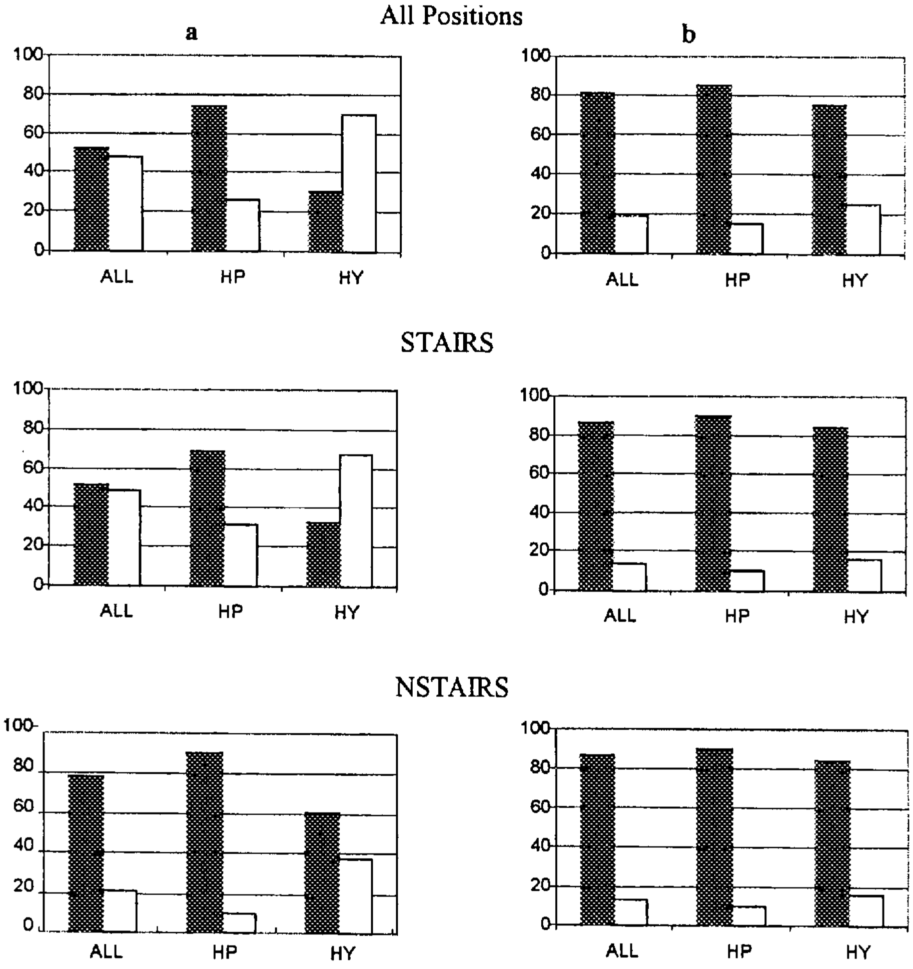

Figure 3 Distribution of residue positions by solvent accessibility.

(a) for the whole database (top), STAIRS (middle), and NSTAIRS (bottom). Percentage of buried residues (black), and exposed residues (white) is shown (see Methods). The distribution is shown for all the residues, as well as for hydrophobic (HP) and for hydrophilic (HY) residues only. (b) Same as (a), but for mutually highly conserved paired positions (only positions with $Z_{ic} \geq 1.65$ for both pair members are included). Residues were divided as in Koehl and Levitt (1999): hydrophobic: A, V, I ,L, F, P, M, W, H, Y, C; hydrophilic : G, S, T, N, Q, D, E, K, R .

Structural and functional roles of STAIRS

We conducted an itemized examination of the four significantly over-represented residues and of proline within the group of mutually conserved STAIRS. Although the overrepresentation of proline was not found to be statistically significant within this group of residues it was added to this survey due to its well known unique structural characteristics. Surveying the 133 proteins in our database for positions that are included in the subset of mutually highly conserved STAIRS has yielded 358 such positions, out of which 204 were occupied with one of these five amino acids. These positions were searched for a possible role in determining secondary or super-secondary structure, and/or participation in a functional site. 191 positions could be assigned a functional or structural role (functional sites were either ligand-binding sites or catalytic sites, as annotated in the PDB files, and structural annotations were based on FSSP). 84 residues participated in functional sites or were close in sequence to functional site residues. Among these histidine stood out as the predominant residue identified in functional sites in our database, and glycine and proline were relatively very prevalent near functional sites, probably playing a role in shaping their structure. 58 residues were located at specific positions in secondary structures, either at the N-termini or C-termini of α helices or β strands, or in turns, where glycines were predominant. As expected, 37 out of 39 cysteines were involved in disulfide bonds. This is obviously due to the special role that disulfide bonds play in stabilizing tertiary structures. Eight prolines were located at the termini of β hairpins, stabilizing this super-secondary structure.

Discussion

As more and more protein structures are solved it has become evident that many proteins having a similar structure share only a very small number of identical residues in structurally aligned positions. This has prompted various lines of research, attempting to identify the common hidden information within these sequences that directs them to assume similar folds (Mirny *et al.*, 1998; Koehl and Levitt, 1999; Mirny and Shakhnovich, 1999; Poupon and Mornon, 1999; Ptitsyn and Ting, 1999). The small fraction of identical residues at corresponding positions merit special attention by themselves, as to whether they have been maintained unchanged just by chance, or because of a special role that they play in determining the structure and/or function of these molecules. In the current study we have attempted to examine this issue by characterizing those structurally aligned residues that are identical (STAIRS).

One approach to determine important positions along a protein sequence is to follow their conservation within the members of the corresponding protein family. The conclusions from such an analysis are reinforced when the conservation patterns of structurally aligned positions in the protein families of two remote sequences are consistent, and mutually highly conserved positions are identified. In general, we find a significant positive linear correlation (r=0.3) between the corresponding conservation scores along the sequences of the two pair members. This correlation is higher for STAIRS and highest for NSTAIRS, STAIRS which are spatially close. NSTAIRS are even better candidates for preserving protein traits, as their spatial proximity suggests a role in maintaining structural integrity or a functional site. Indeed we find that the high frequency of mutually well-conserved positions, exhibited to be more prevalent in STAIRS than in the whole database, is even more prevalent in NSTAIRS. Namely, there is a six-fold increase in the frequency of NSTAIRS in the mutually well-conserved paired positions when compared to their frequency in the whole database (Table 1). This leads to the conclusion that certain well placed residues do possess a trait which may contribute to the protein's evolutionary fitness, and therefore remain unchanged.

It should be noted that 39.5% of STAIRS are *not* well conserved. This may provide an estimate as to the fraction of positions that are coincidentally identical. Caution should therefore be taken before attributing significance to positions that contain identical residues in a pair of SSSD proteins. Thus, among the few identical residues in an SSSD pair, even fewer are considered to have biological significance.

Interestingly, about 50% of the positions with mutually high conservation contain different residues in the two proteins. These positions may contain residues that are similar in their properties and fulfill the same requirements in the two structures. Recall that our analysis was done at the residue level, using an alphabet of 20 amino acids. It is conceivable that by using a redundant alphabet, where the residues are clustered by their physico-chemical properties, these positions would have appeared as containing STAIRS. Alternatively, the conserved residues at corresponding positions may be completely different but form a spatial cluster with compatible residues within the respective structure. Such a finding will support the line of Mirny and Shakhnovich (1999) and Ptytsin and Ting (1999) who proposed the existence of corresponding folding nuclei that are composed of different types of residues, but guide folding into similar structures.

Many previous studies, experimental and theoretical, have shown that positions critical for maintaining protein stability are located in buried positions (Cordes *et al.*, 1996). Our burial analysis, depicted in Figure 3, conforms to those observations. From Figure 3 it is evident that the mutually conserved residues have a higher burial proportion (Figure 3b) than the whole database (Figure 3a). Curiously, the burial percentage of STAIRS is very similar to what was found for the whole database (Figure 3a). Two phenomena contribute to this finding: (1) as 40% of STAIRS are coincidental they are expected to be similar in their characteristics to the general population of residues; (2) many buried residues are interchangeable (Lesk and Chothia, 1980; Lim and Sauer, 1989). By definition these are not included in the group of STAIRS, and therefore on the average the burial of STAIRS is similar to that of the whole population of residues. The latter is supported when looking at Figure 3a: the fraction of hydrophobic buried residues among all positions is even larger than within the STAIRS. The NSTAIRS, however, show a higher fraction of buried residues, consistent with their suspected function in maintaining a spatial entity that is important for the proteins' structure. The same logic applies to the pattern of burial depicted in Figure 3b, that regards the mutually highly conserved positonly. The high fraction of buried positions is evident across the three categories of STAIRS, NSTAIRS and all positions. Recall that 50% of the highly conserved residues are not STAIRS, and the burial percentage within them is remarkable.

The analysis of residue types in STAIRS revealed certain residue types that are significantly over-represented (glycine, leucine and valine). However, when the analysis is restricted to the subset of mutually highly conserved STAIRS, the pattern of conservation changes, and leucine and valine are underrepresented. Again, this is probably due to the interchangeability of these residues in the hydrophobic cores. Our conservation analysis, carried out at the residue level, cannot identify them as highly conserved. Other residues predominate the subset of mutually conserved positions: glycine, cysteine, histidine, and tryptophan. Interestingly, these amino acid residues were also shown to be highly conserved in a set of sequences designed to fit a known structural template (Koehl and Levitt, 1999). Also, lysine and threonine that appeared to be underrepresented among the mutually conserved STAIRS were found also to be underrepresented in that set of designed proteins. Thus, the compatibility between our results and those obtained from designed proteins gives further support to the feasibility of the proposed design procedure (Koehl and Levitt, 1999).

The general features of the positions with mutually conserved STAIRS may suggest their involvement in folding nuclei. This is mostly implied by their low solvent accessibility and spatial closeness, but cannot be definitely proved in the scope of this study. Apart from their potential participation in folding nuclei, our explicit survey shows that mutually conserved STAIRS have crucial structural or functional roles. However, the different ways they assume in achieving this end are quite disparate. This pertains not only for STAIRS in general but also when specific amino acids are considered. While there are certain cases where a conserved residue fulfills the same role through all its occurrences in our data set (such as the involvement of cysteine in disulfide bonds), many other STAIRS show a variety of themes. Thus, there is no common underlying role which can be singled out for mutually conserved STAIRS, rather they are either important in functional sites, or in stabilizing secondary and super-secondary structures.

Methods

Construction of the database of SSSD protein pairs The databases of FSSP (Families of Structurally Similar Proteins, Holm and Sander, 1996) and DAPS (Distant Aligned Protein Sequences, 1998, Rice and Eisenberg, http://siren.bio.indiana.edu/daps, 1998) were used as a starting point for the database of 126 SSSD protein pairs. Briefly, FSSP is a database based on an exhaustive all-versus-all structural alignment of proteins in the PDB database (Bernstein *et al.*, 1977). The classification and alignments are automatically maintained and continuously updated using the DALI program (Holm and Sander, 1993). Each FSSP file has a single structural representative, against which all structurally similar proteins are aligned in decreasing order of structural similarity. The DAPS database is based on FSSP and contains alignments of all protein pairs sharing less than 25% identical residues. These pairs of proteins were based on the PDB_SELECT25 list (Hobohm and Sander, 1994).

For generation of the SSSD database the DAPS database was filtered using the following criteria: 1) Minimal protein length of 30 residues for both pair members. 2) Resolution better than 3.5Å for each pair member. 3) Difference in lengths within a protein pair does not exceed 50% of the shorter member. 4) The alignment length is at least 60% of the longer member's length. 5) The sequences of the pair members should not be well aligned using sequence alignment methods. A good sequence alignment, regardless of compatibility with the FSSP structural alignment, denotes a sequence similarity which we wish to avoid. Each pair was checked for similarity using the BESTFIT program from the GCG package (version 10, Genetics Computing Group, WI USA), an implementation of the Smith-Waterman algorithm (Smith and Waterman, 1981). Protein pairs with statistical significant sequence alignments were excluded.

Methods for structural alignment As different methods may yield different structural alignments (Godzik, 1996), we found it necessary to verify that the determined STAIRS are not dependent on the method used for structural alignment. The STAIRS analyzed in the paper are based on the FSSP alignments obtained by DALI. This is a structural alignment method that uses Monte-Carlo optimization to align Cα-Cα distance matrices. We verified that the results based on these alignments are not different from those based on structural alignments obtained by SSAP (Taylor and Orengo, 1989). SSAP describes the proteins to be aligned as a set of all-versus-all Cβ-Cβ vectors. Optimization of the structural alignment is accomplished by a double-dynamic programming method.

Definition of contact residues Contacting residues were defined as residues whose distance between their side chain beta-carbon atoms was equal to or less than 7Å. For glycine, the alpha-carbon atom was used. Residues consecutive in sequence were excluded from the count. In this manner, corresponding clusters of contacting positions were defined in both pair members. For a position to be included in a cluster of NSTAIRS it is required that it would contact at least one position within this cluster. When looking at contacting residues, we have also used the distance between the centroids of the side-chains as a basis for determining contact residues. When analyzing the clusters, the results were virtually identical to those obtained using the beta-carbon criterion (results not shown).

Information Content (IC) calculations For a single position *j* in a multiple sequence alignment the Information Content would be:

$$IC(j)= \sum_{i=1}^{20} p_{ij} * \log(p_{ij} / q_i),$$

where: p_{ij} is the frequency of residue *i* at position *j*, and q_i is the frequency of residue *i* in the database. Zero frequencies and gaps in the alignment were treated as in Hertz and Stormo (1995; http://www.bionet.nsc.ru/bgrs /thesis/56/index.html).

For each position *j* a normalized value of IC(*j*) was calculated by Z_{ic} (*j*)= (IC(*j*)- $\overline{IC}$)/S_{IC}, where $\overline{IC}$ is the mean of IC(*j*) along the sequence and S_{IC} is its standard deviation

Z_{ic} Distribution By examining the Z_{ic} score distribution in the whole database it was verified that $Z_{ic}\geq1.65$ defines the high 5% scoring positions. However, when building the set of mutually highly conserved STAIRS, the selection criterion was that both aligned positions should have a Z_{ic} above the threshold. By this, high scoring residues that were not aligned with other high scoring residues were excluded, reducing effectively the frequency of mutually highly conserved positions to 3%.

Solvent accessibility Solvent accessibility values in Å^2, were taken from the FSSP database. For each residue, these were divided by the accessible surface area of the extended conformation of that residue (Miller *et al.*, 1987), and expressed in percentages. The analysis was carried out both by using these values and followed by determining the burial threshold at 25% surface accessibility.

Hardware and software Work was performed on Silicon Graphics Indy and Indigo2 workstations, using C and Python languages. The Python PDB handling package used for this work is available on http://www.ls.huji.ac.il/~idoerg/PyStruct.tar.gz

References

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25:3389-3402.

Anfinsen, C.B. 1973. Principles that govern the folding of protein chains. *Science* 181:223-230.

Bernstein, F.C., Koetzle, T.F., Williams, G.J., Meyer, E.E., Jr., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M. 1977. The Protein Data Bank: a computer-based archival file for macromolecular structure. *J Mol Biol* 112:535-542.

Cordes, M.H., Davidson, A.R. and Sauer, R.T. 1996. Sequence space, folding and protein design. *Curr Opin Struct Biol* 6:3-10.

Godzik, A. 1996. The structural alignment between two proteins: is there a unique answer? *Protein Science* 5:1325-1338

Hobohm, U. and Sander, C. 1994. Enlarged representative set of protein structures. *Protein Science* 3:522-524.

Holm, L. and Sander, C. 1993 Protein structure comparison by alignment of distance matrices *J Mol Biol* 233:123-138

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

Koehl, P. and Levitt, M. 1999. De novo protein design. II. Plasticity In sequence space. *J Mol. Biol* 293:1183-1193.

Lesk, A.M. and Chothia, C. 1980. How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. *J Mol Biol* 136:225-270.

Lim, W.A. and Sauer, R.T. 1989. Alternative packing arrangements in the hydrophobic core of lambda repressor. *Nature* 339:31-36.

Markiewicz, P., Kleina, L.G., Cruz, C., Ehret, S. and Miller, J.H. 1994. Genetic studies of the lac repressor. XIV. Analysis of 4000 altered Escherichia coli lac repressors reveals essential and non-essential residues, as well as "spacers" which do not require a specific sequence. *J Mol Biol* 240:421-433.

Milla, M.E., Brown, B.M. and Sauer, R.T. 1994. Protein stability effects of a complete set of alanine substitutions in Arc repressor. *Nat Struct Biol* 1:518-523.

Miller, S., Janin, J., Lesk, A.M. and Chothia, C. 1987. Interior and surface of monomeric proteins. *J Mol Biol* 196:641-656.

Mirny, L.A., Abkevich, V.I. and Shakhnovich, E.I. 1998. How evolution makes proteins fold quickly. *Proc Natl Acad Sci U S A* 95:4976-4981.

Mirny, L.A. and Shakhnovich, E.I. 1999. Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. *J Mol Biol* 291:177-196.

Park, J., Karplus, K., Barrett, C., Hughey, R., Haussler, D., Hubbard, T. and Chothia, C. 1998. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J Mol Biol* 284:1201-1210.

Poupon, A. and Mornon, J.P. 1999. Predicting the protein folding nucleus from a sequence. *FEBS Lett* 452:283-289.

Ptitsyn, O.B. and Ting, K.L. 1999. Non-functional conserved residues in globins and their possible role as a folding nucleus. *J Mol Biol* 291:671-682.

Rennell, D., Bouvier, S.E., Hardy, L.W. and Poteete, A.R. 1991. Systematic mutation of bacteriophage T4 lysozyme. *J Mol Biol* 222:67-88.

Rost, B. 1997. Protein structures sustain evolutionary drift. *Fold Des* 2:S19-24.

Rost, B. 1999. Twilight zone of protein sequence alignments. *Protein Eng* 12:85-94.

Salamov, A.A., Suwa, M., Orengo, C.A. and Swindells, M.B. 1999. Combining sensitive database searches with multiple intermediates to detect distant homologues. *Protein Eng* 12:95-100.

Smith, T.F. and Waterman, M.S. 1981. Identification of common molecular subsequences. *J Mol Biol* 147:95-197.

Suckow, J., Markiewicz, P., Kleina, L.G., Miller, J., Kisters Woike, B. and Muller Hill, B. 1996. Genetic studies of the Lac repressor. XV: 4000 single amino acid substitutions and analysis of the resulting phenotypes on the basis of the protein structure. *J Mol Biol* 261:509-523.

Taylor, W.R., and Orengo, C.A. 1989, Protein structure alignment. *J Mol Biol* 208:1-22.

Abbreviations

df, degrees of freedom; SSSD, Structurally Similar Sequence Dissimilar; STAIRS, STructurally Aligned Identical ResidueS; NSTAIRS, Neighboring STAIRS.