Supporting Text

Methods.

Sequence and structure data. Structural domain definitions for the seed structures of the evolutionary profiles (EPs) were taken from the SCOP database, version 1.65 (2). The ASTRAL database (3), which mirrors the Protein Data Bank (PDB), divides each PDB chain into separate files corresponding to each domain, and the ASTRAL PDB-style files were used as a source of coordinates. If the SCOP domain definition was not available for a particular protein, we used a structural alignment to the most closely related homolog available to define a SCOP-like domain from the PDB chain. A QR factorization-based representative set of structures was derived from all the available structures as described in (1, 4).

Sequences in the EPs were extracted from the well curated Swiss-Prot database (5). The sequences from structure representatives in each of the EPs were used in a BLAST search (6) over the Swiss-Prot database to find the set of all close sequence homologs with E-value $< 10^{-5}$. More distant sequences within the homologous group that could not be found using the initial BLAST search were found by using sequence hits from the initial search as seeds in a second BLAST search. In these searches, enzyme names were used to find the appropriate E-value cutoffs, i.e., all top Blast hits were included until the search hit proteins outside of the homologous group as indicated by their enzyme names, giving typical E-value cutoffs in the range 10^{-3} to 10^{-1}. For example, the only known structure of the PheRS α-chain is from the bacterial organism *Thermus thermophilus* (PDB code 1PYS). The initial BLAST search readily identified all known bacterial orthologs, but only recovered a few of the known archaeal and eukaryotic sequences. In a second round of BLAST, representatives of the eukaryotic and archaeal hits were used queries to retrieve additional orthologs from these domains of life. Sequence alignments of the proteins from Swiss-Prot to the closest homolog with a known structure were used to provide structural domain definitions for the protein sequences, and a representative set of these sequence domains were used to supplement the structural alignments.

Generation of multiple alignments. Structural alignments were computed using the multiple structural alignment program STAMP (7), which uses a dynamic programming procedure in combination with linear-least squares fitting to find the rigid body rotation that simultaneously minimizes the C_α-C_α distance and local main chain conformation for each pair of aligned proteins. While the algorithm does not include sequence dependent information, it uses a progressive multiple alignment procedure with a hierarchical clustering analysis based on structural similarity. The quality of the resultant multiple structural alignment depends to some degree on a set of initial alignments that STAMP computes by "scanning" a selected protein domain against all others in the data set. In difficult alignment cases, e.g., distantly related or highly symmetrical structures, we developed a heuristic algorithm to attempt each scan domain and take the initial alignments from the scan domain that produced the highest alignment scores on average. The initial alignments were executed with the following STAMP parameters: `-scan true -npass 2 -slide 5 -scanscore 6`, and the final multiple structural alignments were computed with default parameters. The original version of STAMP systematically misaligns N- and C-terminal residues, but this has been repaired and will be made freely available through a new multiple structural alignment feature in the next release of the molecular visualization program VMD version 1.8.3 (8). There are fluxional regions in the structures of some proteins, e.g., mobile loops, and the structural alignment of these regions depends on the conformation in which the fluxional region was crystallized. These portions of the alignments were manually corrected to improve the quality of the alignment by comparing them with the sequence-based alignments of closely related sequences, i.e., those with sequence identities of >30%. The structure-based QR factorization (4) was used to select the representatives of known structure from the multiple structural alignment of all known structures in each given protein family or superfamily.

Currently, structural information is not available for each evolutionarily distinct subgroup in a protein family. Sequences were selected to supplement the structural alignments to generate complete evolutionarily profiles for each homologous group of proteins. All sequence alignments were performed with the progressive multiple alignment procedure in CLUSTAL W (9). The progressive alignment method does not ensure the optimal alignment except for a small number of sequences, so the resulting alignments were manually checked and adjusted in regions that were clearly mis-aligned. The alignment quality reduces for diverse sequences when the alignment is based on sequence-based methods alone. Sequence alignments were performed for subgroups, with pairwise sequence percentage identity at >30%. For the AARSs, this cutoff corresponds to homologous groups with the same enzymatic specificity. The class II AARS family, therefore, was divided into 12 subgroups while the lipocalins superfamily was divided into 15 subgroups. Profile-to-profile alignments in CLUSTAL W were used to align proteins from different domains of life for groups displaying canonical or basal canonical phylogenetic patterns. The multiple sequence alignments were generated with the CLUSTAL W parameters: `-infile= -align -outfile=` for multiple alignments or `-profile1= -profile2= -profile -outfile=` for profile-to-profile alignments. Sequence-based QR was applied to these subgroups individually. So that the more accurate structural alignment information would be included in the evolutionary profiles, a constraint was implemented in the QR factorization, see Theory and Methods, that ensured the representative proteins of known structure were retained as members of the EP. The resulting sequence representatives are added to the structure-based profile as discussed below.

Except in the case of the HisA-HisF family, in which the alignments were purely sequence-based, the final alignments used to construct the evolutionary profiles were based on the combination of sequence and structure representatives. The QR factorization-based sequence representatives were added to the structural alignments using the profile-to-sequence alignment method of CLUSTAL W (9). The CLUSTAL W parameters used for supplementing the structural alignments (profile 1) with sequences (profile 2) was: `-profile1= -profile2= -sequences`. If required, these alignments were manually corrected.

Phylogenetic analysis. Trees were drawn using either the neighbor-joining program in Phylip (10) or the unweighted pair group method using arithmetic averages (UPGMA) (11), as implemented in MATLAB 6.5 (Mathworks, Natick, MA). Distance-based sequence trees were computed using sequence identity, normalized by the length of the shortest sequence, as well as sequence similarity according to the PAM substitution model in protdist (10). Simple distance-based phylogenetic methods were used because the QR factorization is based on an orthogonal encoding of the amino acids and is thus most similar to the pairwise sequence identity metric. Amino acids can be numerically encoded, using artificial neural networks, to reflect the amino acid similarities in a particular substitution matrix, as in (12, 13), but we leave this application as well as the application of more sophisticated models of protein evolution, such as maximum-likelihood phylogenetic analysis, for future work. We have previously demonstrated a congruence between sequence and structure-based phylogenies (1, 4) with the application of a measure of structural similarity between homologous proteins, Q_H, which was used to construct the structure-based phylogeny shown in Figure 2.

Database searches. BLAST (6) and HMMER (http://hmmer.wustl.edu/) (14) were used for database searches over the Swiss-Prot, TrEMBL and the National Center for Biotechnology Information's NRDB (15). BLAST uses a position specific scoring matrix (PSSM) profile while HMMER gives a probabilistic treatment, based on a hidden Markov model, for finding amino acids in a given position of the alignment. The main difference between the PSSM-based BLAST approach and the probabilistic HMMER approach is the position-dependent gap penalties in HMMER. HMMER requires approximately 100-fold more computational time than BLAST for an identical database search. The performance of the EPs is typically compared to that of the corresponding Pfam profiles. The Pfam profiles were downloaded from the Pfam 15.0 database (http://pfam.wustl.edu/) (16), and they are computed based on two alignments; the seed alignment and the full alignment, a redundant alignment formed from all the proteins belonging to a particular family in the Swiss-Prot and TrEMBL databases (5).

Homology modeling for putative class II CysRS in a methanogenic archaeon. Database searches using HMMER (14) of the genomes of the archaeal organisms *M. jannaschii* (17), *M. thermoautotrophicus* and *M. kandleri* obtained from Swiss-Prot (5) with the single complete evolutionary profile (QR 40% sequences and structure) for the class II AARS confidently placed the sequences YG60_METJA (in *M. jannaschii*), O27545 (in *M. thermoautotrophicus*), and Q8TY66 (in *M. kandleri*) among the hits for the 10 class II AARSs expected in these genomes. These sequences were found within the trusted cutoff for the profile with the YG60_METJA having an E-value score between those for PheRS α-chain and GlyRS α-chain. The PheRS β-chain and AlaRS are found in the database search with an E-value score greater than that for GlyRS α-chain. The program Modeller (18) was used to build a homology model of the putative class II CysRS, YG60_METJA, based on the structure of the α-chain of PheRS from *T. thermophilus*, PDB code 1pys chain A, with loop optimization set at the maximum level. The initial alignment was performed using the profile-to-profile method of CLUSTAL W of the putative class II CysRS group to the structural alignment of PheRS α-chain, PheRS β-chain, GlyRS α-chain, and AlaRS. These four synthetases form a group that are structurally similar. This alignment was manually modified based on the agreement of secondary structure prediction from PSIPRED (19) and the secondary structure elements in the structural alignment. A number of insertions, specific to the putative class II CysRS group, were found which are not present in the α-chain PheRSs. The larger insertions were removed from the final model as these regions do not have an appropriate template for modeling.

Householder transformation. QR factorization performs a series of orthogonal transformations designed to convert the matrix A to upper triangular form. The Householder transformation is based on the reflection of a given vector such that the desired component(s) of the vector are removed. In other words, while transforming to an upper triangular form, the components of the vector below the diagonal are annihilated (made equal to zero). The Householder transformation H_1 described in Figure 5, is constructed for the vector $\boldsymbol{a}$ such that $H_1\boldsymbol{a} = \boldsymbol{a} - \left(2\frac{\boldsymbol{v}^T\boldsymbol{a}}{\boldsymbol{v}^T\boldsymbol{v}}\right)\boldsymbol{v}$ in which $\boldsymbol{v} = [0\ \ a_2]^T - \alpha\,[1\ \ 0]^T$ and $\alpha = -sign\,(a_1)\,\|a_2\|_2$. For a detailed description of the Householder transformation see (20). This transformation eliminates all of the below diagonal elements of $\boldsymbol{a}$, but preserves the Euclidean norm of $\boldsymbol{a}$ by moving all of the magnitude of $\boldsymbol{a}$ into the component a_1'. The geometric and matrix interpretations of the Householder transformation are shown in Fig. 5.

The model problem shown in Fig. 5 can be thought of as a multiple alignment of three one-dimensional "proteins", each two residues in length with no gapped positions. The pivoting operation is performed to find the "protein" which is most linearly independent of the representative set already formed. This is done after the kth step by finding the protein with the maximum Euclidean norm (or a Frobenius-like p-norm for the multidimensional representation described in the text) of the sub-matrix. The cosine of the angle between two proteins is proportional to the dot product of the two proteins. The angle between two proteins is hence a measure of the number of identical residues that the two proteins have in the same position of the multiple alignment. An alternative formulation of linear independence can be in terms of the angle that the "protein" makes with the space of the representative set already formed. In this case, given that $\boldsymbol{a}$ is chosen as the first representative, $\boldsymbol{c}$ would be chosen as the second representative protein. An advantage of the QR factorization, especially in the higher dimensional representation for real proteins, is that it uses linear independence from the space formed by k proteins to choose the $(k+1)$th protein. While pairwise comparisons are a good measure of linear independence of a pair of vectors, the linear independence of three or more vectors can not be measured just by pairwise comparisons.

Parameter search. The goal of the QR algorithm is to provide an ordering of the sequences in a multiple alignment, ranked by their decreasing linear independence of the previous proteins in the ordering. The ordering can be used to define a non-redundant and representative subset based either on a arbitrary identity threshold or on a desired number of representative sequences. The isoleucyl-tRNA synthetases (IleRSs), for example, are known to display the canonical phylogenetic pattern (21), indicating distinct, well separated and largely monophyletic groups of the Archaea, Eucarya, and Bacteria. The canonical pattern is so robust among the IleRSs that as well as being documented with sequence signature and maximum-likelihood phylogenetic analysis (21), it is also apparent in a simple UPGMA dendrogram, based on sequence identity, as shown in Fig. 6*a*. In this case, therefore, an unbiased, yet representative profile would be three or multiples of three, with equal number of members spanning all the evolutionarily distinct groupings and the QR order must represent this. However, the number of sequences of the IleRS in the Swiss-Prot database is biased towards the bacterial domain of life with 32 of the 49 proteins belonging to bacterial organisms. For the test case shown in Fig. 6*a*, we have taken equal number of proteins belonging to the three domains of life. At a percentage threshold value of 40%, there are three distinct groupings with the three domains of life being represented once and the representative set taken with 40% threshold should have three proteins, one from each domain of life. This is an example of a general property that the QR ordering must obey: at any similarity threshold the maximum number of sequences, with pairwise similarity values below threshold, are chosen as the members of the representative set. If the QR factorization produces an ordering which obeys the above property for any arbitrary similarity threshold, then the ordering is said to be "allowed."

To visualize this property, we introduce the notion that a phylogeny or dendrogram has distinct threshold regions, which are defined by the emergence of a new branch as the threshold increases along the sequence identity scale. The dendrogram in Fig. 6*a* has eight distinct threshold regions. The first distinct threshold can be applied in the range from the initial bifurcation, which splits the sequences into two groups, i.e., (1,4,8) and (2,6,7,3,5,9), to the second bifurcation, which splits the latter group into two subgroups, (2,6,7) and (3,5,9). The range for the second threshold begins with the formation of these three groups and ends at the branch which splits group (1,4,8) into two groups, (1) and (4,8), and the other threshold regions are defined similarly. In summary, each region is associated with a range of cutoff values, represented by $I_{\text{cut}}i$ where the ith threshold separates the proteins into i groups. When the threshold $I_{\text{cut}}i$ is applied, the QR order is defined as allowed if the first i proteins in the QR order represent the i different groups defined at that cutoff, for all i. All other orderings, which can be obtained if, for example, the gaps are weighted too heavily in comparison to the aligned positions, are defined as forbidden. The two adjustable parameters, one of which is the gap scaling parameter, must be defined, therefore, such that the QR factorization gives allowed orderings for all possible thresholds. Fig. 6*a* and *b* show dendrograms for the training sets used to parameterize the QR factorization, and in both trees the sequences are numbered according to an allowed QR ordering. In Figure 6*a*, for example, when $I_{\text{cut}}2$ is applied to the ordering, proteins 1 and 2 represent groups (1,4,8) and (2,6,7,3,5,9), respectively. Similarly, when $I_{\text{cut}}3$ is applied, proteins 1, 2, and 3 represent groups (1,4,8), (2,6,7), and (3,5,9) and so on for all distinct thresholds up to $I_{\text{cut}}9$. This ordering is allowed for all distinct thresholds and is, thus, an allowed ordering.

Our implementation of the multidimensional QR factorization of the alignment matrix depends on two parameters, the gap scaling constant γ and the ordering p-norm. In order to determine an appropriate value for these parameters, two training sets were used, which exhibited well-defined phylogenetic topology, implying a phylogeny that is robust with respect to different tree methods [UPGMA, neighbor joining or maximum likelihood (21)] and distance metrics (similarity or identity). The parameters can also be tuned to give allowed orderings for trees based on more sophisticated phylogenetic methods. Sequences of first training set were selected from the IleRSs, which display the Rossmann fold, with a three layer α-β-α topology and a core made from five parallel β-strands. The proteins of the second training set were selected from the lipocalin fold type, which is a β-barrel made from anti-parallel β-sheets. The second training set was made from a superfamily of proteins while the first training set was made from the closely related IleRS. The second training set has many gaps while the first training set has very few gaps.

The results of the parameter search are given in Fig. 6*c* and *d*. QR factorization is carried out for each pair of candidate parameters over the given ranges. At each value the of candidate parameters, the QR ordering is evaluated to be either forbidden or allowed. The parameter space plots show a large region of overlap where allowed orderings are consistently achieved for both training sets. At low γ values, the gaps are assigned small weights in comparison with the weights assigned to the amino acids, resulting in forbidden orderings for the superfamily level alignments of the lipocalins. In general, the alignment of proteins in the lipocalin superfamily contains more gaps than the IleRS training set. At high gap weights ($\gamma > 2$), the pattern of gaps dominates the QR ordering and forbidden orderings result for the more closely related sequences of the IleRS training set. In general, the superfamily alignment lipocalins favor large gap penalties while the family alignment from IleRS favor smaller gap penalties. The parameters used for this study are taken from the overlapping region and are $\gamma = 1.0$ and $p = 2$. Note that these values are optimal for an orthogonal encoding of the amino acids and can change when the description of the amino acid is altered.

Further Results.

Phylogenetic tree of HisA-HisF family and performance of family profiles with HMMER search. In Fig. 7, the major evolutionarily distinct groups are labeled with brackets, and the thirteen members of the representative set, which form the best performing EP, i.e., QR 40% in Fig. 1, are labeled by their organism name and their rank in the QR order (purple numbers). Although the group of putative, duplicated HisF proteins appears over-sampled, the relationships in this group are more distant than are the relationships between sequences within the other subgroups. The tree includes all members of the HisF-HisA family found in Swiss-Prot, and the bias towards the bacterial organisms in this family is evident.

Fig. 8 shows ROC50 plots for HMMER searches over the Swiss-Prot database for a variety of profiles of the HisA-HisF family. The database search results show that for closely related families, BLAST performs comparably to HMMER. A profile search with HMMER is, however, 100 times more computationally expensive than the corresponding BLAST profile search.

A superfamily level profile of the lipocalins. Proteins of the lipocalin superfamily transport ligands docked to a buried hydrophobic pocket inside its β-barrel fold and have been intensely studied in protein folding experiments. The lipocalin fold has a single superfamily composed of three member families: the retinol binding protein-like family, fatty acid protein-like family, and the thrombin inhibitor family. The diversity of the proteins in the lipocalin superfamily is so large that the sequences cannot be aligned properly by sequence-based methods alone. The members of the lipocalin superfamily can be aligned properly using structure-based information as shown in Fig. 10. From the currently available structures, however, a purely structure-based EP of 20 sequences was constructed. For the lipocalin member subfamilies lacking structure data, QR factorization-based sequence representatives were used to supplement the multiple structure alignment. With a 50% sequence identity threshold applied to the QR ordering of the combined sequence-structure alignment, 54 proteins are required to span the evolutionary space of the lipocalin superfamily. In order to examine the effect of the enhanced alignment quality provided by the structure-based alignment on the database search results, a third EP was built in which these 54 sequences were re-aligned using CLUSTAL W (9). As shown in Fig. 9, once again the sequence supplemented EP outperforms both the Pfam full and seed alignment-base profiles, and does so with 3-fold fewer sequences than the Pfam seed profile. Interestingly, while the EP based on the sequence supplemented structure alignment performs the best, comparable performance is observed in the EP based on the same sequences, but realigned with CLUSTAL, which finishes a near second. In a study by Panchenko and Bryant (22), it was shown that the main advantage of using a profile based on a higher quality seed-alignment is the improved accuracy of the alignment of the databases sequences to the profile.

SUPERFAMILY is a database of hidden Markov models (HMM) profiles for each of the superfamilies in SCOP and is similar in spirit to Pfam (23). For each entry in the database fifty HMMs are given, each one built from a different seed sequence. The SUPERFAMILY protocol suggests that for a single superfamily level database search, fifty separate database searches should be performed and the results summed. Although Gough *et al.* (23) have documented some improvement with this protocol, here we are testing the efficacy of a single profile in a single database search. The first lipocalin superfamily HMM (model no. 0016667) in the list of fifty was chosen for comparison to the EPs. In addition to the HMMs, the SUPERFAMILY database provides multiple alignments of the sequences of a representative set of structures belonging to the superfamily. These sequences are selected on a sequence identity threshold of 40% and the different alignments are based on each of the fifty HMMs. Performance of a profile based on this alignment, corresponding to model no. 0016667, is also shown in Fig. 9. Neither SUPERFAMILY profiles perform as well as the EP based on only the known structures in a single database search. A depiction of structural conservation among the lipocalin superfamily members is shown in Fig. 10.

Class II AARS overlap and performance of EPs with BLAST. The class II AARSs are a diverse family of proteins that cannot be aligned using sequence-based methods alone. The class II AARSs, however, share significant structural similarity and can be aligned using structure-based methods. In Fig. 11*a*, the structural cores of the class II AARSs are shown. The structural core is measured by using a metric of structural similarity at each site which is explained in more detail in ref. 3. The class II AARSs have been divided in to three subclasses based on structural similarity (1). The class II AARSs belong to the $\alpha + \beta$-fold with a prominent antiparallel β-sheet.

In Fig. 11*a*, the secondary structure elements of the class II AARSs are clearly visible. The structural core of the subclass IIA AARSs, which includes the dimeric GlyRS, HisRS, ProRS, SerRS, and ThrRS, is shown in Fig. 11*b*. In addition to the major secondary elements of the class II AARSs, a small β-hairpin (top of Fig. 11*b*) and a small β-sheet (bottom of Fig. 11*b*) are also conserved among the subclass IIA AARSs. These two structural elements seem to be an idiosyncratic insertion which is conserved at the more closely grouped subclass level and is not conserved at the diverse family level of the class II AARSs. Similarly, Fig. 11*c* shows the structural core of the class IIB AARSs, comprising AARSs specific for aspartate (D), lysine (K) and asparagine (N). In addition to the common structural core of the class II AARSs, a small helix is conserved in the class IIB AARSs (top of Fig. 11*c*). Finally, Fig. 11*d* displays the structural core of the subclass IIC AARSs, that are all uniquely tetrameric and specific for glycine (G), alanine (A) and phenylalanine (F). Interestingly, there is a deletion of a β-hairpin in some members of the subclass IIC group that is highly conserved in all other class II AARS (top-right corner of Fig. 11*a*). This β-hairpin is involved in the dimerization of the class IIA AARSs and class IIB AARSs and is not conserved among the tetrameric class IIC AARSs.

1. O'Donoghue, P & Luthey-Schulten, Z. (2003) *Microbiol. Mol. Biol. Rev.* **67**, 550–573.

2. Murzin, A. G., Brenner, S. E., Hubbard, T, & Chothia, C. (1995) *J. Mol. Biol.* **247**, 536–540.

3. Chandonia, J. M., Hon, G., Walker, N. S., Lo Conte, L., Koehl, P., Levitt, M, & Brenner, S. E. (2004) *Nucleic Acids Res.* **32**, D189–D192.

4. O'Donoghue, P & Luthey-Schulten, Z. (2005) *J. Mol. Biol.* **346**, 875–894.

5. Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O'Donovan, C., Phan, I, & *et al.* (2003) *Nucleic Acids Res.* **31**, 365–370.

6. Altschul, S. F., Gish, W., Miller, W., Myers, E. W, & Lipman, D. J. (1990) *J. Mol. Biol.* **215**, 403–410.

7. Russell, R. B & Barton, G. J. (1992) *PROTEINS: Struc. Funct. Genet.* **14**, 309–323.

8. Humphrey, W., Dalke, A, & Schulten, K. (1996) *J. Mol. Graphics* **14**, 33–38.

9. Thompson, J. D., Higgins, H. G, & Gibson, T. (1994) *Nucleic Acids Res.* **22**, 4673–4680.

10. Felsenstein, J. (1989) *Cladistics* **5**, 164–166.

11. Sokal, R. R & Michener, C. D. (1958) *Univ. Kans. Sci. Bull.* **28**, 1409–1438.

12. Jagla, B & Schuchhardt. (2000) *Bioinformatics* **16**, 245–250.

13. Lin, K., May, A. C, & Taylor, W. R. (2002) *J. Theor. Biol.* **216**, 361–365.

14. Eddy, S. R. (1998) *Bioinformatics* **14**, 755–763.

15. Pruitt, K. D., Tatusova, T, & Maglott, D. R. (2003) *Nucleic Acids Res.* **31**, 34–37.

16. Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L. L, & *et al.* (2004) *Nucleic Acids Res.* **32**, D138–D141.

17. Bult, C. J., White, O., Olsen, G. J., Zhou, L., Fleischmann, R. D., Sutton, G. G., Blake, J. A., FitzGerald, L. M., Clayton, R. A., Gocayne, J. D, & *et al.* (1996) *Science* **273**, 1058–1073.

18. Marti-Renom, M. A., Stuart, A., Fiser, A., Sanchez, R., Melo, F, & Sali, A. (2000) *Annu. Rev. Biophys. Biomol. Struct.* **29**, 291–325.

19. Jones, D. T. (1999) *J. Mol. Biol.* **292**, 195–202.

20. Heath, M. T. (2002) *Scientific Computing: An Introductory Survey.* (McGraw-Hill, New York), 2nd edition.

21. Woese, C. R., Olsen, G., Ibba, M, & Söll, D. (2000) *Microbiol. Mol. Bio. Rev.* **64**, 202–236.

22. Panchenko, A. R & Bryant, S. H. (2002) *Protein Sci.* **11**, 361–370.

23. Gough, J., Karplus, K., Hughey, R, & Chothia, C. (2001) *J. Mol. Biol.* **313**, 903–919.