doi:10.1016/j.jmb.2004.11.053 J. Mol. Biol. (2005) xx, 1–20

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Evolutionary Profiles Derived from the QR Factorization of Multiple Structural Alignments Gives an Economy of Information

Patrick O'Donoghue and Zaida Luthey-Schulten*

Department of Chemistry
University of Illinois at
Urbana-Champaign
600 S. Mathews, Urbana
IL 61801, USA

We present a new algorithm, based on the multidimensional QR factorization, to remove redundancy from a multiple structural alignment by choosing representative protein structures that best preserve the phylogenetic tree topology of the homologous group. The classical QR factorization with pivoting, developed as a fast numerical solution to eigenvalue and linear least-squares problems of the form $\boldsymbol{A}x=b$, was designed to re-order the columns of $\boldsymbol{A}$ by increasing linear dependence. Removing the most linear dependent columns from $\boldsymbol{A}$ leads to the formation of a minimal basis set which well spans the phase space of the problem at hand. By recasting the problem of redundancy in multiple structural alignments into this framework, in which the matrix $\boldsymbol{A}$ now describes the multiple alignment, we adapted the QR factorization to produce a minimal basis set of protein structures which best spans the evolutionary (phase) space. The non-redundant and representative profiles obtained from this procedure, termed evolutionary profiles, are shown in initial results to outperform well-tested profiles in homology detection searches over a large sequence database. A measure of structural similarity between homologous proteins, Q_{H}, is presented. By properly accounting for the effect and presence of gaps, a phylogenetic tree computed using this metric is shown to be congruent with the maximum-likelihood sequence-based phylogeny. The results indicate that evolutionary information is indeed recoverable from the comparative analysis of protein structure alone. Applications of the QR ordering and this structural similarity metric to analyze the evolution of structure among key, universally distributed proteins involved in translation, and to the selection of representatives from an ensemble of NMR structures are also discussed.

Keywords: protein structure profiles; evolution; non-redundant set; aminoacyl-tRNA synthetase; OB-fold

*Corresponding author

Introduction

From the complex dynamic interplay of biological molecules life somehow emerges and creates the organizational structures that separate living organisms from their abiotic environment. Molecular biology attempts to understand the fundamental nature of biological organization by addressing questions concerned with how biological organization is maintained, how it evolved and how it continues to evolve. William Astbury, one of the early proponents of this field, argued that molecular biology "is concerned particularly with the *forms* of biological molecules and with the evolution, exploitation and ramification of these forms in the ascent to higher and higher levels of organization".[1] The comparative analysis of information derived from molecular biological form, which includes, among other characters, the sequence of genes and genomes and the sequence and three-dimensional structure of gene products, has taken on a prominent role in understanding the evolution and function of biological molecules.

Zuckerkandl & Pauling introduced the notion

Abbreviations used: HMM, hidden Markov model; hQR, hierarchical QR; HGT, horizontal gene transfer; IF, initiation factor.

E-mail address of the corresponding author: zan@uiuc.edu

that molecular sequences contain evolutionary information which can be accessed by comparative analysis.[2] Woese and his colleagues were "the first to fully exploit the full power of molecular phylogenetics"[3] by comparing sequences of the ubiquitous and slowly evolving small subunit ribosomal RNA to produce the first reliable universal phylogeny, depicting the evolutionary relationships of all life on Earth, and to discover a third unexpected primary division of life, which came to be called the Archaea.[4] This work restructured our understanding of the evolution of living organisms and fulfilled Darwin's prophetic statement that "the time will come, I believe, though I shall not live to see it, when we shall have very fairly true genealogical trees of each great kingdom of Nature".[5]

Initially, molecular phylogenies were based on pairwise sequence alignments. Feng & Doolittle introduced the progressive multiple alignment algorithm and showed that phylogenetic trees derived from multiple sequence alignments were in better accord with the expected taxonomic groupings of higher eukaryotes.[6] The structure and organization of biological form was created by an evolutionary process, and multiple alignments are data types that better represent this process than pairwise alignments. In the statistical analysis of proteins from comparative studies, the paradigm of using a database of pairwise comparisons is shifting to using a database of multiple alignments of evolutionarily related groups.[7–9] Gribskov *et al*.[10] were the first to formulate a statistical representation of a multiple alignment, known as a profile. In their most basic form, profiles encode multiple alignments by recording position-specific amino acid substitution and gap probabilities. With the construction of the Pfam database, this method was subsequently extended by representing multiple alignments as hidden Markov model (HMM) profiles and by providing a comprehensive database of multiple alignments and the resulting HMMs for all known protein domain families.[11] Profile-to-profile alignment techniques have been very successful in locating unexpected homologies in protein domain substructure,[12,13] and the inclusion of secondary and tertiary structure information has been shown to give profiles with increased sensitivity for remote homology detection.[14] Recently, O'Sullivan *et al*. have developed an automated method to combine sequence and structure profiles.[15] In the realm of protein structure prediction, CASP5 clearly showed that a combination of multiple structure and sequence alignments and derived profiles most accurately capture the sequence-structure patterns in related and especially in distantly related protein groups.[16,17] Multiple alignments reveal patterns of sequence or structure conservation and variability among groups of homologous molecules and are useful in many applications, including: determining which residues are important for structure,[18] stability, function or folding;[19] constructing knowledge-based potentials for protein structure prediction and design, for reviews see Hardin *et al*.[20] Russ & Ranganathan,[21] generating profiles for genome annotation[11] and for constructing molecular phylogenies.

Each of these examples, as well as many others not mentioned, relies on the composition of the multiple alignment and derived profiles, but which and how many sequences or structures should be included? A key issue in almost every area of bioinformatics, and especially concerning multiple alignments, is redundancy. Redundancy in the molecular data stems from research bias and the amenability of the system to laboratory conditions. In the area of genome sequencing, most organisms that inhabit the biosphere are microbes and most of those cannot be cultured in the laboratory.[22,23] Although environmental sequencing methods are now beginning to ameliorate this particular source of research bias,[3,24,25] we must continually remind ourselves that only approximately 1% of extant microbial organisms have been identified.[23]

While the emerging, widespread use of multiple alignments has put bioinformatics on an evolutionary footing to some extent, the problem of redundancy and representativeness in multiple alignments has yet to be addressed in the context of evolution. What should a non-redundant set represent? Certainly the set should not merely represent the database, which we already know to be biased, rather the non-redundant subset should best represent the evolutionary history of a group of homologous molecules; the subset should span the evolutionary space. In large databases, which serve to collect and curate structure and sequence data, redundancy has been addressed by application of a sequence identity threshold.[26–28] The algorithm implied by a sequence identity threshold typically involves all-on-all pairwise alignments between sequences in the database followed by computation of sequence identity values. Of any pair of sequences that are above threshold, one of the pair is arbitrarily chosen, or chosen under the constraint of some heuristic such as minimum sequence length or, for protein structures, lowest crystallographic resolution, and removed. Although an arbitrary decision at each step of this algorithm is clearly undesirable, the method is intended to be used with large collections of sequences or structures in order to give an approximately non-redundant set of molecules. Chothia and co-workers recently showed that a non-redundant database culled at 50% identity has a higher effective information content than the full database.[29] Databases of multiple alignments, such as HOMSTRAD[9] and Pfam,[11] use either sequence identity cutoffs or sequence weighting to address redundancy and representativeness. Figure 10 compares the composition of the non-redundant set or seed alignment of Pfam to that derived using our evolutionary based method detailed below for a subclass of aminoacyl-tRNA synthetases. Note in this example that while sequence weighting can account for bias due to

over or under-representation, it cannot account for missing data.

Even though sequence identity cutoff algorithms, because of their speed, are sensible for a large database of sequences, the arbitrariness of the algorithm is particularly problematic in the context of a multiple alignment. Sequence weighting methods,[30,31] which give proportionally more weight to outliers and less to more "common" sequences, suffer from the lack of a general theory for weighting, and weighting schemes give no assurance that the representative sequences span the evolutionary space, see Figure 10. May recently presented an algorithm which partitions sequences from a multiple alignment into, typically two, maximally divergent clusters based on a sequence entropy measure over aligned regions.[32] Although the method is of interest, ignoring the effect of gaps is highly undesirable, especially for distantly related protein groups, and the method is not readily applicable to treating redundancy among protein structures.

Here, we present an algorithm which applies the multidimensional QR factorization to multiple alignments of protein structures. By equating redundancy to numerical linear dependence, the algorithm re-orders the protein structures by increasing linear dependence, and allows selection of a non-redundant set of structures with respect to a user defined similarity threshold, subject to the constraint that at any threshold the resultant non-redundant set best represent the evolutionary history of the protein group. The non-redundant representative multiple alignments, and derived statistical representations, which result from this method are termed evolutionary profiles. After introducing the theoretical background of the QR algorithm, we show how the derived structure-based phylogenies agree with established sequence-based phylogenies and how proper comparison of structures allows investigation of the most distant evolutionary events; namely, those events that predate the split between the main lines of descent (bacterial *versus* archaeal plus eukaryotic)[33] as represented by the root node of the universal molecular phylogenetic tree. Since the structural databases do not always allow the formation of complete evolutionary profiles and because the association of sequences with structures is at the core of the structure prediction problem, we illustrate how multiple alignments of structures and sequences can be combined in such a way that the resulting non-redundant sequence-structure profiles well represent the evolutionary space in both sequence and structure.

Theory

We begin by introducing the STAMP algorithm used to superpose protein structures, and then define a measure of structural similarity between homologous proteins, Q_H. As the multiple alignment of structures and sequences requires a multidimensional matrix encoding, we will first demonstrate an algorithm to find a non-redundant set *via* a model problem of one-dimensional proteins. After a brief description of the QR factorization, we detail how the algorithm is modified for multidimensional factorization, and the accompanying parameter search is carried out to assure the derived subsets best represent the evolutionary history of the group. Phylogenetic analysis methods used for constructing protein structure-based dendrograms and a hierarchical formulation of the multidimensional QR factorization are also discussed. Although a brief sketch of some of the techniques was presented,[33] here we present the algorithm in greater detail.

Structural alignment

Alignments were computed using the multiple structural alignment program STAMP,[34] 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. The algorithm does not include sequence-dependent information. This program uses a progressive multiple alignment procedure in which all possible pairwise alignments are computed, and then a hierarchical clustering analysis based on structural similarity is used to build the multiple alignment. The program aligns the most similar structures first, and moves along a structural dendrogram to add groups of aligned structures to the multiple alignment. 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. The initial alignments were executed with the following STAMP parameters: -scan true -npass 2 -slide 5 -scanscore 6. The final multiple structural alignments were computed with default parameters. The original version of STAMP systematically misaligns N and C-terminal residues, but this fault 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.[35]

Structural similarity measure Q_H

We derive a structural similarity measure between homologous proteins which is based on the structural identity measure, Q, developed by Wolynes, Luthey-Schulten and coworkers[36] in the

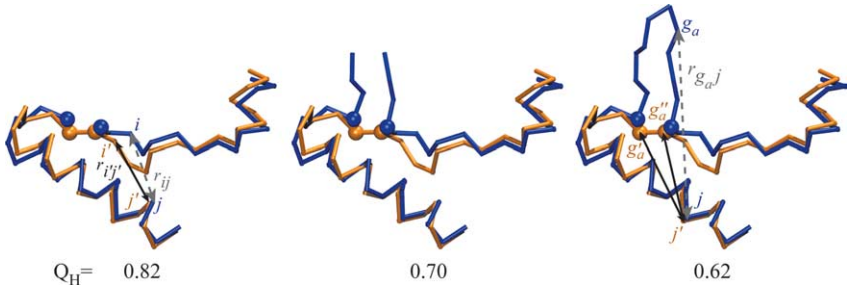

Figure 1. Illustration of Q_H, specifically demonstrating the effect of the q_{gap} term. The structural overlap shown is of two homologous fragments from d1efwa3 (aspartyl-tRNA synthetase, orange) and d1bbwa2 (lysyl-tRNA synthetase, blue). The left panel corresponds to only the computation of q_{aln} as the insertion in d1bbwa2 has been removed. As the insertion increases to its full length (left to right), the value of Q_H decreases. Notation is explained in Structural similarity measure Q_H, and all molecular structures were drawn using VMD.[35]

field of protein folding. Our adaptation of Q is referred to as Q_H, and the measure is designed to account for the presence of gaps and how insertions perturb the aligned core structure:

$$Q_H = \aleph^{-1}[q_{aln} + q_{gap}]$$

$\aleph$ is the normalization, specifically given below. Q_H is composed of two components. The term q_{aln} is identical in form to the un-normalized Q-measure of Eastwood *et al.*,[36] and it computes the fraction of similar contact distances between the aligned residues of a pair of proteins.

$$q_{aln} = \sum_{i \le j-2 \vee i' \le j'-2} \exp\left[-\frac{(r_{ij} - r_{i'j'})^2}{2\sigma_{ij}^2}\right]$$

This term computes the fraction of C_α–C_α pair distances that are the same or similar between two aligned structures. r_{ij} is the spatial C_α–C_α distance between residues i and j in the protein "a", and $r_{i'j'}$ is the C_α–C_α distance between residues i' and j' in the protein "b". This term is restricted to aligned positions, e.g. where i is aligned to i' and j is aligned to j', and the summation is over all unique, non-nearest neighbor residue pairs (see Figure 1). The symbol $\vee$ is an "inclusive or", which accounts for the perturbation of insertions to aligned gap edges (see below).

The normalization, $\aleph$, accounts for the contributions to Q_H from both the aligned regions and from contacts between the gap residues and each aligned position. In this way, initially gaps are strictly penalized, but the q_{gap} term reduces this penalty for smaller gaps that are in closer contact to the core structure and essentially maintains the initial penalty for long gaps that wander far from the core. Insertions that perturb the aligned structure less are also ascribed a lower penalty. The q_{gap} term is expressed as:

$$q_{gap} = \sum_{g_a}^{G_a} \sum_{j}^{N_{aln}} \max\left\{ \exp\left[-\frac{(r_{g_a j} - r_{g'_a j'})^2}{2\sigma_{g_a j}^2}\right], \exp\left[-\frac{(r_{g_a j} - r_{g''_a j'})^2}{2\sigma_{g_a j}^2}\right] \right\} + \sum_{g_b}^{G_b} \sum_{j}^{N_{aln}} \max\left\{ \exp\left[-\frac{(r_{g_b j} - r_{g'_b j'})^2}{2\sigma_{g_b j}^2}\right], \exp\left[-\frac{(r_{g_b j} - r_{g''_b j'})^2}{2\sigma_{g_b j}^2}\right] \right\}$$

where g_a and g_b are the residues in insertions in both protein "a" and protein "b" respectively. As shown in Figure 1, each insertion residue is associated with a C-terminal and N-terminal gap edge. The gap edge is formed by aligned residues on either side of the inserted residues. In Figure 1, the gap edge residues are marked with spheres. In constructing the q_{gap} term, we hypothesized that the more the gap residues deviated from the nearest gap edge, the lower the value of structural similarity between the two proteins. In protein "a", therefore, the contact distance, $r_{g_a j}$, between a residue j and the gap residue g_a, is compared with the contact distances, $r_{g'_a j'}$ and $r_{g''_a j'}$, between residue j' of protein "b", which is aligned to residue j, and the gap edges, represented by residues g'_a and g''_a in protein "b". The "max" function takes whichever gap edge, g'_a or g''_a, that produces a larger contribution to Q_H. The outer summation is over all inserted residues in protein "a", g_a, while the inner summation is over all non-nearest neighbor aligned residues. The definition is analogous for insertions in protein "b".

The normalization and the σ_{ij}^2 terms are computed as:

$$\aleph = \tfrac{1}{2}(N_{aln} - 1)(N_{aln} - 2) + G^{(0)} N_{aln} + G^{(1)}(N_{aln} - 1) + G^{(2)}(N_{aln} - 2) + N_g$$

$$\sigma_{ij}^2 = |i - j|^{0.15}$$

N_{aln} is the number of aligned residues. The total

number of residues in gaps, in both proteins "a" and "b", is equal to $G^{(0)}+G^{(1)}+G^{(2)}$. The number of residues in gaps with no gap edge residues as nearest neighbors is $G^{(0)}$. The number of gap residues with one nearest neighbor in a gap edge is $G^{(1)}$, and the number of gap residues with two nearest neighbors that are gap edges, i.e., gaps of one residue in length, is $G^{(2)}$. These three cases of gap residues are separately defined because contact distance comparisons involving residues that are nearest neighbors are excluded from the computation of Q_H. Each of the $G^{(0)}$ residues make N_{aln} contributions to Q_H, the $G^{(1)}$ residues, because they each have one nearest neighbor in an aligned pair, make only $N_{aln}-1$ contributions. Similarly, the $G^{(2)}$ residues each make $N_{aln}-2$ contributions to Q_H. In reference to the q_{gap} formula (above), the total number of residues in gaps in proteins "a" and "b" is $G_a=(G_a^{(0)}+G_a^{(1)}+G_a^{(2)})$ and $G_b=(G_b^{(0)}+G_b^{(1)}+G_b^{(2)})$ respectively.

The only exception to the exclusion of counting nearest neighbor contact comparisons involves the comparison of the contacts between the four residues of each gap edge. This additional comparison measures the perturbation of the insertion to the aligned gap edges, and the number of these contributions to Q_H is N_g, which sum of the number of insertions in protein "a", the number of insertions in protein "b" and the number of simultaneous insertions (referred to as bulges or c-gaps[37]). Gap-to-gap contacts and intra-gap contacts do not enter into the computation, and terminal gaps are also ignored. σ_{ij}^2 is a slowly growing function of sequence separation of residues i and j, and this serves to stretch the spatial tolerance of similar contacts at larger sequence separations. Q_H ranges from 0 to 1 where $Q_H=1$ refers to identical proteins. If there are no gaps in the alignment, then Q_H becomes $Q_{aln}=\aleph q_{aln}$, which is identical to the Q-measure described.[36] See Figure 1.

QR factorization of model alignments

An enormous variety of problems in scientific computing can be mapped onto the eigenvalue problem, $Ax=\lambda x$, or the least-squares problem, $Ax=b$. In the 1950s, the emergence of computers allowed for the possibility of treating large systems of equations. A key step forward, however, came with the "recognition that matrices could be reduced, by orthogonal transformations, in a finite number of steps, to some special reduced form that lends itself more efficiently to further computations".[38] In the case of multiple structural alignments, the alignment matrix, A, contains redundant information and the problem is to find a reduced set that is most representative of the structural data. With respect to the eigenproblem and the least-squares problem, the matrix A can be reduced to a special form, e.g. upper triangular, after which the problem can be solved by simple back substitution. The QR factorization by successive application of elementary reflectors, called Householder

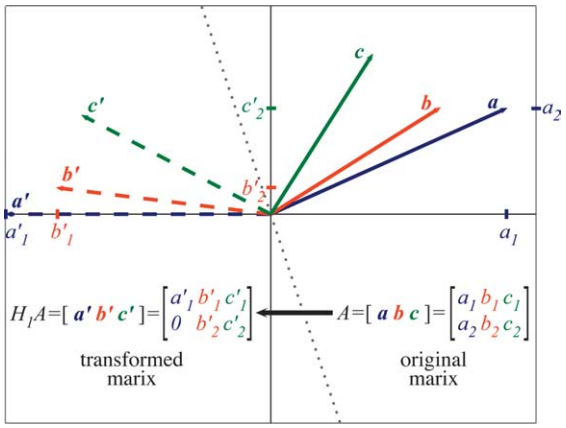

Figure 2. A depiction of the Householder transformation. The 2×3 matrix A contains three vectors, $\boldsymbol{a}$, $\boldsymbol{b}$ and $\boldsymbol{c}$, e.g. the x-coordinates of three proteins, each of two residues in length. As indicated by the algebraic form, the effect of the Householder transformation, H_1, is to reflect the vector $\boldsymbol{a}$ about the dotted line onto the x-axis. The dotted line bisects the angle between the vector $\boldsymbol{a}$ and the x-axis, i.e. the axis of the first unit vector. H_1 also has an effect on vectors $\boldsymbol{b}$ and $\boldsymbol{c}$, which are concomitantly reflected about the dotted line. The reflection of $\boldsymbol{a}$ is referred to as $\boldsymbol{a}'$, likewise for vectors $\boldsymbol{b}$ and $\boldsymbol{c}$. The transformation reveals which vector, $\boldsymbol{b}$ or $\boldsymbol{c}$, has the largest component in common with $\boldsymbol{a}$. In this simple problem, it is clear that $\boldsymbol{c}$ is more dissimilar to $\boldsymbol{a}$ than $\boldsymbol{b}$ and that $\boldsymbol{b}$ would be pivoted to the right.

transformations,[39] was quickly recognized as the most efficient route to triangularizing a matrix.

The QR factorization of the $m\times n$ matrix A is expressed as:

$$Q^{\mathrm{T}}(Ax)=Q^{\mathrm{T}}(b)$$

$$Q^{\mathrm{T}}Ax=\tilde{R}x=Q^{\mathrm{T}}b=\begin{bmatrix}c_1\\ c_2\end{bmatrix}$$

where $\tilde{R}=[\,R\quad 0\,]^{\mathrm{T}}$ is introduced as compact notation, R is an $n\times n$ upper triangular matrix and Q^{T} is an orthogonal transformation matrix which triangularizes A. Q^{T} is also applied to the right-hand side of the least-squares equation. The resulting least-squares problem, $Rx=c_1$ is then solved by back substitution with a minimum residual Euclidean norm of $\|c_2\|_2$.

The orthogonal transformation, Q^{T}, is a product of the successive application of Householder transformations, H_j, such that, $Q^{\mathrm{T}}=H_n\ldots H_2H_1$. Each Householder transformation, H_k, is an elementary reflector which is designed to reflect the vector represented in the kth column of A onto the axis of the kth unit vector while preserving the length, or more precisely the Euclidean norm, of the kth column vector. Equivalently, the elementary reflector, H_k, is designed to zero all entries below the diagonal of the kth column and to preserve the Euclidean norm of the kth column by inserting the annihilated magnitude of the below diagonal

entries into the diagonal entry. Thus, *via* successive application of H_k for $k=1, 2,\ldots n$, the matrix A is reduced to triangular form if $m=n$ or the form of $\tilde{R}$ if $m>n$. Both the geometric and algebraic interpretations of the Householder transformation are shown in Figure 2. The Householder transformation H_1 described in Figure 2, is constructed for the vector $\mathbf{a}$ such that $H_1\mathbf{a} = \mathbf{a} - (2\mathbf{v}^T\mathbf{a}/\mathbf{v}^T\mathbf{v})\mathbf{v}$ in which $\mathbf{v} = [0 \quad a_2]^T - \alpha[1 \quad 0]^T$ where $\alpha = -\text{sign}(a_1)\|a_2\|_2$. For a detailed description of the Householder transformation, see the work done by Heath.[40]

The matrix A may include linearly dependent or nearly linear dependent columns which contain redundant data. Golub developed the QR factorization with column pivoting (QRP) algorithm which orders the columns of A by increasing linear dependence from left to right, and allows an approximate solution to the least-squares problem by only retaining the first r ordered columns of A.[41] Prior to the application of H_k, i.e. the kth step of the QR factorization, the kth column of the matrix A is exchanged with a column j having maximum subvector Euclidean norm, defined as $\max_{j=k,\ldots,n}(s_j^{(k)})$ where:

$$s_j^{(k)} = \left(\sum_{i=k}^{m} a_{ij}^2\right)^{1/2}$$

The column exchange process is encoded in the permutation matrix P and the factorization is rewritten as

$$Q^T AP = \tilde{R}$$

The above description implies that the matrix A is rank deficient, i.e. $\text{rank}(A)<n$. Let $A^{(r)}=H_r\ldots H_1 AP_1\ldots P_r$, which represents the matrix A after the rth step of the QRP algorithm. Note that $A^{(n)} = \tilde{R}$, and due to the pivoting strategy the diagonal entries of $A^{(n)}$ obey the relationship $a_{11}^{(n)} > a_{22}^{(n)} > \cdots > a_{nn}^{(n)}$. This procedure can be used to numerically determine $\text{rank}(A)=r$ with respect to the threshold τ such that $a_{11}^{(r)} > a_{22}^{(r)} > \cdots > a_{rr}^{(r)} > \tau > a_{r+1r+1}^{(r+1)} \cdots > a_{nn}^{(n)}$. For this reason, the QRP algorithm is also referred to as the rank-revealing QR factorization. This strategy is used to determine the rank of A, and to define a minimal basis, or maximally linearly independent set, from the columns of A. Although there are many standard descriptions of the QR factorization including Householder's original paper[39] and Wilkinson's famous text,[42] Heath presents an accessible and modern treatment of both the QR and QRP algorithms.[40]

Matrix encoding of protein structure

Since our goal is to use the QRP algorithm to define a non-redundant set of protein structures from a multiple alignment, we encode the data in the alignment matrix, A which is of dimension $m_{\text{aln}}\times n_{\text{proteins}}\times d$. Each column of A corresponds to a single protein structure, and the multiple alignment is defined by the rows of A. The total length of the multiple alignment is m_{aln}, and n_{proteins} is the number of proteins in the alignment. The multiple structural superposition provides a set of rotated coordinates for each protein in the alignment. The rotated real space coordinates of the C^α positions for the proteins are encoded in the first three components of the d-dimension, giving the matrices X, Y and Z. Gapped positions are accounted for by the fourth component, the matrix G, of the d-dimension, so $d=4$. G is a binary matrix in which zero represents aligned positions and 1 represents gaps, while $\tilde{G} = gG$ is the scaled gap matrix. The four components of the alignment matrix, A, are expressed as $A_{(1)}=X$, $A_{(2)}=Y$, $A_{(3)}=Z$, $A_{(4)}=\tilde{G}$. The aligned positions are described by a 4-vector of the type (x_{C^α}, y_{C^α}, z_{C^α}, 0), while gapped positions are encoded by the 4-vector (0, 0, 0, g). In the current application, all gaps are treated with the same weight, namely the gap scale parameter g which is defined as:

$$g = \gamma \frac{||X||_{F_4} + ||Y||_{F_4} + ||Z||_{F_4}}{||G||_{F_4}}$$

where γ is a constant and one of two parameters that we determine numerically in a procedure defined below. The notation $||X||_{F_4} = \left(\sum_{i,j} |x_{ij}|^4\right)^{1/4}$ defines the Frobenius-like matrix 4-norm. The form of the equation for the gap scale parameter is motivated by the notion that any gap position should be weighted equivalently to any real space position.

Multidimensional QR factorization

The QR factorization, in the classical form, is designed to operate on an $m\times n$ matrix, but the alignment matrix, which encodes a multiple alignment of protein structures, simply cannot be expressed in fewer dimensions than $m_{\text{aln}}\times n_{\text{proteins}}\times 4$. Although there is an effort, based on knot theory, which is attempting to reduce the number of dimensions that are required to define protein topology,[43] this formulation of protein structure appears to be too coarse-grained for application in the present context. A novel multidimensional QR factorization, however, was recently formulated by Heck, Olkin & Naghshineh in the field of active noise control.[44,45] The multidimensional QR is, for a matrix of dimension $m\times n\times d$, essentially the simultaneous QR factorization of d-matrices each of size $m\times n$. The algorithm is formally expressed as:

$$Q_{(d)}^T A_{(d)} P = \tilde{R}_{(d)}$$

in which the alignment matrix, A, is of the form $A_{(1)}=X$, $A_{(2)}=Y$, $A_{(3)}=Z$, $A_{(4)}=\tilde{G}$.The key point is that the permutation matrix, P, is independent of the d-dimension, so that the columns of A, where each "column" is now actually a matrix representing an aligned protein, are not scrambled during pivoting operations. This property is the result of

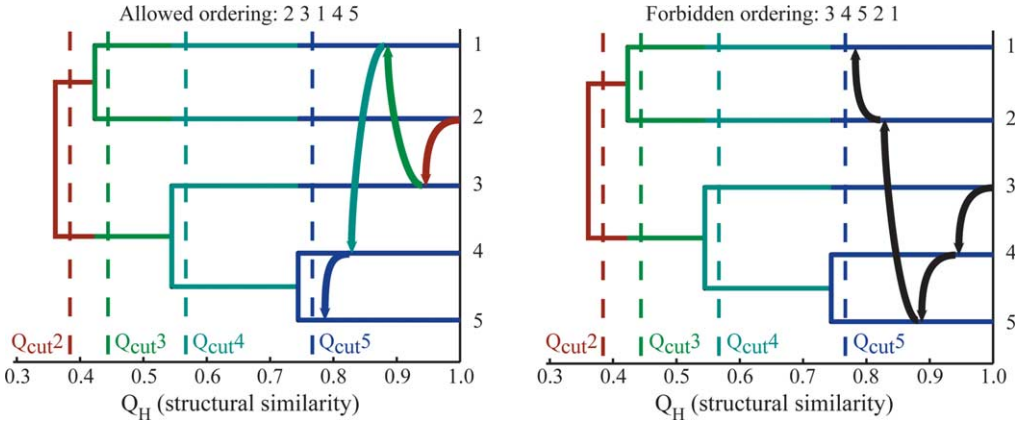

Figure 3. Allowed (left) and forbidden orderings (right) shown for an example dendrogram.

the pivoting rule, which takes the form of a Frobenius norm over d-space. At the kth step in the factorization, prior to the application of the Householder transformation, $H^{(k)}_{(d)}$, for each of the d-matrices, the permutation $P^{(k)}$ is constructed to exchange the kth column of A, over each d-dimension simultaneously, with the column of maximum Frobenius-like matrix p-norm, $\max_{j=k,\ldots,n_{\text{proteins}}}(||a_j||_{F_p})$ where:

$$||a_j||_{F_p} = \left(\sum_{d=1}^{4} \sum_{i=k}^{m_{\text{aln}}} |a_{ijd}|^p \right)^{1/p}$$

The integer, p, is a constant and the second of the two parameters that we optimize numerically in a procedure defined below. A new alignment matrix is generated, $\tilde{A} = AP$, in which the proteins in $\tilde{A}$ are ordered by increasing linear dependence from left to right. Since we assume that redundancy in a multiple alignment is directly related to linear dependence between the aligned proteins, trimming proteins from right to left in $\tilde{A}$, to a desired level of redundancy, gives a reduced set of proteins which form the non-redundant multiple structural alignment. We have also implemented this procedure for generating non-redundant multiple sequence alignments.[46]

The QR ordering is used to define a non-redundant set. Quite often this is done by incrementally including proteins, according to the QR ordering, until the maximum Q_{H} value, or some other pairwise similarity metric, is greater than a user-specified threshold. Although the following procedure is not a necessary part of the algorithm, in the next section we describe how a threshold can be naturally defined by computing a phylogenetic tree.

Phylogenetic analysis and evaluation of the QR ordering

Structure-based phylogenetic trees are drawn using either the neighbor-joining program in Phylip,[47] as in Figures 6–10, or the unweighted pair group method using arithmetic averages (UPGMA),[48] in Figures 3, 5 and 11, as implemented in MATLAB 6.5 (Mathworks, Inc.), and the measure of distance is $1-Q_{\text{H}}$. An example phylogeny of five protein structures, labeled 1–5, is shown in Figure 3. In this phylogeny, there are four distinct threshold partitions where each partition is defined by the emergence of a new branch. The first partition ranges from the initial bifurcation, splitting the five proteins into two subgroups, i.e. (1,2) and (3,4,5), to the bifurcation splitting protein 1 and protein 2. The second threshold partition begins with the formation of three groups, namely (1), (2) and (3,4,5), and ends at the bifurcation between proteins 3 and proteins (4,5). The third and fourth partitions are defined similarly. Each partition is associated with a Q_{H} cutoff value, Q_{cut^i}, where the ith threshold separates the proteins into i groups. The thresholds can be applied to the pre-computed QR ordering such that the first i proteins in the QR order represent each of the i groups defined by the threshold.

Unlike typical pairwise similarity threshold algorithms, the goal of the QR factorization is not merely to produce a set of proteins with pairwise similarity values below a given threshold, but rather to order the proteins by increasing linear dependence. Any arbitrary threshold can then be applied to the precomputed ordering and a "below threshold" set is produced by simply adding to or subtracting proteins from the representative set according to the QR ordering. One advantage of this algorithm is that rather than search through all pairwise relationships for values that violate the threshold each time a new threshold is applied, as is

Table 1. The allowed orderings for the example phylogeny in Figure 3

1 3 2 4 5	2 3 1 4 5	3 2 1 4 5	4 2 1 3 5
1 3 2 5 4	2 3 1 5 4	3 2 1 5 4	4 1 2 3 5
1 4 2 3 5	2 4 1 3 5	3 1 2 4 5	5 2 1 3 4
1 5 2 3 4	2 5 1 3 4	3 1 2 5 4	5 1 2 3 4

If the number of proteins is N, then there are $N!$ total possible orderings. In this example, of the 120 possible orderings, 16 are allowed and 104 are forbidden.

the case for typical pairwise similarity threshold algorithms, one simply follows the QR ordering, beginning with the first pair, and adds proteins to the representative set until a kth protein is added that violates the threshold; the representative set is then defined as the first $k-1$ proteins in the QR order. Since the QR factorization will yield different orderings given different values of the above mentioned adjustable parameters, p and γ, the QR must be parameterized to consistently give "allowed" orderings. The QR ordering is defined as "allowed" if it satisfies the following criterion: At each distinct threshold, the QR ordering incorporates the maximum number of proteins with pairwise similarity values less than the specified threshold. Non-allowed or forbidden orderings will result if, for example, the gaps are over-weighted with respect to the Cartesian coordinates, i.e. the aligned positions.

In Figure 3, the ordering, 2 3 1 4 5, is an example of an allowed ordering. When $Q_{\text{cut}}2$ is applied to this ordering, protein 2 represents the (1,2) cluster, and protein 3 represents the cluster (3,4,5). If the threshold is set at $Q_{\text{cut}}2$, we expect a minimal set of two proteins. Applying the threshold $Q_{\text{cut}}3$ would imply taking the first three proteins in the order, 2 3 1. Note that proteins 4 and 5 should not be included because they are too similar to protein 3, as $Q_{H_{3,4}}$ and $Q_{H_{3,5}}$ are both greater than the threshold, $Q_{\text{cut}}3$. Protein 2 is now representing its own branch, protein 3 still represents the (3,4,5) cluster and protein 1 represents itself. Continuing this procedure for the two remaining thresholds shows that the ordering, 2 3 1 4 5, is allowed for all distinct thresholds and is therefore declared to be an allowed ordering.

On the right hand side of Figure 3, a forbidden ordering, 3 4 5 2 1, is shown. The above procedure reveals that this ordering is allowed only, and trivially so, at $Q_{\text{cut}}5$ while this ordering is forbidden at all other thresholds. For example, consider the first two proteins in the order, 3 and 4, and apply the $Q_{\text{cut}}2$ threshold. Clearly $Q_{H_{3,4}} > Q_{\text{cut}}2$, so proteins 3 and 4 are representing only one of the two branches at the $Q_{\text{cut}}2$ level of similarity. As in this case, a forbidden ordering occurs when, for a specified threshold, below threshold branches of the tree are missed or others are over-sampled. The allowed orderings, listed in Table 1 for the example phylogeny in Figure 3, are degenerate, but small in number, as compared to the possible number of forbidden orderings.

Hierarchical multidimensional QR factorization

The multidimensional QR can fail to yield an allowed ordering when the information content of the alignment matrix, A, diminishes beyond some lower limit. Although there is not an adequate theoretical justification for this behavior or for the value of the lower limit, in practice this can occur in alignments represented by the phylogeny shown in Figure 5, where there are very large structural differences in the same alignment with proteins displaying very subtle structural differences. This failure is most simply resolved by application of a user-specified threshold of allowed redundancy, e.g. a Q_{Hcut}, and the following hierarchical QR factorization procedure.

As proteins are added to the non-redundant set, the pairwise similarity relationships are checked, and if any of those values exceed the threshold, Q_{Hcut}, the factorization halts. The proteins included in the first non-redundant set are removed and the remaining components of the alignment matrix, A, are set to their initial values. The procedure is repeated until no more proteins remain or until no sub-sets can be formed which contain pairwise similarity values below threshold.

Parameter search

Although the classical QR factorization naturally employs the 2-norm in the pivoting operations, the

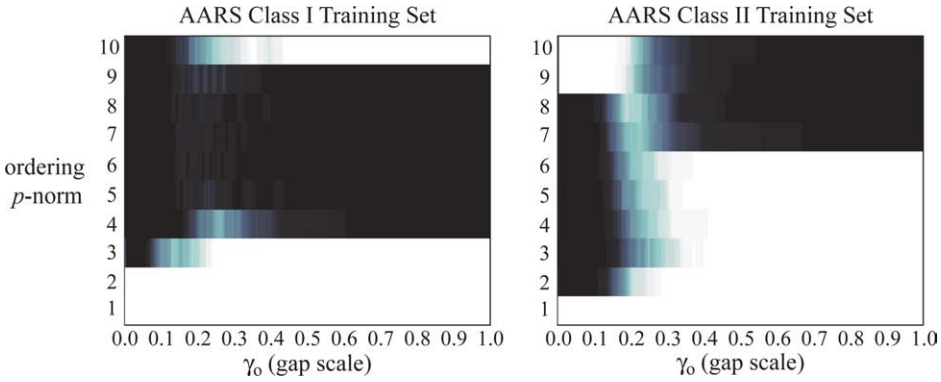

Figure 4. Result of the parameter search for the aminoacyl-tRNA synthetase (AARS) class I and class II training sets, see Figure 5. Black denotes forbidden QR-factorization orderings while white regions mark allowed orderings. For each candidate pair of parameters, p and γ, 50 random rotations and translations were performed, and the average of allowed and forbidden orderings for these linear transformations are plotted above. The abscissa in each of the above plots has been scaled such that $\gamma_0 = 2.83\gamma$. The parameters used in this study are chosen from the overlapping allowed regions in the above plots: $\gamma = 1.1$ ($\gamma_0 = 0.4$) and $p = 2$.

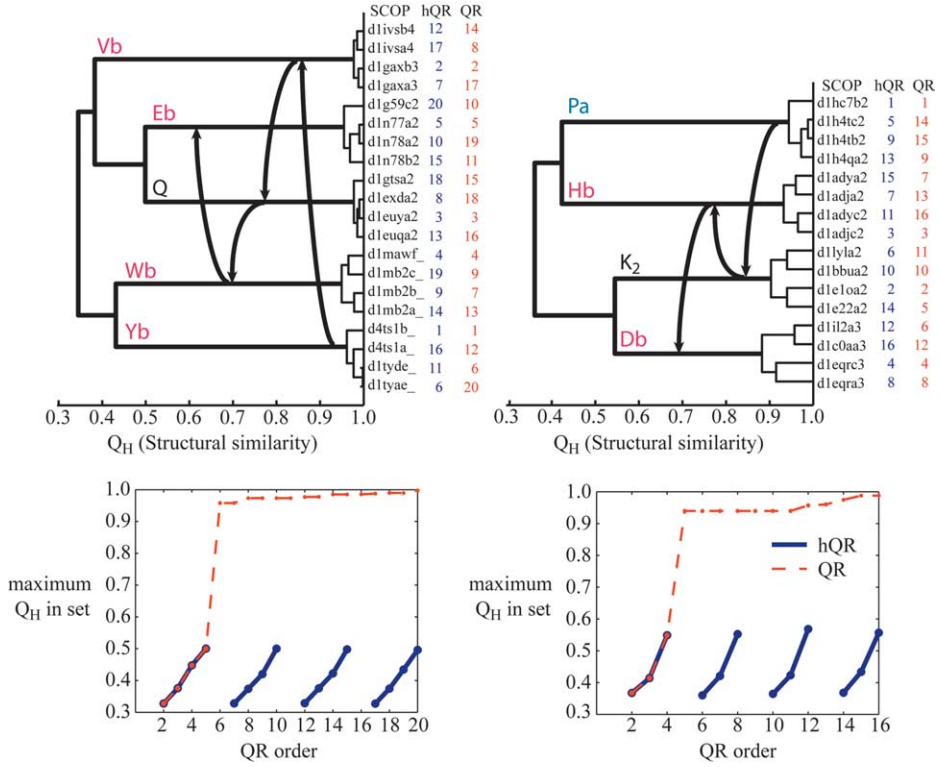

Figure 5. Phylogenies (above) are shown for the aminoacyl-tRNA synthetase (AARS) class I (left) and class II (right) training sets used in the parameter search. The function of the selected AARSs is indicated by the one-letter amino acid code, e.g. the branch labeled K_2 is occupied by four different crystallographic structures of the class II lysyl-tRNA synthetase. In those selected AARSs that exhibit the full canonical or basal canonical phylogenetic pattern,[50] the letter a or b indicates that the branch represents the archaeal or bacterial type, respectively, e.g. Eb is the bacterial type glutamyl-tRNA synthetase. The leaves of the phylogeny are labeled by the seven letter SCOP/ASTRAL domain codes[65,66] and by the placement of that protein structure in the QR ordering. Arrows depict the path of the QR order through the phylogenies. The dependence of maximum Q_H in the set is shown (below) *versus* the QR ordering. If the set is composed of only the first two proteins in the QR ordering then the maximum Q_H value is identical with the structural similarity value between the two proteins. At a set size of three, the Q_H value plotted is the maximum of all pairwise similarity values among all three proteins, and so on. The plots show the full QR ordering (red, broken) and the hQR ordering with $Q_{Hcut}=0.7$ (blue, continuous), which gives four separate non-redundant sets in each case, for the class I (left) and class II (right) AARS training sets.

derivation of the multidimensional QR factorization does not specify a particular p-norm or how to balance contributions from each matrix across the d-dimension. The notion that information in the coordinate and gap matrices should be evaluated with the same weight suggests the Frobenius-like matrix p-norm as a pivoting metric. The value of p is not obvious *a priori*, so we scale this integer from $p=1$ to $p=10$ in the parameter search procedure. In general, higher p-norms emphasize the outliers in the structural data.

Olkin *et al.*, who first implemented the multi-dimensional QR in active noise control problems, also studied the choice of matrix norm used in the pivoting operation.[44] They were not confronted with heterogeneous data, namely the combination of Cartesian coordinates with gap location data, so there was no need to consider a second scaling parameter, such as gap scaling parameter, γ, defined above. Since gap position is a binary data type, each position in the alignment is either a gap or not a gap, there must be a scaling operation such that the aligned positions and gap positions both contribute an appropriate weight to the final ordering of the multidimensional QRP algorithm. The value $\gamma=1$ gave promising initial results, but the occurrence of forbidden orderings, in some cases, motivated a larger search of γ-space. In the parameter search, γ is varied from 0.017 to 2.8 in increments of 0.017. Empirically, this space extends from a regime where gaps do not contribute enough to the final ordering, to a limit where the pattern of gaps completely dominates the ordering. The two adjustable parameters in our implementation of the multi-dimensional QR factorization can only be defined numerically. For each set of candidate values for the parameters p and γ, a multidimensional QR factorization is computed. The resulting ordering is defined as allowed or forbidden by the criteria give in Phylogenetic analysis and evaluation of the QR ordering. The results of this parameter search, for two training sets with totally different fold motifs, are discussed in Technical analysis of the QR factorization algorithm (see Figure 4).

The definition of allowed and forbidden orderings is dependent on the method used to construct the phylogenetic tree. In the case of structure-based phylogenies, we have thus far employed distance-based methods, as opposed to maximum-likelihood or parsimony. Indeed, distance based methods may be ideal for describing homology among protein structures, since the similarity metric, Q_H, is a comparison measure of physical distances, unlike sequence "distances" between aligned amino acid residues which are based on empirical substitution probabilities. In this work, we fit the adjustable parameters to agree with phylogenies based on the Q_H metric. If a different metric for inferring structure-based phylogenetic trees is desired, the parameters can be tuned to produce allowed orderings for the alternative trees.

Results and Discussion

Technical analysis of the QR factorization algorithm

Defining the QR parameters

Our implementation of the multidimensional QR factorization of the alignment matrix, A, depends on two adjustable parameters, the gap scaling constant γ and the ordering p-norm (see Theory). In order to determine appropriate parameters for the QR algorithm, two training sets were used, both of which exhibit a well-defined phylogenetic topology, meaning that the branch order is robust and the structures are sufficiently distinct that the same groupings would be obtained whether we used Q_H or RMSD as a metric or whether we used UPGMA or neighbor-joining as a tree-drawing method. Note that this is only the case for the training sets shown in Figure 5. The phylogenies shown in subsequence sections are better represented as neighbor-joining trees. The advantage of using a group of homologs with well-defined phylogenies for the training set, is that allowed and forbidden orderings, which result from the QR factorization, can be confidently defined. One of the training sets contains a selection of protein domains from the AARS class I family, of the Rossmann fold type, and the proteins in the second set are selected from the AARS class II family, a novel α–β fold with anti-parallel β-sheets. The class I and class II AARSs are unrelated in sequence and structure,[33] and phylogenies for the two training sets are shown in Figure 5.

The results of the parameter search, in which for each pair of candidate parameters a QR factorization is computed and the ordering is determined to be allowed or forbidden, are shown in Figure 4. An additional complication is that the QR factorization will not necessarily yield the same ordering if linear transformations, rotations and translations, are applied to the alignment matrix, and, though the QR algorithm is algebraically guaranteed to be scale invariant, numerical error may lead to aberrant orderings if the alignment matrix is scaled. Since protein structure crystallization reference frames are arbitrary, the parameters must be chosen so that the QR factorization is robust to linear transformations; an allowed ordering is obtained even after arbitrary linear transformations on the alignment matrix. For each pair of candidate parameters, therefore, the alignment matrix was randomly rotated, translated and scaled 50 times, and for each of the random linear transformations the QR factorization was computed. The results in Figure 4 show the average, over the 50 transformations, of allowed and forbidden orderings. The AARS class I and class II training sets show a generous overlap of allowed orderings in the parameter space. The parameters chosen from this region and used in the following applications are $\gamma = 1.1$ and $p = 2$.

Interpreting the QR ordering in a well-defined phylogeny

QR ordering of the class I and class II AARS training sets, which results from applying the above parameters, is depicted in Figure 5. Note that in both training sets the QR factorization gives an allowed ordering (see Theory), which means that each major branch in the phylogeny is visited once before it is repeated. In addition, any threshold, or similarity cutoff in Q_H, applied to the order will include the appropriate and maximum number of structures with all pairwise relationships below threshold. The advantage of the allowed ordering generated by the QR factorization is that any arbitrary similarity threshold can be applied without the need to recompute the factorization. This feature is in sharp contrast to typical similarity cutoff algorithms (see Introduction), which do require a complete re-computation of representative set members if the similarity threshold is adjusted. For the QR factorization method, adjusting the similarity threshold just requires adding or subtracting one or more proteins from the pre-computed ordering.

While the allowed ordering is observed for the first five structures in the class I set and the first four structures in the class II set, after all of the major groups are visited, the QR ordering begins to break down. At this point, the partially factorized alignment matrix has lost too much information to produce an allowed ordering for the remaining proteins. This behavior emerges when one representative from a group of nearly identical molecules, i.e. different crystal structures of the same protein sequence showing very little difference in conformation, is used to construct a Householder transformation. Since the copies of the representative are so nearly identical, and thus have near exact linear dependence to the representative, the Householder transformation has the effect of completely annihilating the structural information describing those copies in the alignment matrix. As mentioned in Hierarchical multidimensional QR factorization,

this situation is rescued by applying the hierarchical QR (hQR) factorization, which requires specification of a Q_H threshold value. As soon as the ordering incorporates a protein which pushes the maximum Q_H value of the set above threshold, the previous proteins in the order are assigned to the first representative set, and the alignment matrix is re-initialized, excluding the proteins assigned to the first representative set. The factorization is then allowed to continue on the re-initialized alignment matrix until the second representative set is similarly defined, and additional representative sets are defined until no more proteins remain. See Figure 5.

Application to the analysis of protein structure evolution

Congruence of sequence and structure phylogenetic trees

Since Pauling first discussed the concept of molecular evolution,[49] an enormous field of research was founded on the idea that the sequences of nucleic acids and proteins contain a trace of the evolutionary course of genes and sometimes organisms.[50] Visualizing overlapped structures evokes a similar idea, that the structure of a biological molecule contains evolutionary information. The challenge in recovering the potential evolutionary information from protein structure alone is first addressed by deriving a measure of structural similarity between homologous proteins. The metric should consider the similarity of the aligned three-dimensional structure of the molecules and also account for the effect of gaps, i.e. insertions and deletions.

We have used such a measure, Q_H (see Theory), to compute structure-based phylogenies for the class I and class II aminoacyl-tRNA synthetases (AARSs).[33] By showing agreement with maximum-likelihood sequence-based analysis of the AARSs,[51] we demonstrated that, if the gaps are properly considered, evolutionary information is indeed recoverable from protein structure alone, allowing accurate computation of protein structure-based phylogenies.[33] In Figure 6, we present an updated structure-based phylogeny for the aspartyl and asparaginyl-tRNA synthetase (AspRS–AsnRS) group, with the addition of a newly release structure, PDB code 1n9w, and make direct comparison to the sequence-based phylogeny for the same group. Structural differences between the bacterial and archaeal AspRSs are also highlighted. While the archaeal-eukaryotic structure signatures are seen mainly in the elongation of helices, features unique to the bacterial genre of AspRS include a large inserted domain, a β-loop-β motif and a small inserted helix. *Thermus thermophilus*, a member of the *Deinococcus-Thermus* phylum, has the "native" bacterial type AspRS (*Deinococcus-Thermus* 1) as well as a second AspRS (*Deinococcus-Thermus* 2) of the archaeal genre, acquired through horizontal gene transfer (HGT). The HGT event is detectable in both sequence[51] and structure.

In the crystal structure d1n9wb2, 40 residues are not resolved, which are well conserved in both sequence and structure in all members of the AspRS–AsnRS group. Due to the high level of conservation, the region was straightforwardly modeled with the Modeler program,[52] using d1b8aa2 and d1asza2 as scaffolds for the unresolved regions. The original crystal structure still groups clearly with other structures of the archaeal genre and has all the archaeal-type structure signatures.

Evolution of structure in aspartyl-tRNA synthetase

Since protein structure contains evolutionary information and because it is more highly conserved than sequence,[53,54] the comparative analysis of structures allows the investigation of evolutionary events that pre-date the split between the main lines of descent in the universal phylogenetic tree,[33] such as those events marked i and ii in Figure 7. This appears to correspond to an era of rapid evolutionary change in which the basic protein functions were evolved, including most of the AARS enzymatic specificities,[33] and represents the evolution of the last common ancestral state itself.

The QR algorithm can be used to obtain an unbiased profile of structural conservation at different levels of similarity: class, subclass, enzymatic specificity, domain-of-life genre and species.[33] Based on these profiles, in Figure 7 we constructed candidate structures for AspRS from the last common ancestral state of the class II AARSs to the present enzyme in *Escherichia coli* and mapped changes in contacts between the proposed ancestral structures and the modern cognate tRNA. These candidate ancestral structures are simply a depiction of the portions of the molecule showing high-structural conservation, $Q_H > 0.4$, at the different levels of diversity. We hypothesize that the common conserved portions of the molecule are ancestral, but caution that we cannot depict portions of the ancestral proteins that have been lost or that are not well conserved among the modern forms. According to this analysis, tRNA recognition occurred only through the catalytic domain early in evolution and only later included the anticodon binding domain of the OB-fold type.

A representative set of OB-folds involved in translation

The OB-folds involved in translation participate in RNA–protein interactions that are part of the core fabric of the biological cell, as they play important roles in at least three major components of the translation machinery, including aminoacylation of tRNAs, translation initiation and the ribosome itself. Homologs of the OB-fold anticondon binding domain observed in AspRS, AsnRS and LysRS are present in the C-terminal domain of MetRS and domain B2 of PheRS, though not as anticodon

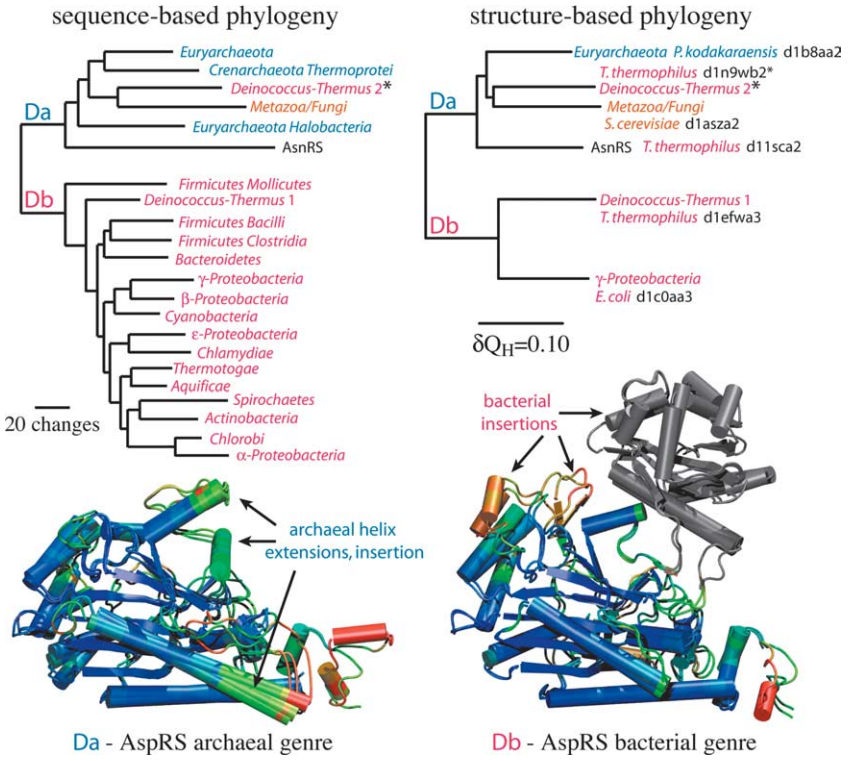

Figure 6. Congruence between the sequence[51] and structure-based phylogenies[33] for the AspRS–AsnRS group (both trees are rooted by LysRS). The deep division between the archaeal (Da) and bacterial (Db) genre has been documented in sequence by the presence of specific sequence signatures and phylogenetic analysis. This division is also apparent in the differences in the protein structures. The structure signatures that distinguish the archaeal and bacterial genres are shown in structural overlaps below the phylogenetic trees. A case of HGT from the archaeal-eukaryotic group to the *Deinococcus-Thermus* group (*) is clear in both sequence and structure. The structure d1n9wb2 (*Deinococcus-Thermus* 2) is shown in the overlap with other structures of the archaeal type (Da), and the structure d1efwa3 (*Deinococcus-Thermus* 1) is shown overlapped with the bacterial type AspRS from *E. coli*. Structures of the AspRS–AsnRS catalytic domains are color-coded by structural conservation. Here and throughout the phyla, organism names are color-coded according to their respective domain of life as Archaea (blue), Eucarya (gold) and Bacteria (red).

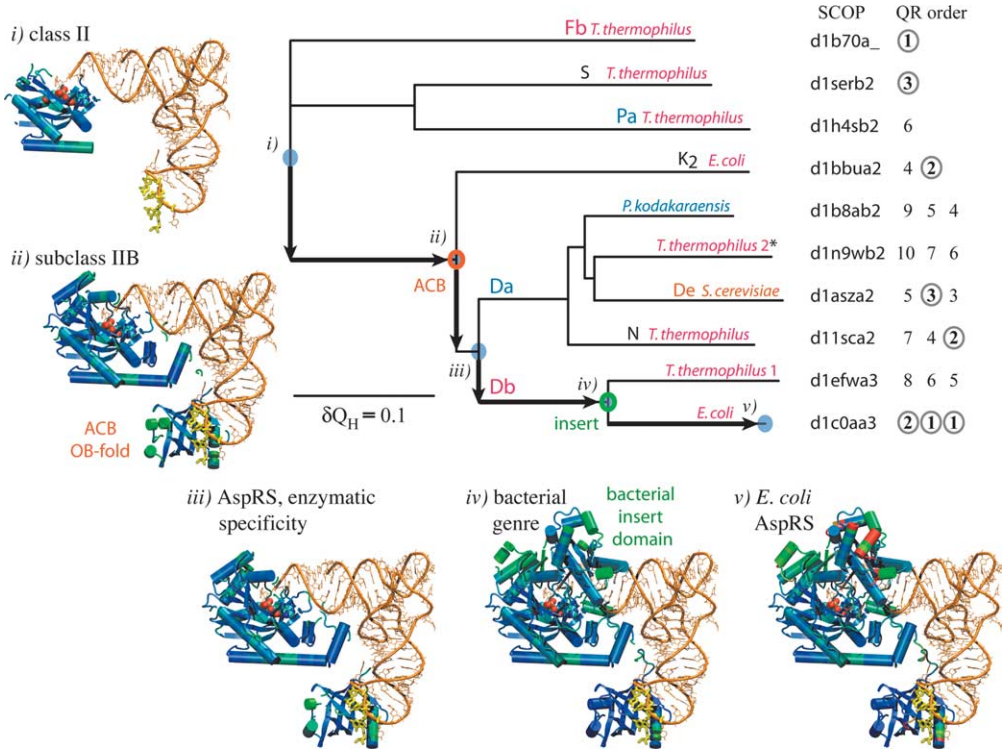

Figure 7. The evolution of aspartyl-tRNA synthetase (AspRS) from *E. coli* is depicted by showing the most conserved portions of the structure, $Q_H > 0.4$, from the class II AARS level to the modern enzyme found in *E. coli*. At each evolutionary stage (i–v), the degree of structural conservation is computed based upon a QR-derived representative set (circled numbers in the QR order). The phylogeny shown above is abbreviated at the class II AARS level of similarity,[33] showing only PheRS (Fb), representing subclass IIC, and SerRS (S) and ProRS (Pa), representing subclass IIA. As this AspRS enzyme "evolves" (i–v), note that contacts with the tRNA become increasingly specific and intricate. The OB-fold type anticodon binding (ACB) domain is an ancestral feature of subclass IIB, added at point ii, but not of the entire AARS class II family.

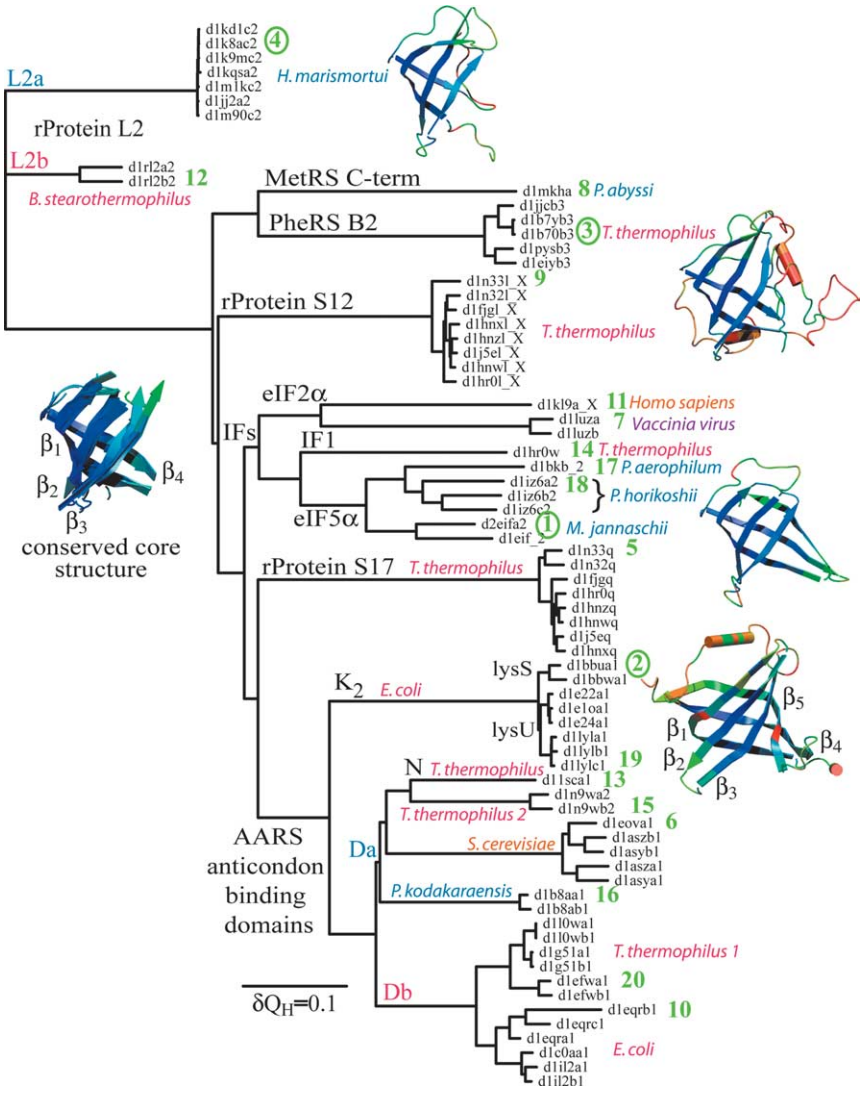

Figure 8. Structure-based phylogeny for OB-folds involved in translation. The QR ordering for the first 20 proteins in the order is indicated by a green number adjacent to the SCOP domain code. The first four structures in the QR order (circled numbers) are shown next to their position in the tree. These structures display the canonical five-stranded β-barrel topology observed among other OB-folds, except ribosomal protein L2, which is missing β5 and half of β4. The structures are color-coded according to structural conservation among the nine members of the non-redundant set, and the overlap of the conserved core ($Q_H > 0.4$) is also shown. The neighbor joining tree was computed with Phylip[46] based on a distance matrix of pairwise $1-Q_H$ values. The L2 proteins were used as the outgroup.

binding domains in these two AARSs, protein domains (eIF2α, IF1 and eIF5α) that contribute to the translation initiation factor (IF) assembly, and universally distributed ribosomal proteins in both the large (L2) and small (S12 and S17) ribosomal subunits. While beyond the scope of our current investigations, studying the distant evolutionary relationships between these particular protein structures may reveal new clues as to how the complex process of translation evolved from some simpler state. Since the OB-folds have been recruited to these various components of the translation machinery, their evolution may mirror the evolution of the translation apparatus itself. Sequence-based analysis in this direction has already revealed that the rudiments of the translation initiation complex are indeed universal in distribution, and thus, already present in the last common ancestral state of all life.[55,56]

The phylogeny in Figure 8 depicts the evolution of structure among OB-fold domains involved in translation. These proteins are found in all forms of life, and are responsible for binding and recognizing single-stranded RNA. Most of these proteins show the standard OB-fold topology of five β-strands wrapped into a β-barrel fold such that β1, β2 and β3 form a three-strand sheet on one face of the protein. The opposite face is also a three-strand sheet formed by β1, which wraps almost completely around the barrel, β4 and β5. Interestingly, ribosomal protein L2 lacks this opposite face, as it only contains sheets β1, β2, β3 and approximately half of β4. L2 is one of the smallest OB-folds with only 65 residues, whereas most OB-folds are approximately 100 residues in length. These proteins typically share a canonical RNA-binding interface, formed by β2 and β3, that usually binds single-stranded RNA in a "standard polarity" where the RNA strands runs 5′ to 3′, beginning near β4 and β5 and ending at β2.[57] The RNA-binding face in the OB-folds is also the most structurally conserved region, and this and other features of the OB-fold discussed above are shown in Supplementary Figure 1.

The OB-phylogeny presented here (see Figures 8 and 9) displays nine major branches, which are represented, in the QR ordering and with representative structure labels given in parentheses, as: (1) initiation factor-1/eukaryotic initiation factor-5α

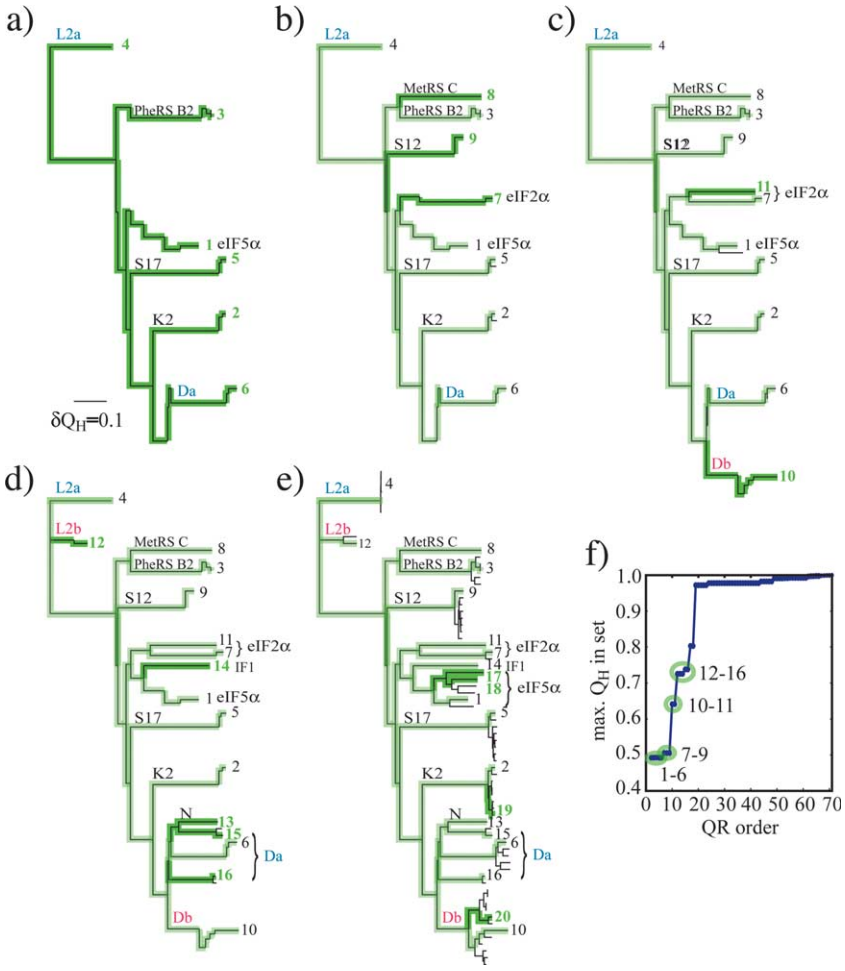

Figure 9. The above trees depict the incremental addition ((a)–(e)) of proteins in the QR factorization-based ordering for the OB-fold type proteins of known structure that are involved in translation. Dark green branches depict newly added representatives, light green indicates proteins previously included in the ordering. Figure 9(a) is but a skeleton of the complete tree, which nonetheless spans the breadth of the evolutionary space, Figure 9(e) includes the detail of all known structures. The plot in Figure 9(f) shows that the set of proteins becomes increasingly redundant, with increasing maximum Q_H value in the set, as proteins are added in the QR order.

group (eIF5α); (2) class II LysRS anticodon binding domain (K2); (3) PheRS domain B2 (PheRS B2); (4) ribosomal protein L2 (L2); (5) ribosomal protein S17 (S17); (6) AspRS anticodon binding domain (Da); (7) eukaryotic initiation factor-2α (eIF2α); (8) MetRS C-terminal domain (MetRS C-term); and (9) ribosomal protein S12 (S12). Some of these major groups are nearly equidistant in structural similarity, i.e. they are related to one another *via* short branchings in the phylogenetic tree. This indicates that the OB-fold phylogeny, based solely on the available structures, is not well defined at all nodes, and additional structures, such as archaeal examples of S12 and S17, would likely give better support to the branching order. The AARS anti-codon binding domains, the initiation factors (IFs), ribosomal proteins L2 and the MetRS–PheRS domains do, however, each form separate and well-defined clusters.

In such a phylogeny, where several of the major groups are equidistant, there is no precise way to define allowed or forbidden QR-based orderings as in the simple and well-defined training sets (see Interpreting the QR ordering in a well-defined phylogeny). We can, however, expect the QR ordering to visit each of the major branches before returning to any one of them. With this more relaxed notion of an appropriate QR ordering, the result of the QR factorization in the case of the OB-folds involved in translation can be properly interpreted. The tree diagrams in Figure 9 are included to graphically illustrate the effect of adding proteins in the QR order in this somewhat complicated phylogeny. Figure 9(a) shows the tree with just the first six proteins in the QR order, and these are indeed the six most distantly related of the nine major groups mentioned above. In other words, of any subset of six proteins from the entire set of known structures shown in Figure 8, these six best span the evolutionary space of the all proteins in the group. In Figure 9(b), the nine major branches are completely represented without any being represented more than once and without missing one of the major branches. In the following panels of the Figure, branches are added which represent increasing structural similarity to those previously added. While Figure 9(a) is but a skeleton of the complete tree, which nonetheless spans the breadth of the evolutionary space, Figure 9(e) includes the detail of all known structures.

In summary of the QR order, the first proteins in the order represent the major functions, followed by the inclusion of proteins with similar functions to each representative. The next proteins added are distantly related species-specific variants of the major functions, like the addition of the bacterial

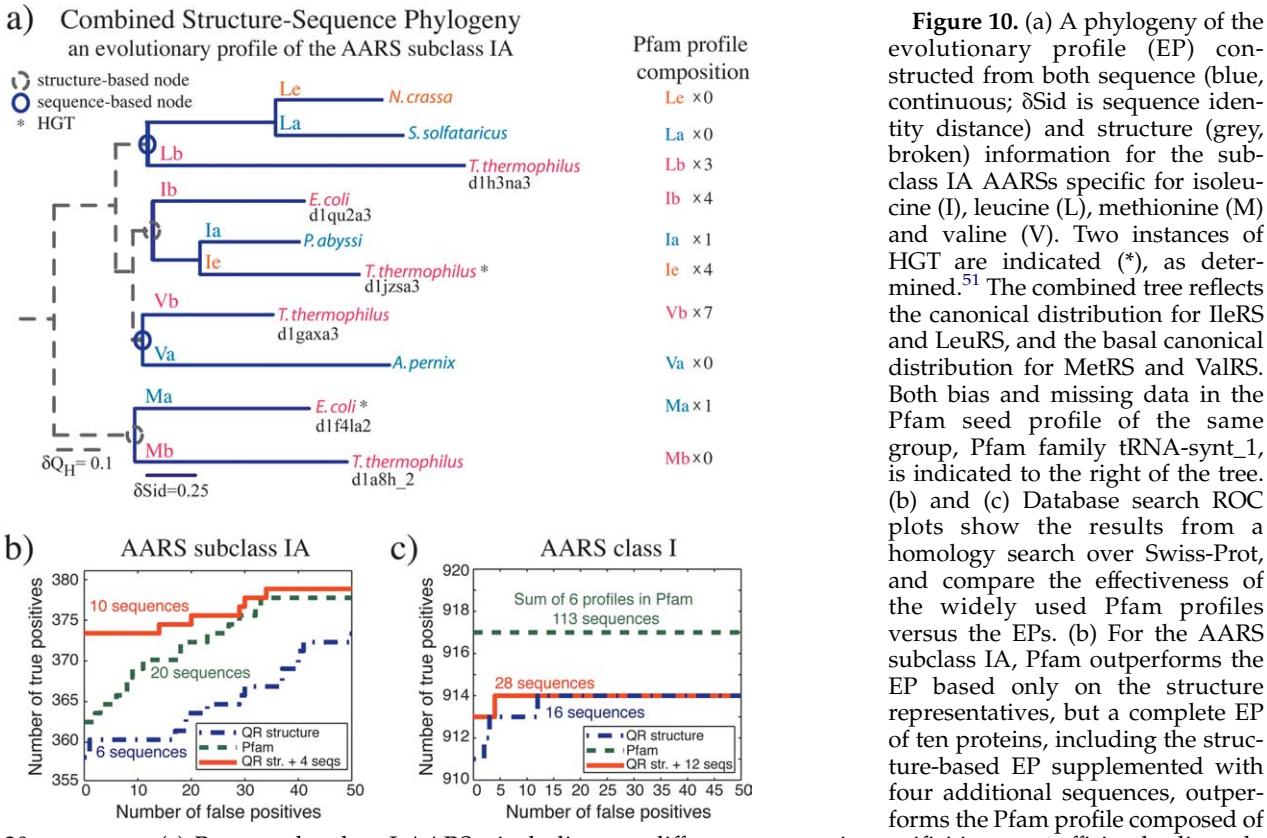

Figure 10. (a) A phylogeny of the evolutionary profile (EP) constructed from both sequence (blue, continuous; δSid is sequence identity distance) and structure (grey, broken) information for the subclass IA AARSs specific for isoleucine (I), leucine (L), methionine (M) and valine (V). Two instances of HGT are indicated (*), as determined.[51] The combined tree reflects the canonical distribution for IleRS and LeuRS, and the basal canonical distribution for MetRS and ValRS. Both bias and missing data in the Pfam seed profile of the same group, Pfam family tRNA-synt_1, is indicated to the right of the tree. (b) and (c) Database search ROC plots show the results from a homology search over Swiss-Prot, and compare the effectiveness of the widely used Pfam profiles versus the EPs. (b) For the AARS subclass IA, Pfam outperforms the EP based only on the structure representatives, but a complete EP of ten proteins, including the structure-based EP supplemented with four additional sequences, outperforms the Pfam profile composed of 20 sequences. (c) Because the class I AARSs, including ten different enzymatic specificities, are sufficiently distantly related, Pfam requires six separate subclass level profiles to describe the class I AARSs. A single database search using our combined structure-sequence EP for the class I AARSs gives a result very close to that of the combined results of six Pfam database searches. The six Pfam database searches find only three additional proteins not found by our single profile search, and these are small protein fragments.

versions of AspRS (tenth position in the order) and L2 (12th position) after the archaeal versions have already been represented (sixth and fourth positions, respectively). Closely related species versions of the major groups are added next, e.g. additional archaeal structures of eIF5α which are 17th and 18th in the order, and lastly different crystal structures of an identical, previously represented protein fill out the complete tree. Some proteins appear to come out of order with respect to the general trend described above. Note that after the AspRS branch is initially represented by the archaeal type (Da) at the sixth position in the ordering, the bacterial version (Db) is added at the tenth position even before IF1 (14th in the order) has been added at all. Although IF1 does not have the same function as its closest relative in the group, namely eIF5α, IF1 has a closely related function to eIF5α. In fact, IF1 is more closely related to eIF5α than Da is to Db, even though Da and Db are of the same function. There are other cases, such as this one and, for example, the relationship between the archaeal AspRS and AsnRS (see Figure 6), where proteins of a different but related function are more closely related than proteins of the same function found in different species.

An economy of information from structure-based evolutionary profiles

The same QR methodology can be applied to a multiple sequence alignment. Instead of Cartesian coordinates describing the protein structures in the alignment matrix, A, protein sequences are described by an orthogonal encoding, with each amino acid type assigned to a unit vector in a 21 space (20 amino acid residues and one gap dimension). This gives an alignment matrix, A, of dimensions $21 \times m_{\text{alignment-length}} \times n_{\text{proteins}}$. The sequence and structure QR factorization algorithms can be used in concert to provide profiles from sequence and structure-based multiple alignments of representative proteins that best span the evolutionary space in both sequence and structure. The complete details of the algorithm are beyond the scope of the current study and will be presented elsewhere.

Here, we provide an example of two combined sequence-structure profiles, the first of which is of the subclass IA tRNA synthetases, including AARSs specific for valine, isoleucine, leucine and methionine. This profile began with a seed structural alignment of the QR factorization based representatives of the known structures giving six

representatives, Lb, Ib, Ie, Vb, Ma and Mb as labeled in Figure 10. Because LeuRS and IleRS display the full canonical phylogenetic pattern, these groups should each be represented by three sequences, one from each of the primary domains-of-life, while MetRS and ValRS, which display the basal canonical pattern, are best represented by only two sequences, one bacterial and one archaeal. This analysis indicates that four additional sequences are required to complete the EP, representing Le, La, Ia and Va. Supplemental sequences were added to the LeuRS, IleRS and ValRS groups by computing a sequence-based QR factorization on multiple sequence alignments of the SWISS-PROT sequences for each of these three groups separately. The sequences of the structure representatives were retained and the next sequence in the QR order, in the cases of IleRS and ValRS, and the next two sequences in the QR ordering of LeuRS were retained as the four supplemental sequences. The second profile encompasses the greater level of diversity of the class I AARS family, including synthetases specific for half of the standard amino acid residues, for a review see the work done earlier.[33] The supplemental sequences for the class I AARS profile were chosen in a procedure similar to that outlined above for the subclass IA profile. Three profiles from both groups were tested in a homology detection search over the SWISS-PROT database, one from the Pfam seed alignment, one of QR factorization based representatives of the known structures, and a structure-based alignment supplemented with sequence representatives, see Figure 10. Although others have recently presented methods to combined sequence and structure-based multiple alignments,[15,58] our goal is to provide combined non-redundant sequence-structure profiles, which we term evolutionary profiles (EPs), that best span the evolutionary space in both sequence and structure.

The subclass IA profile is named tRNA-synt_1 in Pfam, and its composition is shown on the right of the combined sequence-structure phylogeny of this subclass in Figure 10. This composite phylogeny depicts the distances between the QR factorization based representatives of the available structures in regions of more distant similarity; namely, in the so-called "twilight zone" of sequence similarity (less than 20% sequence identity). Above that level of similarity, the tree depicts sequence distances in relationships between orthologs of the same AARS enzymatic specificity. When using the HMMER program[59] for database searching, the bias in the Pfam profiles can be ameliorated with sequence weighting options, but sequence weighting cannot account for missing data. In the Pfam profile of the four AARS enzymatic specificities in the subclass IA, three have phylodomain level representatives missing, such as the lack of archaeal (La) or eukaryotic (Le) representatives in the LeuRS group. Weighting schemes cannot use the bacterial representative to account for sequence motifs or signatures that are idiosyncratic to the eukaryotic or archaeal groups. Default options were used in the HMMER database search.

An accepted test of the effectiveness of a profile is to compute the specificity and sensitivity of the profile in homology searches over large sequence databases. Sensitivity, the number of true positives (within the homologous group), is plotted *versus* the specificity, the number of false positives (outside of the homologous group) found in the database by the profile, and is usually presented as a ROC (receiver-operating characteristics) plot. In Figure 10, we provide results from a search over the SWISS-PROT database[60] using profiles from the subclass IA and the full class I AARSs, where increasingly specific and sensitive profiles find more true positives before the search hits false positives. In the case of the subclass IA group, compared to the Pfam profile[7] for the same group of homologs, we observe improved sensitivity and specificity with our evolutionary profile (EP) by supplementing non-redundant structure-based alignments with an appropriate number of representative sequences. These additional sequences represent major phyletic groups for which there is no three-dimensional structure available. The performance of such a combined profile is shown in Figure 10(b). Most of the true positives that the Pfam profile does not find before hitting false positives are from the Le and Mb groups, two of the major phyletic groups that are not represented in the Pfam profile.

The class I AARSs can only be properly aligned as a single group with the aid of structural alignments, and this profile can be used to find all class I AARS homologs in a sequence database or in unannotated genomes in a single search. Pfam provides profiles at the level of the sublcasses for the class I AARSs, so in order to find all class I AARSs with the Pfam profiles, six database searches are required. We tested the search accuracy of the EP for the class I

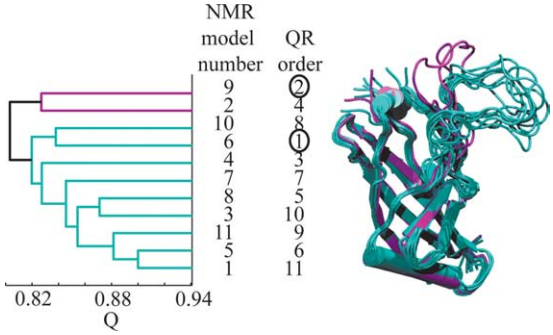

Figure 11. The superposition is shown for the 11 NMR structure models for Small protein B (SmpB), which plays a crucial role in tagging incompletely translated proteins in bacteria.[67] The QR factorization identifies two major conformationally distinct groups, one with the loop in a horizontal position (cyan, first in QR order) and one set of structures with the loop in a vertical position (purple, second in QR order) with respect to the β-barrel. The structure-based dendrogram also indicates the two major conformational groups.

AARSs, against that of the combined results of six database searches with the Pfam profiles. In comparison with the EPs, the Pfam profiles find three additional class I AARS sequence fragments, but at six times the computational expense, see Figure 10(c). A single evolutionarily well-balanced profile of combined sequence-structure multiple alignments performs comparably to a collection of profiles of the more closely related subclasses. Interestingly, the EPs perform better than or comparably with the Pfam profiles, at least in these initial results, but do so with many fewer sequences. Finally, note that the representatives of the EPs were chosen based on the results from the combined sequence-structure QR factorization procedure, without adding additional sequences in an iterative attempt to improve the database search accuracy.

Further applications of the QR algorithm

Representative structures of an NMR ensemble

In addition to the utility of the QR factorization for choosing representatives in structure or sequence which best span an evolutionary space, the algorithm can also be used to generate a representative set of structures which best span a conformational space, such as that explored in an molecular dynamics simulation or from a number of conformers in an NMR ensemble. Sutcliffe investigated the usefulness of representing an NMR structure ensemble by a single structure, either a minimized average structure or a single most representative structure, and the study concluded that it is often best to study the ensemble as a whole.[61] The result from the QR factorization presented here, see Figure 11, indicates that NMR ensembles can be adequately represented, not by a single structure, but by some small number or subset of representative structures which well span the conformational space. Such subsets should be considered when generating an average structure or set of averaged structures for use in a protein structure prediction application or analysis of the conformations observed in an NMR structure determination experiment. The essential dynamics method of Amadei and colleagues[62] uses a diagonalization of a covariance matrix of atomic fluctuations to separate large-scale conformational changes from small fluctuations. The method is similar in spirit to the QR factorization when applied to NMR structures or MD trajectories, but it is not clearly generalizable to treat the presence of gaps in multiple structural alignments as in the evolutionary applications mentioned above. The advantage of the QR algorithm, therefore, is that it is a single technology that is able to deal with conformational or evolutionary changes or both simultaneously.

Incorporating the QR factorization in other bioinformatics applications

The multidimensional QR factorization method described here can also be applied to previously established bioinformatic algorithms. Below, we briefly discuss how the QR algorithm could be used to enhance two recently developed bioinformatics algorithms, the evolutionary trace and the 3dHMM methods. Cohen and colleagues recognized that there are different patterns of sequence conservation at different levels of sequence diversity. While molecules that have the same function in different species should show species-specific differences, two homologous proteins with different functions should show conservation patterns that reflect the changes in, say, active site residues. The evolutionary trace method[63] was designed to automatically categorize such differences by dividing a phylogenetic tree into some number of evolutionary stages, according to functional classifications or a set of evenly spaced sequence identity thresholds. At the border between each evolutionary stage, consensus sequences, i.e. patterns of sequence conservation, are compiled. As the number and size of the partitions of the phylogenetic tree are arbitrarily chosen, this method could be enhanced by selecting the partitions directly from the QR ordering, i.e. adding a new sequence from the order adds a representative of the next partition or evolutionary stage. The number of partitions could then simply be defined by a single upper limit sequence or structural similarity threshold.

Gerstein and co-workers recently developed a hidden Markov model (HMM) method to compute a profile for a multiple alignment of protein structures, 3dHMM.[64] In their method, each aligned position is represented by a gaussian function, centered at the mean position of the aligned C^{α} atoms, and gaps are treated as insertion and deletion states, as in typical sequence-based HMMs. The principal application of 3dHMM is to search structure databases for structural relatives to the group of proteins described by the structure-based profile. The QR factorization could be applied to the set of structures used to train the HMM, thus giving an unbiased model for database searching.

Conclusion

We have presented here a method to obtain evolutionary profiles (EPs) from the multidimensional QR factorization of structural alignments. We have described how the QR factorization can be used to select a spanning set of representatives in a structure, sequence or combined sequence-structure profile and also in an NMR or molecular dynamics conformational ensemble. In addition, by applying a structural similarity measure which accounts for the presence of gaps, an interesting congruence between sequence and structure-based phylogenies results in the AspRS–AsnRS group.

While there is widespread agreement that there is

a need for representative and non-redundant sets of protein sequences and structures, we assert that evolution should be the basis or organizational framework for this type of bioinformatic analysis and, perhaps, for bioinformatics in general. By removal of biasing from the "data deluge" resulting from genome sequencing and structural genomics projects, the EPs offer not only an economy of information but promise to improve the performance of structure and sequence profiles used in database searches. As the EPs are multiple alignments of evolutionarily related groups, they also allow researchers to investigate evolutionary questions. Often structures needed to obtain distant evolutionary profiles are missing, and while we have illustrated how appropriately selected sequences can be incorporated to complete the EP, we hope that the incomplete profiles will motivate the structure determination of the missing proteins that represent major evolutionary transitions.

On a final note, this work has been carried out at the domain level, which is the appropriate length scale because the protein domain is the largest common evolutionarily shared segment between the proteins we investigated. As we look into more recent evolutionary events we should move to the multi-domain level, while the most distant evolutionary events, yet to be probed, may be better addressed on a length scale smaller than even that of the single domain.

Acknowledgements

We are grateful to Carl Woese for stimulating discussions on the aminoacyl-tRNA synthetases and other evolutionary matters, and many thanks are due to Michael Heath for discussions concerning the QR factorization. We thank John Eargle for coding the algorithms presented here into a new multiple structural alignment feature in VMD version 1.8.3,[35] and Anurag Sethi for providing the analysis detailed in Figure 10. P.O'D. was supported on an NIH Institutional NRSA in Molecular Biophysics (5T32GM08276) with additional support from the NSF grant MCB04-46227.

Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2004.11.053

References

1. Astbury, W. T. (1952). *Harvey Lectures 1950–51*, Thomas, Springfield.
2. Zuckerkandl, E. & Pauling, L. (1965). Molecules as documents of evolutionary history. *J. Theor. Biol.* **8**, 357–366.
3. DeLong, E. F. & Pace, N. R. (2001). Environmental diversity of bacteria and archaea. *Syst. Biol.* **50**, 470–478.
4. Woese, C. R. & Fox, G. E. (1977). Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc. Natl Acad. Sci. USA*, **74**, 5088–5090.
5. Darwin, C. (1887). *The Life and Letters of Charles Darwin, Including an Autobiographical Chapter*, John Murray, London.
6. Feng, D. F. & Doolittle, R. F. (1987). Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J. Mol. Evol.* **25**, 351–360.
7. Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S. *et al*. (2004). The Pfam protein families database. *Nucl. Acids Res.* **32**, D138–D141.
8. Pandit, S. B., Bhadra, R., Gowri, V. S., Balaji, S., Anand, B. & Srinivasan, N. (2004). Supfam: a database of sequence superfamilies of protein domains. *BMC Bioinformat.* **5**, 28.
9. Stebbings, L. A. & Mizuguchi, K. (2004). HOMSTRAD: recent developments in the homologous protein structure alignment database. *Nucl. Acids Res.* **32**, D203–D207.
10. Gribskov, M., McLachlan, M. & Eisenberg, D. (1987). Profile analysis: detection of distantly related proteins. *Proc. Natl Acad. Sci. USA*, **84**, 4355–4358.
11. Sonnhammer, E. L. L., Eddy, S. R. & Durbin, R. (1997). Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins: Struct. Funct. Genet.* **28**, 405–420.
12. Lupas, A. N., Ponting, C. P. & Russell, R. B. (2001). On the evolution of protein folds: are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world? *J. Struct. Biol.* **134**, 191–203.
13. Sadreyev, R. & Grishin, N. (2003). COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance. *J. Mol. Biol.* **326**, 317–336.
14. Tang, C. L., Xie, L., Koh, I. Y. Y., Posy, S., Alexov, E. & Honig, B. (2003). On the role of structural information in remote homology detection and sequence alignment: new methods using hybrid sequence profiles. *J. Mol. Biol.* **334**, 1043–1062.
15. O'Sullivan, O., Suhre, K., Abergel, C., Higgins, D. G. & Notredame, C. (2004). 3DCoffee: combining protein sequences and structures within multiple sequence alignments. *J. Mol. Biol.* **340**, 385–395.
16. Tramontano, A. & Morea, V. (2003). Assessment of homology-based predictions in CASP5. *Proteins: Struct. Funct. Genet.* **53**, 352–368.
17. Petrey, D., Xiang, Z., Tang, C. L., Xie, L., Gimpelev, M., Mitros, T., Soto, C. S. *et al*. (2003). Using multiple structure alignments, fast model building, and energetic analysis in fold recognition and homology modeling. *Proteins: Struct. Funct. Genet.* **53**, 430–435.
18. Bickel, P. J., Kechris, K. J., Spector, P. C., Wedemayer, G. J. & Glazer, A. N. (2002). Finding important sites in protein sequences. *Proc. Natl Acad. Sci. USA*, **99**, 14764–14771.
19. Shakhnovich, B. E., Dokholyan, N. V., DeLisi, C. & Shakhnovich, E. I. (2003). Functional fingerprints of folds: evidence for correlated structure-function evolution. *J. Mol. Biol.* **326**, 1–9.

20. Hardin, C., Pogorelov, T. V. & Luthey-Schulten, Z. (2002). *Ab initio* protein structure prediction. *Curr. Opin. Struct. Biol.* **12**, 176–181.
21. Russ, W. P. & Ranganathan, R. (2002). Knowledge-based potential functions in protein design. *Curr. Opin. Struct. Biol.* **12**, 447–452.
22. Amann, R. I., Ludwig, W. & Schleifer, K. H. (1995). Phylogenetic identification and *in situ* detection of individual microbial cells without cultivation. *Microbiol. Rev.* **59**, 143–169.
23. Pace, N. R. (1997). A molecular view of microbial diversity and the biosphere. *Science*, **276**, 734–740.
24. Tyson, G. W., Chapman, J., Hugenholtz, P., Allen, E., Ram, R. J., Richardson, P. M. *et al*. (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, **428**, 37–43.
25. Venter, J. C., Remington, K., Heidelberg, J. F., Halpern, A. L., Rusch, D., Eisen, J. A. *et al*. (2004). Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, **304**, 66–74.
26. Holm, L. & Sander, C. (1998). Removing near-neighbour redundancy from large protein sequence collections. *Bioinformatics*, **14**, 423–429.
27. O'Donovan, C., Martin, M. J., Glemet, E., Codani, J.-J. & Apweiler, R. (1999). Removing redundancy in SWISS-PROT and TrEMBL. *Bioinformatics*, **15**, 258–259.
28. Wang, G. & Dunbrack, R. L., Jr (2003). PISCES: a protein sequence culling server. *Bioinformatics*, **19**, 1589–1591.
29. Park, J., Holm, L., Heger, A. & Chothia, C. (2000). RSDB: representative protein sequence databases have high information content. *Bioinformatics*, **16**, 458–464.
30. Vingron, M. & Argos, P. (1989). A fast and sensitive multiple sequence alignment algorithm. *Comput. Appl. Biosci.* **5**, 115–121.
31. Heniko, S. & Heniko, J. G. (1994). Position-based sequence weights. *J. Mol. Biol.* **243**, 574–578.
32. May, A. C. W. (2001). Optimal classification of protein sequences and selection of representative sets from multiple alignments: application to homologous families and lessons for structural genomics. *Protein Eng.* **14**, 209–217.
33. O'Donoghue, P. & Luthey-Schulten, Z. (2003). Evolution of structure in aminoacyl-tRNA synthetases. *Microbiol. Mol. Biol. Rev.* **67**, 550–573.
34. Russell, R. B. & Barton, G. J. (1992). Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. *Proteins: Struct. Funct. Genet.* **14**, 309–323.
35. Humphrey, W., Dalke, A. & Schulten, K. (1996). VMD–visual molecular dynamics. *J. Mol. Graph.* **14.1**, 33–38.
36. Eastwood, M. P., Hardin, C., Luthey-Schulten, Z. & Wolynes, P. G. (2001). Evaluating protein structure-prediction schemes using energy landscape theory, IBM. *J. Res. Dev.* **45**, 475–497.
37. Koretke, K., Luthey-Schulten, Z. & Wolynes, P. (1996). Self-consistently optimized statistical mechanical energy functions for sequence structure alignment. *Protein Sci.* **5**, 1043–1059.
38. Golub, G. H. & Van der Vorst, H. A. (2000). Eigenvalue computation in the 20th century. *J. Comput. Appl. Math.* **123**, 35–65.
39. Householder, A. S. (1958). Unitary triangularization of a nonsymmetric matrix. *JACM*, **5**, 339–342.
40. Heath, M. T. (2002). *Scientific Computing: An Introductory Survey,* (2nd edit.), McGraw-Hill, New York.
41. Golub, G. (1965). Numerical methods for solving linear least squares problems. *Numer. Math.* **7**, 206–216.
42. Wilkinson, J. H. (1965). *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford.
43. Rogen, P. & Bohr, H. (2003). A new family of global protein shape descriptors. *Math. Biosci.* **182**, 167–181.
44. Olkin, J. A., Heck, L. P. & Naghshineh, K. (1996). Automated placement of transducers for active noise control: performance measures. In *IEEE Proceedings of International Conference on Acoustics, Speech, and Signal Processing, Atlanta, GA*, vol. 2, pp. 969–972.
45. Heck, L. P., Olkin, J. A. & Naghshineh, K. (1998). Transducer placement for broadband active vibration control using a novel multidimensional qr factorization. *J. Vib. Acoust.* **120**, 663–670.
46. Sethi, A., O'Donoghue, P. & Luthey-Schulten, Z. (2005). Evolutionary profiles from the QR factorization of multiple sequence alignments. *Proc. Natl Acad. Sci. USA,* in the press.
47. Felsenstein, J. (1989). PHYLIP–phylogeny inference package (version 3.2). *Cladistics*, **5**, 164–166.
48. Sokal, R. R. & Michener, C. D. (1958). A statistical method for evaluating systematic relationships. *Univ. Kansas Sci. Bull.* **28**, 1409–1438.
49. Pauling, L. (1964). Molecular disease and evolution. *Bull. NY Acad. Med.* **40**, 334–342.
50. Woese, C. R. (2002). On the evolution of cells. *Proc. Natl Acad. Sci. USA*, **99**, 8742–8747.
51. Woese, C. R., Olsen, G., Ibba, M. & Söll, D. (2000). Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process. *Microbiol. Mol. Bio. Rev.* **64**, 202–236.
52. Marti-Renom, M. A., Stuart, A., Fiser, A., Sanchez, R., Melo, F. & Sali, A. (2000). Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.* **29**, 291–325.
53. Chothia, C. & Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5**, 823–826.
54. Gan, H. H., Perlow, R. A., Roy, S., Ko, J., Wu, M., Huang, J. *et al*. (2002). Analysis of protein sequence/structure similarity relationships. *Biophys. J.* **83**, 2781–2791.
55. Kyrpides, N. C. & Woese, C. R. (1998). Universally conserved translation initiation factors. *Proc. Natl Acad. Sci. USA*, **95**, 224–228.
56. Kyrpides, N. C. & Woese, C. R. (1998). Archaeal translation initiation revisited: the initiation factor 2 and eukaryotic initiation factor 2b alpha–beta–delta subunit families. *Proc. Natl Acad. Sci. USA*, **95**, 3726–3730.
57. Theobald, D. L., Mitton-Fry, R. M. & Wuttke, D. S. (2003). Nucleic acid recognition by OB-fold proteins. *Annu. Rev. Biophys. Biomol. Struct.* **32**, 115–133.
58. Al-Lazikani, B., Sheinerman, F. B. & Honig, B. (2001). Combining multiple structure and sequence alignments to improve sequence detection and alignment: application to the SH2 domains of Janus kinases. *Proc. Natl Acad. Sci. USA*, **98**, 14796–14801.
59. Eddy, S. R. (1998). Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
60. Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.-C., Estreicher, A., Gasteiger, E. *et al*. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucl. Acids Res.* **31**, 365–370.

61. Sutcliffe, M. J. (1993). Representing an ensemble of NMR-derived protein structures by a single structure. *Protein Sci.* **2**, 936–944.
62. Van Aalten, D. M. F., de Groot, B. L., Findlay, J. B. C., Berendsen, H. J. C. & Amadei, A. (1997). A comparison of techniques for calculating protein essential dynamics. *J. Comput. Chem.* **18**, 169–181.
63. Lichtarge, O., Bourne, H. R. & Cohen, F. E. (1996). An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.* **257**, 342–358.
64. Alexandrov, V. & Gerstein, M. (2004). Using 3D hidden Markov models that explicitly represent spatial coordinates to model and compare protein structures. *BMC Bioinformat.* **5**, 2.
65. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536–540.
66. Chandonia, J. M., Hon, G., Walker, N. S., Lo Conte, L., Koehl, P., Levitt, M. & Brenner, S. E. (2004). The ASTRAL compendium in 2004. *Nucl. Acids Res.* **32**, D189–D192.
67. Someya, T., Nameki, N., Hosoi, H., Suzuki, S., Hatanaka, H., Fujii, M. *et al*. (2003). Solution structure of a tmRNA-binding protein, SmpB, from *Thermus thermophilus*. *FEBS Letters*, **535**, 94–100.

Edited by C. R. Matthews

(Received 11 August 2004; received in revised form 11 November 2004; accepted 17 November 2004)