Current Topics in Biochemical Research
Vol. 16, 2014

Review

Assessing the contribution of coevolving residues to the stability of proteins by computational means

Domenico L. Gatti[1,2,*]

[1]Department of Biochemistry and Molecular Biology, [2]Cardiovascular Research Institute, Wayne State University School of Medicine, Detroit, Michigan, USA.

ABSTRACT

In the multiple sequence alignment (MSA) of a protein family, non-conserved positions can be very important because the destabilizing effects of a given amino acid at one position can be compensated by the stabilizing effect of another amino acid at a different position. As a consequence these positions are often coevolving. Several methods are available for the detection of coevolving positions from the analysis of MSAs. Information about coevolution in combination with information on the changes in folding free-energy produced by all possible point mutations of each residue can be very valuable to understand the protein mechanism and dynamic properties, and to design mutagenesis studies. Using an example based on the family of KDO8P synthase, an enzyme involved in the synthesis of bacterial endotoxin, we describe a general strategy to assess the contribution of coevolving residues to protein stability by computational means.

KEYWORDS: coevolution, covariation, folding free energy, methods, multiple sequence alignments, stability

INTRODUCTION

The net stabilization of the folded state of proteins relative to the unfolded state is usually so small, that all positions, conserved and non-conserved, contribute to protein stability. The existence of physical and functional interactions between sites in protein sequences leads to non-independence of their evolution; in other words, two (or more) positions in a protein sequence could be coevolving, and for any mutation to become fixed at such sites, compensatory mutations are needed at the related sites. When trying to extract the coevolution history of different residues inside a protein from the multiple sequence alignment (MSA) of its family it is important to realize that a high background of different interacting factors often hides the coevolutionary relationships between amino acid sites. A simple model to explain the correlation C_{ij} between two sites *i* and *j* in a sequence alignment was proposed by Atchley *et al.* [1, 2]:

$$C_{ij} = C_{phylogeny} + C_{structure} + C_{function} + C_{interaction} + C_{stochastic} \quad (1)$$

In this model $C_{phylogeny}$ is the correlation originating from phylogenetic relationships between homologous sequences that belong to the same branch of an evolutionary tree. For example, a mutation in an ancestral protein, which is clearly a single evolutionary event, appears in the MSA as an independent event that occurred in each of the proteins that descended from that ancestor. $C_{structure}$ and $C_{function}$ represent the correlation originating from structural and functional constraints. $C_{interaction}$ describes both the interaction between the aforementioned sources of correlation, and the correlation originating from atomic interactions in homo-oligomeric proteins. Finally, $C_{stochastic}$ represents the correlation originating from casual co-variation and/or from uneven or incomplete sequence sampling. Low-quality and poorly populated MSAs typically produce a high

*Email id: dgatti@med.wayne.edu

degree of false coevolution signals as a result of the significant effect of stochasticity [3-5].

A wide variety of algorithms have been developed to detect coevolving positions from an MSA (reviewed in [6-11]). Some of these methods use χ^2-tests [12, 13], some are perturbative [14-16], others employ amino acid substitution matrices [17], and many work within the frame of information theory [18-32]. We recall here that information entropy, $H(X)$, is a measure of the uncertainty associated with a discrete random variable X that assumes values $\{x_1, ..., x_n\}$:

$$H(X) = -\sum_{x \in X} p(x) \log_b p(x) \quad (2)$$

where b is the base of the logarithm used and p is the probability mass function of the variable X [33, 34]. Related to $H(X)$, mutual information, $MI(X;Y)$, measures the mutual dependence of two discrete random variables X and Y:

$$MI(X;Y) = MI(Y;X) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log_b \frac{p(x,y)}{p(x)p(y)} \quad (3)$$

where $p(x,y)$ is the joint probability mass function of X and Y, and $p(x)$ and $p(y)$ are the marginal probability mass functions of X and Y, respectively. Intuitively, MI measures how much knowing one of the two variables reduces the uncertainty about the other. In an MSA, the amino acids in a given column can be considered as a set of observations (x_i) of a random variable X. An estimate of the entropy $H(X)$ is obtained by using the observed amino acid frequencies, $f(x_i)$, in place of the underlying probabilities, $p(x_i)$; likewise, $MI(X;Y)$ for a pair of columns can be derived using the frequencies, $f(x_i,y_j)$, of all ordered pairs occurring in the two columns. In practice, MI between positions (columns in an MSA) reflects the extent to which knowledge of the amino acid at one position allows us to predict the identity of the amino acid at the other position [2, 20, 21]. If amino acids occur independently at the two sites, the theoretical value for MI is zero; conversely, MI is high if the two positions are correlated.

A problem shared by most coevolution detection methods is that many structurally distant pairs appear to be strongly correlated. One source of this correlation is the propagation of statistical dependencies along chains of co-evolving contacts [15, 35]. In general, given residues A,B,C, when pair AB and pair BC are among the top pairs and represent true structural contacts based on protein geometry, we may find pair AC as highly covarying (yet distant in geometric structure) as an induced coupling produced by pairs AB and BC. This kind of induced coupling can extend along chains of contacts. For example if A contacts B, B contacts C, C contacts D, ...N-1 contacts N, covariation maps may show some level of covariation between A and N. On the other hand, two positions do not have to be part of a chain of contacts to appear correlated; what is important in these cases of covariation is not the presence of a direct physical interaction, but the fact that residues exposed to like forces (e.g, the hydrophobic interior or the hydrophilic surface), respond in a correlated fashion to the changes that affect the global fit function (which includes both stability and mechanism). Disentanglement of these two types of interactions (local/direct *versus* global/indirect) was attempted with the MIp [22], Zres [24] and Zpx [36] corrections of mutual information (MI) statistics, with the application of Bayesian network modeling in the logR method [26], with Direct Coupling Analysis (DCA) [28, 29, 32, 37], a maximum entropy method, with the use of sparse inverse covariance estimation in PSICOV [27], by employing a pseudolikelihood approach with plmDCA [29], gplmDCA [30], and GREMLIN [31], or by extending mutual information from 2 to 3 dimensions, as in 3D_MI [38].

From the point of view of protein stability the organization of enzyme active sites is inherently unstable because these sites are optimized for catalysis, which means they are pre-organized to stabilize the transition state(s), rather than the protein [39, 40]. Thus, the substitution of a catalytic side chain (most often to alanine) will typically increase the overall protein stability, while sacrificing function [41, 42]. Conversely, most mutations that introduce a new function are destabilizing [43, 44]. The generality of this stability-function tradeoff must be viewed within the context of the fact that regardless of their effect on functions most mutations are destabilizing [45-48].

In this context, it becomes natural to ask in what way are correlated mutations at different sites in a

protein affecting protein stability. During the past few years we have addressed this question as it applies to the family of KDO8P synthase (KDO8PS), a bacterial enzyme that synthesizes KDO8P from phosphoenolpyruvate (PEP) and arabinose 5-phosphate (A5P). This reaction is of significant biological relevance, as KDO8P is the phosphorylated precursor of KDO, which is an essential component of the endotoxin of Gram negative bacteria [49]. A combination of tools from information theory and structural modeling has provided an avenue to quantify the contribution of coevolving residues to the stability of KDO8P synthase [50]. We review here the methodology used in this study as it may be generally applicable to all protein families.

METHODS

Multiple sequence alignments

Multiple sequence alignments (MSAs) of 1056 sequences of KDO8PS were calculated independently with Muscle [51], and Mafft [52] and then merged together with T-Coffee [53].

Molecular dynamics simulations

A complete three-dimensional model of *Nm.* KDO8PS was built with Prime 2.1 (Schrodinger, LLC) using primarily the X-ray structure of *Nm.* KDO8PS (PDB 2QKF) as template, and that of *Aquifex aeolicus* KDO8PS (PDB 1FWW) to build the residues missing in the *Nm.* structure. MD simulation were carried out with Desmond (D.E. Shaw Research) [54].

Coevolution analysis

Covariation scores were calculated with the MSAvolve v3.0a Toolbox for Matlab available for download at our website (http://146.9.23.191/~gatti/coevolution/).

FoldX calculations of protein stability

$\Delta\Delta G$ changes associated with introducing any one of the 20 possible amino acids at each position in all four monomers of a tetramer (the biological unit) of *Nm.* KDO8PS were calculated with FoldX v3.0b4 [46, 55] following the procedure described in [56]. Each calculation was carried out in duplicate to ensure convergence: in this case the FoldX algorithm repeated the same mutations twice changing the rotamer set used and the order of moves such that alternative solutions could be explored.

A general strategy to assess by computational means the contribution of coevolving residues to protein stability

Obtaining a good MSA

The first step in coevolution analysis is to obtain the best possible multiple sequence alignment of the protein family of interest. In our case, sequences for 8753 members of the KDO8PS family were downloaded as a single 'fasta' file from the UniProt database (http://www.uniprot.org/help/uniprotkb) [57]. Many of these sequences were highly redundant (for example originating from different strains of the same organisms), or incorrectly included in the family, and it was necessary to screen the sequences contained in the downloaded file based on the selection of a 'reference' sequence. In our example, we used the structure of KDO8PS from *Neisseria meningitidis* (*Nm.*) (Uniref Q9JZ55, PDB 2QKF) [58]), as the representative structure of the family, and its sequence as the 'reference' sequence. The original dataset was reduced to a reliable 'core' by retaining only the regions of each sequence similar to the reference sequence, by removing outliers (sequences with less than 30% average alignment accuracy with all the other sequences in the dataset), and by reducing redundancy (the number of sequences that would allow pairwise alignments with no more than a given percentage of identity) to 95%. At the end of this screening the number of KDO8PS sequences decreased from 8753 to 1111. Independent MSAs were then calculated with Muscle [51], and Mafft [52], and then merged with T-Coffee [53]. While most of the sequences in the alignment were of length comparable to that of the *Nm* protein (280 residues), the merged MSA contained 474 columns due to the presence of multiple gaps. The MSA was then 'trimmed' in order to remove all the columns that did not correspond to a residue of the reference sequence, and sequences with too many gaps were removed leading to a final MSA consisting of 1056 rows (out of the initial 8753), and 280 columns.

Calculating coevolution maps and validating against experimental distance maps

We start by noticing that the words 'coevolution' and 'covariation' are often used interchangeably in this type of studies because the statistical 'covariation' of amino acid symbols in the columns of an MSA

is taken as an indication of the 'coevolution' of those positions in the protein family. Recently we have reviewed the performance of several methods [9, 38] for covariation analysis with both experimental and synthetic data sets, and found there is significant variability in their performance with different proteins [9]. Since a large fraction of the positions that coevolve are due to residues that are close to each other in space (local/direct interactions), one can expect significant overlap between the coevolution map of a protein family and the contact map of a representative 3D structure for that family. The strong statistical correlation between coevolving positions and positions that are within 8 Å of each other in space, but are separated by at least 10 positions along the linear sequence, forms the basis for the spectacular structure predictions that have been recently accomplished with the DCA and PSICOV methods using only sequence data [59-61].

It is also important to realize that even algorithms whose overall performance with a given protein family is similar on a statistical basis, share no more than 2/3 of all the pairs among the top covariation scores [38]. For these reasons we recommend analyzing the MSA of interest with several methods, and then selecting the most effective one based on the method capacity to predict the close contacts observed in the representative X-ray structure.

In our study of KDO8PS, coevolution maps were calculated with 3D_MI [38], hpPCA [32], plmDCA [29], and GREMLIN [31]. We refer the readers to the original reports for details of the algorithms used in each method. Programs implementing these algorithms can be downloaded from the respective authors websites, but are also available (with their original unmodified code) in our Matlab Toolbox MSAvolve v3.0a (http://146.9.23.191/~gatti/coevolution/) for the simulation and analysis of coevolution in proteins. 3D_MI applies by default first an 'average product correction' (APC or MIp correction) [22] to remove entropic and phylogenetic bias, and then a ZPX correction [4, 24, 25, 62], which further improves the accuracy of covariation detection particularly in MSAs containing some degree of misalignment. In contrast, plmDCA, hpPCA, and GREMLIN, only apply by default the APC correction. We have applied a ZPX correction also to these methods, as without it their performance is significantly decreased.

At this time we do not recommend using some older covariation methods including Observed Minus Expected Squared Covariance (OMES) [12, 13], McLachlan Based Substitution Correlation (McBASC) [17, 63], Explicit Likelihood of Subset Co-variation (ELSC) [14], and Statistical Coupling Analysis (SCA) [15, 16, 64], which were shown in a recent survey [9] not to be very effective in the detection of true covarying pairs.

All four methods used produced covariation maps that superimposed well with the contact map derived from the X-ray structure of *Nm.* KDO8PS (Figure 1). To quantify the detection of close contacts, we measured what fraction of all residue pairs separated by less than 8 Å in the X-ray structure was represented in the top covarying pairs identified by each method. A number of pairs equal to the number of residues L in each sequence was considered. This result was further filtered to include either all the pairs (Figure 2A) or only pairs whose component residues are separated by at least 20 positions in sequence space (Figure 2B): in general, correct prediction of pairs separated by at least 20 positions in sequence space is a better indicator of the method performance than the prediction of all pairs. In the example shown here, similar results were obtained with all four methods, but plmDCA and 3D_MI were about 5 times faster than GREMLIN or hpPCA.

Calculating the protein stability landscape

In an earlier study we have used the experimentally validated FoldX algorithm [46, 47, 55] to calculate the folding ΔΔG changes associated with introducing any one of the 20 possible amino acids at each position of the structure of *Nm.* KDO8PS [50]. This type of calculation was initially introduced by Tokuriki *et al.* [65] to study the overall distribution of stability effects for all possible mutations in a large set of different single domain globular proteins. In our case we used an entire tetramer of *Nm.* KDO8PS, which is known to be the biological unit of the enzyme [58], and the individual mutations were introduced simultaneously in all four subunits. Thus, the calculated ΔΔG changes account also for the effects of mutations at the interface between subunits. Furthermore, as ΔΔG changes can be dependent on a particular conformation of the enzyme trapped in the crystal environment, the three-dimensional model of tetrameric *Nm.* KDO8PS derived from the X-ray structure [58] was relaxed by means of a molecular dynamics (MD) simulation under solvated conditions [50]. The MD run progressively eliminated possible errors in the original model and

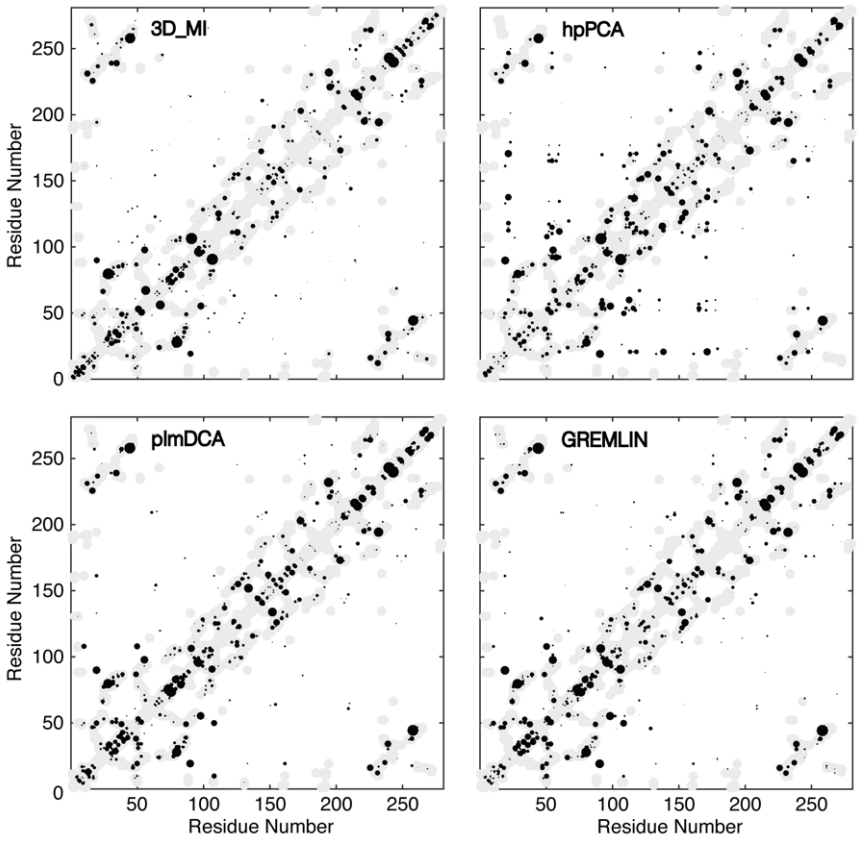

Figure 1. Correspondence between the distance map of *Nm.* KDO8PS and the coevolution maps of the KDO8PS family obtained with different methods. Contact predictions by 3D_MI, hpPCA, plmDCA, and GREMLIN are shown as spots of size proportional to the covariation score. Gray regions represent the native distance map of *Nm.* KDO8PS X-ray structure with a cutoff of 8 Å on the distance between the centroids of different residues.

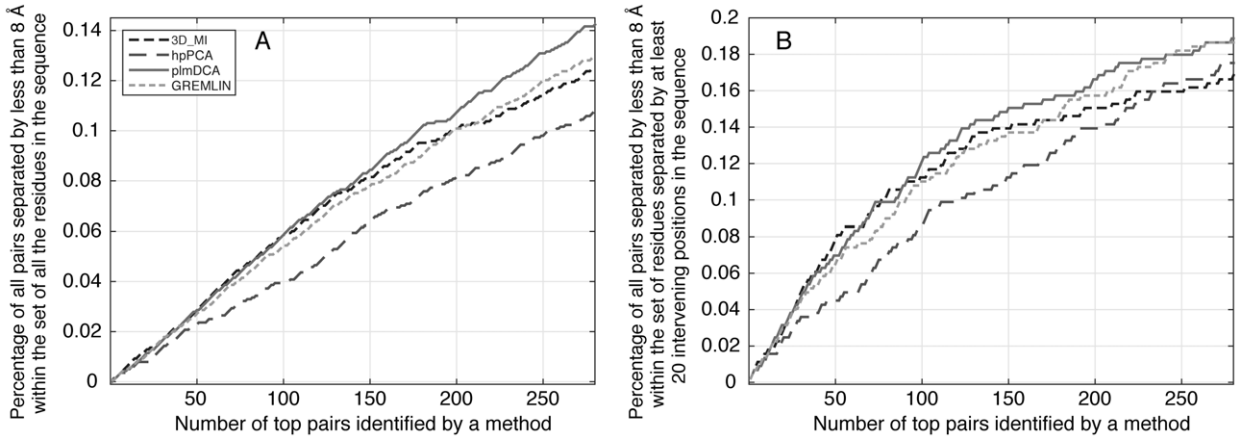

Figure 2. Detection of close contacts by coevolution maps. A. Each trace shows what fraction of all residue pairs separated by less than 8 Å in the reference X-ray structure is present in the top *L* covarying pairs identified by each method. **B.** Only pairs whose residues are separated by at least 20 intervening positions in sequence space are included in the analysis. True positives (covarying pairs corresponding to structural pairs < 8 Å apart) appear as upward displacements in the traces; false positives appear as horizontal segments in the traces.

assured that the equilibrium structure of *Nm.* KDO8PS used in the FoldX calculation was as close as possible to the native structure in solution.

The outcome of this calculation was a 2D-matrix of ΔΔG values that depicts the "stability landscape" of KDO8PS (Figure 3); the peaks in the landscape represent positions in the protein where introduction of a certain amino acid in the *Nm.* KDO8PS would significantly increase the folding free-energy ΔG, and therefore decrease the overall stability. While the energies derived from FoldX are not on an absolute scale [65], the relative trends are expected to be correct [66, 67]. In general, it can be seen how bulky aromatic residues (W,Y,F,H) tend to decrease stability (increase energy) at every position, and in four positions (21,68,231,232) any residue besides glycine or alanine decreases stability dramatically. These effects appear to be due to very large energy terms derived from van der Waals clashes of these residues with their surroundings.

Calculating the contribution of coevolving positions to protein stability

In order to determine the effect of coevolution on stability we compared the covariation score of each pair with the contribution to stability by that pair

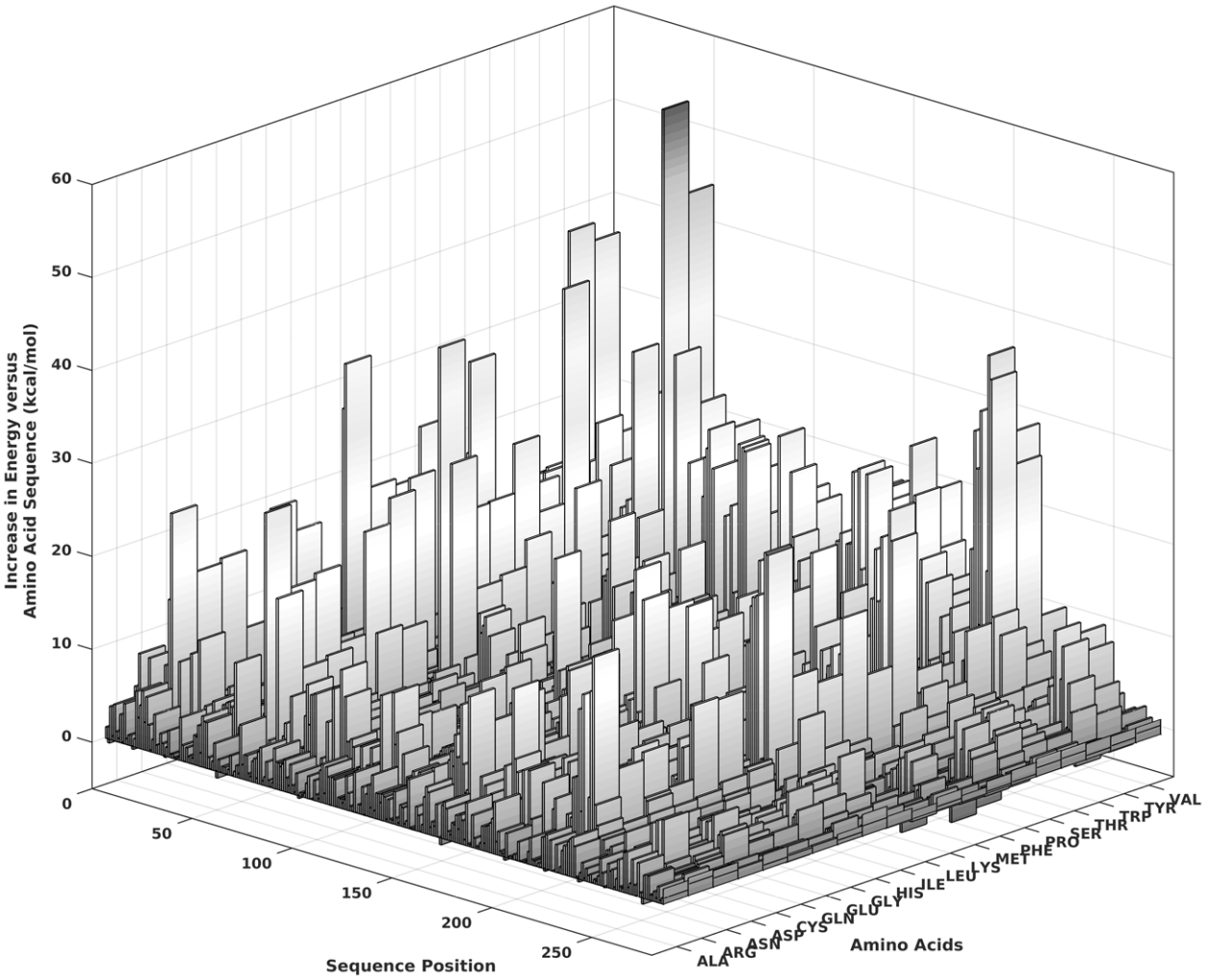

Figure 3. Stability landscape of KDO8PS. ΔΔG changes associated with introducing any one of the 20 possible amino acids at each position of all four subunits of the structure of *Nm.* KDO8PS were calculated using the FoldX algorithm to produce a "stability landscape" of KDO8PS: the peaks in the landscape represent positions in the protein where introduction of a certain amino acid would significantly decrease the overall stability.

in each sequence of the MSA with respect to the *Nm.* sequence; this was accomplished by applying to each sequence in the MSA the information contained in the stability matrix shown in Figure 3. For example, based on that matrix inserting a proline into *Nm.* KDO8PS would increase the folding free-energy by ~5 kcal/mol at position 31, and decrease it by ~1 kcal/mol at position 58. Thus, for every position of KDO8PS we can calculate the folding free-energy change associated with converting the *Nm.* sequence

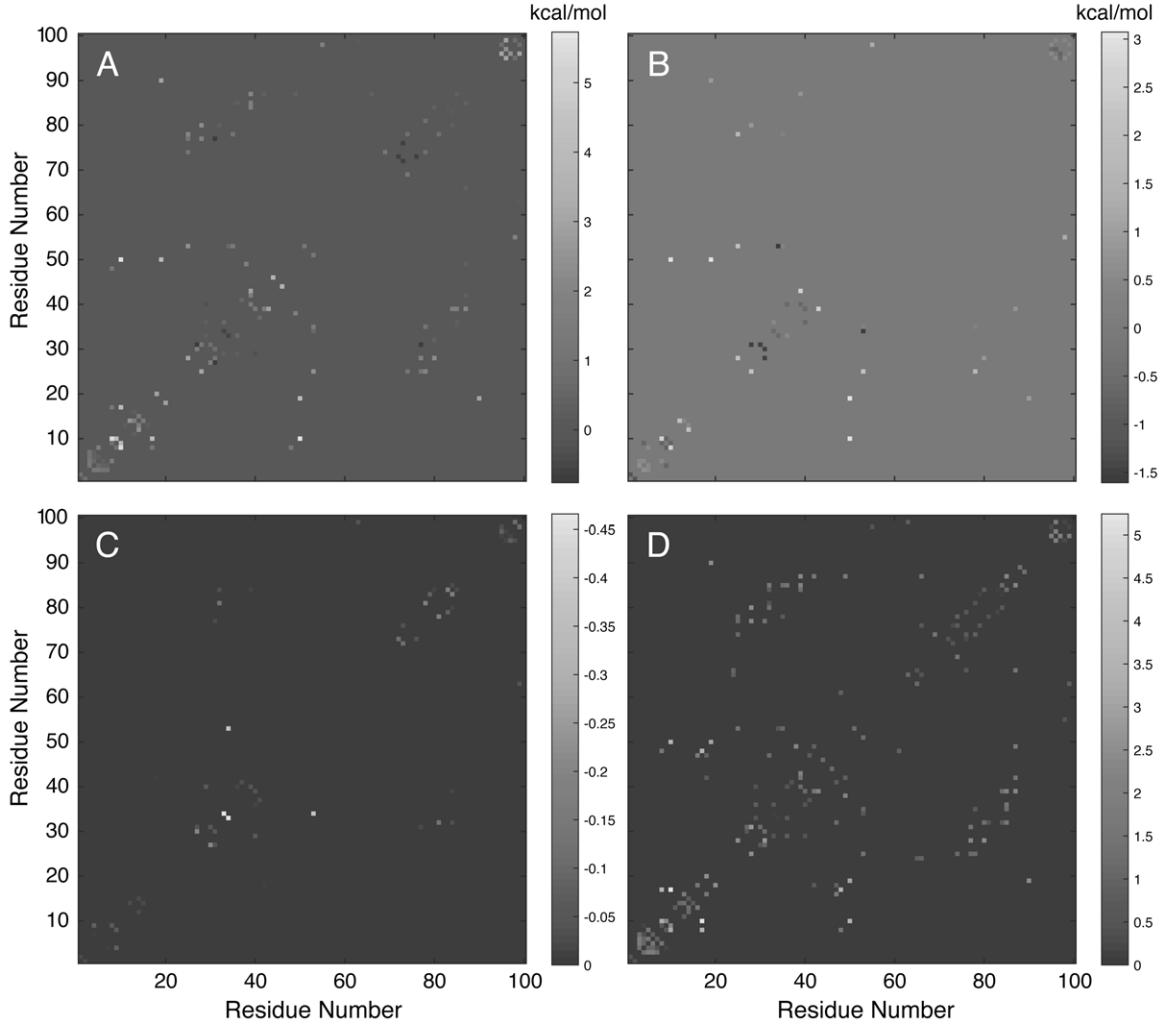

Figure 4. Contribution of coevolving positions to the stability of *Nm.* KDO8PS. A. The first 100 residues of the 3D_MI coevolution map are shown with gray shades proportional to the ΔΔG values for the transition *Nm.* → *Hh.* KDO8PS, as calculated from the stability landscape shown in Figure 3. Positive and negative values of the *i*,*j* pairs indicate increase (destabilization) or decrease (stabilization) in energy, respectively. **B.** The same 100 residues of the 3D_MI coevolution map are shown with gray shades proportional to the values of $\Delta\Delta G_{ij} = (|\Delta\Delta G_i + \Delta\Delta G_j| - |\Delta\Delta G_i - \Delta\Delta G_j|)$. If the value of this difference is larger than 0, it means that the energy changes at the two positions of the *i*,*j* pairs point in the same direction (the two mutations are enhancing each other regardless of whether they are stabilizing or destabilizing); conversely if the value of this difference is smaller than 0, it means that the energy changes at the two positions of the *i*,*j* pairs point in opposite direction (the two mutations are partially neutralizing each other). **C.** Average effect on stability of a pair change from its composition in *Nm.* KDO8PS to its composition in a subset of sequences in the MSA: only pairs whose covariation would produce an increase in stability are shown. **D.** Same as **C**, but only pairs whose covariation would produce a decrease in the stability of the *Nm.* protein are shown.

to any other sequence in the MSA. This information can be stored in a 3D-matrix, in which the 1st and 2nd dimensions provide the *i,j* identity of a covarying pair, and the 3rd dimension provides the free-energy change associated with changing that pair from its composition in *Nm.* KDO8PS to its composition in every other sequence. Since we are interested in coevolving positions we can fill the matrix for every possible *i,j* pair of residues that coevolution maps (calculated by any of the four methods) show to be highly covarying. We are interested not only in the sum ($\Delta\Delta G_i + \Delta\Delta G_j$) of the stability contributions of each member of an *i,j* pair (as these contributions can be considered approximately additive), but also in their difference ($\Delta\Delta G_i - \Delta\Delta G_j$), which provides additional information on whether the two contributions have similar or opposite effects. For example, we could look at the predicted energy changes associated with mutating the pairs with the highest covariation scores in the 3D_MI map from the *Nm.* sequence to the sequence of *Halobacteroides halobius* (*Hh*.) KDO8PS. For ease of visualization we show in Figure 4A only the first 100 residues of the 3D_MI coevolution map gray shaded according to $\Delta\Delta G$ values for the transition *Nm.* $\rightarrow$ *Hh.* KDO8PS. Positive and negative values of the *i,j* pairs indicate increase (destabilization) or decrease (stabilization) in energy, respectively. In Figure 4B the values of $\Delta\Delta G_{ij} = (|\Delta\Delta G_i + \Delta\Delta G_j| - |\Delta\Delta G_i - \Delta\Delta G_j|)$ are shown. If the value of this difference is larger than 0, it means the energy changes at the two positions of the *i,j* pairs point in the same direction (the two mutations are enhancing each other regardless of whether they are stabilizing or destabilizing); conversely if the value of this difference is smaller than 0, it means the energy changes at the two positions of the *i,j* pairs point in opposite direction (the two mutations are partially neutralizing each other). It is also possible to look at the average effect on stability of a pair change from its composition in *Nm.* KDO8PS to its composition in all the other sequences of the MSA combined. For example, in Figure 4C we see pairs whose covariation in a subset of all sequences would produce an average increase in stability, and in Figure 4D pairs whose covariation in a different subset of all the sequences would produce an average decrease in stability of the *Nm.* protein. Finally, we can further analyze a specific covarying pair for increased (or decreased) stability, and explore its composition throughout the alignment (this is a single 'pencil' in the 3D-matrix). For example, it turns out that pair 34,53 of Figure 4C has always an Ala at position 34 and either an Ile (-1 kcal/mol), Met (-0.8 kcal/mol), or Val (-0.7 kcal/mol) at position 53. The same pair of Figure 4D can instead have G,A,S,T at position 34, and A,C,I,M,S,T,V at position 53 with an average positive energy change of ~0.6 kcal/mol.

Visualizing coevolving positions in the representative structure of a protein family

It is usually very informative to locate the high scoring pairs directly in the reference structure. In Figure 5A we show the structure of *Nm.* KDO8PS as a tube whose thickness in different points is proportional to the level of covariation in that part of the structure. Coevolving pairs that are within 12 Å of each other are connected by dashed bonds: these pairs account for 49 out of the top 50 scoring pairs in the coevolution map calculated with 3D_MI. Pairs whose covariation is most likely to increase the stability of *Nm.* KDO8PS are shown separately in Figure 5B as dotted van der Waals surfaces.

DISCUSSION

Due to the strong statistical correlation between coevolving positions and positions that are close in space, coevolution analysis can be a powerful tool for the prediction of protein folds from sequence data only, and there are already many reports of coevolution maps derived from MSAs of both soluble and membrane proteins, that were of sufficient quality to generate accurate 3-dimensional structures [59-61]. However, very often structure prediction is not the goal and we are interested instead in structure-function relationships. In these cases coevolution analysis can be instrumental in identifying coevolving positions as pointers of residues involved in specific catalytic activities and/or in protein stability. For example, a network of positions consisting of both catalytic and non-catalytic residues was recently identified in homing endonucleases using computational methods to predict coevolving residues [68]. In these enzymes, variants of catalytic residues with low activity could be rescued by restoring an optimal coevolving network with compensatory mutations of the non-catalytic residues. In another example, Wang *et al.* [69] used coevolving sites saturation mutagenesis (CCSM) to enhance the thermal stability of *Bacillus subtilis* α-amylase by 8 °C, a result that

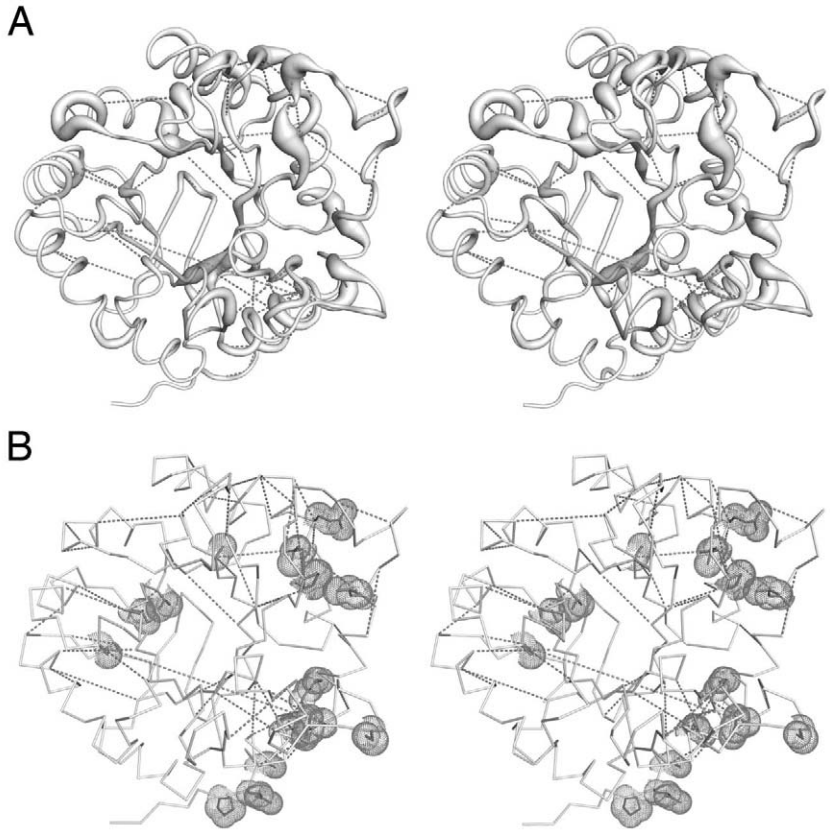

Figure 5. Location of coevolving pairs in the X-ray structure of *Nm.* KDO8PS. A. Stereo-view of a monomer of *Nm.* KDO8PS (PDB entry 2QKF): the structure is shown here as a tube whose thickness is proportional to the maximum covariation score of each column of the coevolution map (corresponding to a position in the structure) calculated with the 3D_MI method. Pairs of residues separated by less than 12 Å are connected by dashed bonds. **B.** Stereo-view of a monomer of *Nm.* KDO8PS in the same orientation as in panel A: pairs whose covariation to the residues present in other sequences of the MSA are predicted to increase the stability of *Nm.* KDO8PS are shown as thin sticks inside dotted van der Waals surfaces. Dashed bonds connect covarying pairs separated by less than 12 Å (as in panel A). This subset of covarying pairs is only partially overlapping the subset of pairs that increase stability.

could not be achieved by any ordinary rational introduction of single or double point mutations or by random mutagenesis.

Since, in addition to structure prediction, coevolution analysis can be very effective to design mutagenesis strategies, in this article we aimed to provide the reader with a well tested strategy to a) identify coevolving residues from the analysis of MSAs, and b) predict the effect on stability of different pairs of residues at specific coevolving positions. To this end, we have used our recent study of the contribution of coevolving residues to the stability of *Nm.* KDO8P synthase [50] as a practical example.

A key step in our study was the derivation of the stability landscape of *Nm.* KDO8PS (Figure 3) by calculating the folding ΔΔG changes associated with introducing any one of the 20 possible amino acids at each sequence position of the X-ray structure of this protein. Free-energy calculations were carried out with the FoldX algorithm, which uses a full atomic description of the structure of proteins and whose different energy terms have been weighted using empirical data obtained from protein engineering experiments [46, 55, 70]. Recent comparisons of various methods designed to predict the ΔΔG changes associated with mutations, ranked FoldX among the best performing ones [66, 67].

Calculation of the stability landscape (using FoldX or other comparable method) can be an important tool to integrate structural data with information

theory for the purpose of understanding how proteins increase their fitness in terms of both function and stability. We recall here that if an MSA is composed of independent sequences all producing approximately the same stable fold, and if individual residues contribute additively to stability [71, 72], then the stability contribution $\Delta\Delta G_{a,i}$ of a particular amino acid *a* at a given position *i* is roughly a logarithmic function of its frequency $f_{a,i}$ in the MSA [73]:

$$\Delta\Delta G_{a,i} \approx -\ln f_{a,i} \quad (4)$$

Approaches based on this idea have been generally successful in engineering more stable proteins [74-77]. However, what is lacking in these mutational strategies is a rationale for linking multiple mutations at different sites. Far too often massive mutagenesis studies based on the introduction of multiple point mutations at random or conserved sites, or on the generation of recombination libraries [72, 78] are carried out to obtain this information. For example, both rational and 'brute force' directed evolution studies were used extensively to map the evolutionary history of serine β-lactamases [79-86], so that useful predictions could be made with respect to the future appearance of particularly active/stable forms of these enzymes. With the merging of coevolution and folding free-energy calculations, we provide an avenue for the identification of multi-site mutations that increase (or decrease) stability. We submit that specific combinations of residues at coevolving positions represent unique signatures associated with particular levels of protein fitness, and that there is no need for massive projects of directed evolution to identify cooperating positions, because this information can be more efficiently extracted from the record of the past evolutionary history of proteins, which is hidden in their multiple sequence alignment.

CONCLUSION

The computational strategy outlined in this article provides a non brute-force approach to predict the future evolution of proteins based on their past history. The predicted outcome of this strategy as applied to a host of different proteins will be the identification of specific patterns of co-mutations, some of which have already appeared, while others may appear in future. This information can be collected in a database of *signature* sequences at known coevolving positions associated with specific fitness levels of each protein.

ACKNOWLEDGEMENTS

The author wishes to acknowledge earlier collaborations with Drs. Sharon Ackerman and Elisabeth Tillier that led to the design of some of the tools described in this review.

AUTHOR'S CONTRIBUTIONS

The author wrote all the Matlab scripts and the manuscript.

CONFLICT OF INTEREST STATEMENT

The author declares no competing interests. This research was supported by United States Public Health Service grant GM69840 and by a Wayne State University Research Enhancement Program in Computational Biology grant to the author.

SUPPLEMENTARY MATERIAL

Supplementary Material to this article, in the form of Matlab, FoldX, and Pymol scripts that can be used as general templates for the generation of high quality MSAs, the analysis of positional coevolution, and the analysis of folding energy matrices with any protein family, can be requested directly from the author.

REFERENCES

1. Wollenberg, K. R. and Atchley, W. R. 2000, Proc. Natl. Acad. Sci. USA, 97(7), 3288-3291.
2. Atchley, W. R., Wollenberg, K. R., Fitch, W. M., Terhalle, W. and Dress, A. W. 2000, Mol. Biol. Evol., 17(1), 164-178.
3. Fares, M. A. and Travers, S. A. 2006, Genetics, 173(1), 9-23.
4. Dickson, R. J. and Gloor, G. B. 2012, PLoS One, 7(6), e37645.
5. Dickson, R. J. and Gloor, G. B. 2014, Methods in Molecular Biology, 1123, 223-243.
6. Horner, D. S., Pirovano, W. and Pesole, G. 2008, Brief Bioinform, 9(1), 46-56.
7. Caporaso, J. G., Smit, S., Easton, B., Hunter, L., Huttley, G. and Knight, R. 2008, BMC Evolutionary Biology, 8(1), 327.
8. Codoner, F. M. and Fares, M. A. 2008, Evol. Bioinform Online, 4, 29-38.

9. Ackerman, S. H., Tillier, E. R. and Gatti, D. L. 2012, PLoS One, 7(10), e47108.
10. de Juan, D., Pazos, F. and Valencia, A. 2013, Nature Reviews Genetics, 14(4), 249-261.
11. Marks, D. S., Hopf, T. A. and Sander, C. 2012, Nature Biotechnology, 30(11), 1072-1080.
12. Kass, I. and Horovitz, A. 2002, Proteins, 48(4), 611-617.
13. Fodor, A. A. and Aldrich, R. W. 2004, Proteins, 56(2), 211-221.
14. Dekker, J. P., Fodor, A., Aldrich, R. W. and Yellen, G. 2004, Bioinformatics, 20(10), 1565-1572.
15. Lockless, S. W. and Ranganathan, R. 1999, Science, 286(5438), 295-299.
16. Halabi, N., Rivoire, O., Leibler, S. and Ranganathan, R. 2009, Cell, 138(4), 774-786.
17. Gobel, U., Sander, C., Schneider, R. and Valencia, A. 1994, Proteins, 18(4), 309-317.
18. Reza, F. M. 1994, An Introduction to Information Theory, Dover Publications, Inc., New York.
19. Tillier, E. R. and Lui, T. W. 2003, Bioinformatics, 19(6), 750-755.
20. Martin, L. C., Gloor, G. B., Dunn, S. D. and Wahl, L. M. 2005, Bioinformatics, 21(22), 4116-4124.
21. Gloor, G. B., Martin, L. C., Wahl, L. M. and Dunn, S. D. 2005, Biochemistry, 44(19), 7156-7165.
22. Dunn, S. D., Wahl, L. M. and Gloor, G. B. 2008, Bioinformatics, 24(3), 333-340.
23. Brown, C. A. and Brown, K. S. 2010, PLoS One, 5(6), e10779.
24. Little, D. Y. and Chen, L. 2009, PLoS One, 4(3), e4762.
25. Gloor, G. B., Tyagi, G., Abrassart, D. M., Kingston, A. J., Fernandes, A. D., Dunn, S. D. and Brandl, C. J. 2010, Mol. Biol. Evol., 27(5), 1181-1191.
26. Burger, L. and van Nimwegen, E. 2010, PLoS Comput. Biol., 6(1), e1000633.
27. Jones, D. T. Buchan, D. W., Cozzetto, D. and Pontil, M. 2012, Bioinformatics, 28(2), 184-190.
28. Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D. S., Sander, C., Zecchina, R., Onuchic, J. N., Hwa, T. and Weigt, M. 2011, Proc. Natl. Acad. Sci. USA, 108(49), E1293-1301.
29. Ekeberg, M., Lovkvist, C., Lan, Y., Weigt, M. and Aurell, E. 2013, Physical Review E, Statistical, Nonlinear, and Soft Matter Physics, 87(1), 012707.
30. Feinauer, C., Skwark, M. J., Pagnani, A. and Aurell, E. 2014, PLoS Comput. Biol., 10(10), e1003847.
31. Kamisetty, H., Ovchinnikov, S. and Baker, D. 2013, Proc. Natl. Acad. Sci. USA, 110(39), 15674-9.
32. Cocco, S., Monasson, R. and Weigt, M. 2013, PLoS Comput. Biol., 9(8), e1003176.
33. Shannon, C. E. 1948, Bell System, Technical Journal, 27, 379-423.
34. Shannon, C. E. 1948, Bell System, Technical Journal, 27, 623-656.
35. Suel, G. M., Lockless, S. W., Wall, M. A. and Ranganathan, R. 2003, Nat. Struct. Biol., 10(1), 59-69.
36. Fernandes, A. D. and Gloor, G. B. 2010, Bioinformatics, 26(9), 1135-1139.
37. Weigt, M., White, R. A., Szurmant, H., Hoch, J. A. and Hwa, T. 2009, Proc. Natl. Acad. Sci. USA, 106(1), 67-72.
38. Clark, G. W., Ackerman, S. H., Tillier, E. R. and Gatti, D. L. 2014, BMC Bioinformatics, 15(1), 157.
39. Warshel, A. and Parson, W. W. 2001, Q. Rev. Biophys., 34(4), 563-679.
40. Roca, M., Liu, H., Messer, B. and Warshel, A. 2007, Biochemistry, 46(51), 15076-15088.
41. Nagatani, R. A., Gonzalez, A., Shoichet, B. K., Brinen, L. S. and Babbitt, P. C. 2007, Biochemistry, 46(23), 6688-6695.
42. Beadle, B. M. and Shoichet, B. K. 2002, J. Mol. Biol., 321(2), 285-296.
43. Wang, X., Minasov, G. and Shoichet, B. K. 2002, J. Mol. Biol., 320(1), 85-95.
44. Bloom, J. D., Labthavikul, S. T., Otey, C. R. and Arnold, F. H. 2006, Proc. Natl. Acad. Sci. USA, 103(15), 5869-5874.
45. Matthews, B. W. 1993, Annu. Rev. Biochem., 62, 139-160.

46. Guerois, R., Nielsen, J. E. and Serrano, L. 2002, J. Mol. Biol., 320(2), 369-387.
47. Schymkowitz, J. W., Rousseau, F., Martins, I. C., Ferkinghoff-Borg, J., Stricher, F. and Serrano, L. 2005, Proc. Natl. Acad. Sci. USA, 102(29), 10147-10152.
48. Tokuriki, N., Stricher, F., Serrano, L. and Tawfik, D. S. 2008, PLoS Comput. Biol., 4(2), 1-7.
49. Raetz, C. R. and Whitfield, C. 2002, Annu. Rev. Biochem., 71, 635-700.
50. Ackerman, S. H. and Gatti, D. L. 2011, PLoS One, 6(3), e17459.
51. Edgar, R. C. 2004, Nucl. Acids Res., 32(5), 1792-1797.
52. Katoh, K., Misawa, K., Kuma, K-i. and Miyata, T. 2002, Nucl. Acids Res., 30(14), 3059-3066.
53. Notredame, C., Higgins, D. G. and Heringa, J. 2000, J. Mol. Biol., 302(1), 205-217.
54. Bowers, K. J., Chow, E., Xu, H., Dror, R. O., Eastwood, M. P., Gregerson, B. A., Klepeis, J. L., Kolossvary, I., Moraes, M. A., Sacerdoti, F. D., Salmon, J. K., Shan, Y. and Shaw, D. E. 2006, Scalable Algorithms for Molecular Dynamics Simulations on Commodity Clusters. In: ACM/IEEE Conference on Supercomputing (SC06): November 11-17; Tampa, Florida, Los Alamitos, CA, USA, IEEE Computer Society, 43.
55. Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F. and Serrano, L. 2005, Nucleic Acids Res., 33(Web Server issue), W382-388.
56. Kiel, C. and Serrano, L. 2006, J. Mol. Biol., 355(4), 821-844.
57. Consortium, T. U. 2014, Nucleic Acid Res., 42(D1), D191-D198.
58. Cochrane, F. C., Cookson, T. V., Jameson, G. B. and Parker, E. J. 2009, J. Mol. Biol., 390(4), 646-661.
59. Marks, D. S., Colwell, L. J., Sheridan, R., Hopf, T. A., Pagnani, A., Zecchina, R. and Sander, C. 2011, PLoS One, 6(12), e28766.
60. Hopf, T. A., Colwell, L. J., Sheridan, R., Rost, B., Sander, C. and Marks, D. S. 2012, Cell, 149(7), 1607-1621.
61. Nugent, T. and Jones, D. T. 2012, Proc. Natl. Acad. Sci. USA, 109(24), E1540-E1547.
62. Dickson, R. J., Wahl, L. M., Fernandes, A. D. and Gloor, G. B. 2010, PLoS One, 5(6), e11082.
63. Olmea, O., Rost, B. and Valencia, A. 1999, J. Mol. Biol., 293(5), 1221-1239.
64. Russ, W. P., Lowery, D. M., Mishra, P., Yaffe, M. B. and Ranganathan, R. 2005, Nature, 437(7058), 579-583.
65. Tokuriki, N., Stricher, F., Schymkowitz, J., Serrano, L. and Tawfik, D. S. 2007, J. Mol. Biol., 369(5), 1318-1332.
66. Potapov, V., Cohen, M. and Schreiber, G. 2009, Protein Eng. Des. Sel., 22(9), 553-560.
67. Khan, S. and Vihinen, M. 2010, Hum. Mutat., 31(6), 675-684.
68. McMurrough, T. A., Dickson, R. J., Thibert, S. M., Gloor, G. B. and Edgell, D. R. 2014, Proc. Natl. Acad. Sci. USA, 111(23), E2376-2383.
69. Wang, C., Huang, R., He, B. and Du, Q. 2012, BMC Bioinformatics, 13, 263.
70. Mendes, J., Guerois, R. and Serrano, L. 2002, Curr. Opin. Struct. Biol., 12(4), 441-446.
71. Horovitz, A. 1996, Fold Des., 1(6), R121-126.
72. Tracewell, C. A. and Arnold, F. H. 2009, Curr. Opin. Chem. Biol., 13(1), 3-9.
73. Ohage, E. C., Graml, W., Walter, M. M., Steinbacher, S. and Steipe, B. 1997, Protein Sci., 6(1), 233-241.
74. Polizzi, K. M., Bommarius, A. S., Broering, J. M. and Chaparro-Riggers, J. F. 2007, Curr. Opin. Chem. Biol., 11(2), 220-225.
75. Chaparro-Riggers, J. F., Polizzi, K. M. and Bommarius, A. S. 2007, Biotechnol. J., 2(2), 180-191.
76. Lehmann, M., Pasamontes, L., Lassen, S. F. and Wyss, M. 2000, Biochim. Biophys. Acta, 1543(2), 408-415.
77. Lehmann, M. and Wyss, M. 2001, Curr. Opin. Biotechnol., 12(4), 371-375.
78. Romero, P. A. and Arnold, F. H. 2009, Nature Reviews Molecular Cell Biology, 10(12), 866-876.
79. DePristo, M. A., Hartl, D. L. and Weinreich, D. M. 2007, Mol. Biol. Evol., 24(8), 1608-1610.

80. Meyer, M. M., Hiraga, K. and Arnold, F. H. 2006, Curr. Protoc. Protein Sci., Chapter 26, Unit 26 22.
81. Kather, I., Jakob, R. P., Dobbek, H. and Schmid, F. X. 2008, J. Mol. Biol., 383(1), 238-251.
82. Barlow, M. and Hall, B. G. 2002, Genetics, 161(3), 1355B-1355.
83. Bershtein, S., Goldin, K. and Tawfik, D. S. 2008, J. Mol. Biol., 379(5), 1029-1044.
84. Bershtein, S. and Tawfik, D. S. 2008, Mol. Biol. Evol., 25(11), 2311-2318.
85. Bershtein, S. and Tawfik, D. S. 2008, Curr. Opin. Chem. Biol., 12(2), 151-158.
86. Mroczkowska, J. E. and Barlow, M. 2008, Antimicrob. Agents Chemother., 52(7), 2340-2345.