Understanding conserved amino acids in proteins

Nikolay V. Dokholyan[a,*], Leonid A. Mirny[a,b],
Eugene I. Shakhnovich[a]

[a] Department of Chemistry and Chemical Biology, Harvard University, 12 Oxford Street, Cambridge,
MA 02138, USA
[b] Harvard-MIT Division of Health Sciences and Technology, Massachusetts Institute of Technology,
77 Massachusetts Ave, 16-343, Cambridge, MA 02139, USA

Abstract

It has been conjectured that evolution exerted pressure to preserve amino acids bearing thermodynamic, kinetic, and functional roles. In this letter we show that the physical requirement to maintain protein stability gives rise to a sequence conservatism pattern that is in remarkable agreement with that found in natural proteins. Based on the physical properties of amino acids, we propose a model of evolution that explains conserved amino acids across protein families sharing the same fold.
© 2002 Elsevier Science B.V. All rights reserved.

PACS: 87.15.−v; 87.90.+y

Keywords: Protein evolution; Conserved amino acids; Model

Molecular evolution is a sumptuous natural laboratory that provides an invaluable source of information about the structure, dynamics and function(s) of biomolecules. This information has already been widely used to understand the folding kinetics, thermodynamics, and function of proteins (e.g. [1–3]). The basic belief behind the majority of such studies is that evolution optimizes, to a certain extent, the properties of proteins, so that they become more sufficiently stable, and have better folding and functional properties.

---

* Corresponding author. Tel.: +1-617-496-5499; fax: +1-617-495-3075.
E-mail address: dokh@wild.harvard.edu (N.V. Dokholyan).

Recent studies [2,4] identified positions in several common protein folds where amino acids are universally conserved within each family of proteins having that fold. Such positions are localized in structure, and their unusually strong conservatism may be due to functional reason (e.g. super-site), or folding kinetics (folding nucleus) [5–8]. In contrast to function and folding kinetics, evolutionary pressure to maintain stability may be applied "more evenly" because all amino acids contribute, to a lesser or greater extent, to protein stability via their interaction with other amino acid residues and with water.

In this letter we develop a model that provides a rationale for conservatism patterns caused by selection for stability. Our model is of *equilibrium* evolution that maintains stability and other properties achieved at an earlier, prebiotic stage. To this end we propose that stability selection accepts only those mutations that keep energy of the native protein, E, below a certain threshold E_0 necessary to maintain an energy gap [9–12]. The requirement to maintain an energy threshold for the viable sequences makes the equilibrium ensemble of sequences analogous to a microcanonical ensemble. In analogy with statistical mechanics, a more convenient and realistic description of the sequence ensemble is a canonical ensemble, whereby strict requirements on energy of the native state is replaced by a "soft" evolutionary pressure that allows energy fluctuations from sequence to sequence but makes sequences with high energy in the native state unlikely. In the canonical ensemble of sequences, the probability of finding a particular sequence, $\{\sigma\}$, in the ensemble follows the Boltzmann distribution [10–13]

$$P(\{\sigma\}) = \frac{\exp(-E\{\sigma\}/T)}{Z} \,, \tag{1}$$

where T is the effective temperature of the canonical ensemble of sequences that serves as a measure of evolutionary pressure and $Z = \sum_{\{\sigma\}} \exp(-E\{\sigma\}/T)$ is the partition function taken in sequence space.

Next, we apply a mean-field approximation that replaces all multiparticle interactions between amino acids by interaction of each amino acid with an effective field Φ acting on this amino acid from the rest of the protein. This approximation presents $P(\{\sigma\})$ in a multiplicative form as $\prod_{k=1}^{N} p(\sigma_k)$ of probabilities to find an amino acid σ at position k. $p(\sigma_k)$ also obeys Boltzmann statistics

$$p(\sigma_k) = \frac{\exp(-\Phi(\sigma_k)/T)}{\sum_{\sigma} \exp(-\Phi(\sigma_k)/T)} \,. \tag{2}$$

The mean field potential $\Phi(\sigma_k)$ is the effective potential energy between amino acid σ_k and all amino acids interacting with it, i.e.,

$$\Phi(\sigma_k) = \sum_{i \neq k}^{N} U(\sigma_k, \sigma_i) \Delta_{ki} \,. \tag{3}$$

Matrix U describes energy parameters in contact approximation and matrix Δ is a contact matrix for protein native structure (see footnotes 1 and 2 for more

details).[1,2] The potential Φ is similar in spirit to the protein profile introduced by Bowie et al. [14] to identify protein sequences that fold into a specific 3D structure.

For each member, m, of the fold family (FSSP database [15]) we compute the mean-field probability, $p_m(\sigma_k)$, using Eq. (2) and then, we compute the average mean-field probability over all N_s families sharing the same fold,

$$p_{MF}(\sigma_k) = \frac{1}{N_s} \sum_{m=1}^{N_s} p_m(\sigma_k) \,. \tag{4}$$

Eqs. (2)–(4), along with properly selected energy function, U, make it possible to predict probabilities of all amino acid types and sequence entropy $S_{MF}(k)$ at each position k

$$S_{MF}(k) = -\sum_{\sigma} p_{MF}(\sigma_k) \ln p_{MF}(\sigma_k) \tag{5}$$

from the native structure of a protein. The summation is taken over all possible values of σ.

Theoretical predictions from statistical-mechanical analysis can be compared with data on real proteins. In order to determine conservatism in real proteins we note that the space of sequences that fold into the same protein structure presents a two-tier system, where homologous sequences are grouped into families and there is no recognizable sequence homology between families despite the fact that they fold into closely related structures [2,16,17].

Using the database of protein families with close sequence similarity (HSSP database [18]), we compute frequencies of amino acids at each position, k, of aligned sequences, $P_m(\sigma_k)$, for a given, mth, family of proteins. We average these frequencies *across* all N_s families sharing the same fold that are present in FSSP database [15]:

$$P_{acr}(\sigma_k) = \frac{1}{N_s} \sum_{m=1}^{N_s} P_m(\sigma_k) \,. \tag{6}$$

---

[1] We represent interactions in a protein in a C_β approximation—each pair of amino acids interact if their C_βs (C_α in the case of Gly) are within the contact distance 7.5 Å [23]. The total potential energy of the protein can be written as follows: $E = \frac{1}{2} \sum_{i \neq j}^{N} U(\sigma_i, \sigma_j) \Delta_{ij}$, where N is the length of the protein, and σ_i is the amino acid type at the position $i = 1, \ldots, N$. $U(\sigma_i, \sigma_j)$ is the corresponding element of the matrix of pairwise interactions between amino acids σ_i and σ_j. Δ_{ij} is the element of the contact matrix that is defined to be 1 if there is a contact between amino acids i and j, and is 0 otherwise. The gaps are treated as "ghost" amino acids that do not contribute to the total potential energy.

[2] Due to the similarities in the properties of the 20 types of amino acids one can classify these amino acids into 6 distinct groups: aliphatic $\{AVLIMC\}$, aromatic $\{FWYH\}$, polar $\{STNQ\}$, positive $\{KR\}$, negative $\{DE\}$, and special (reflecting their special conformational properties) $\{GP\}$. We construct the effective potential of interaction, $U_6(\hat{\sigma}_i, \hat{\sigma}_j)$, between six groups of amino acids, $\hat{\sigma}$, by computing the average interaction between these groups, i.e., $U_6(\hat{\sigma}_i, \hat{\sigma}_j) = 1/(N_{\hat{\sigma}_i} N_{\hat{\sigma}_j}) \sum_{\sigma_k \in \hat{\sigma}_i, \sigma_l \in \hat{\sigma}_j} U_{20}(\sigma_k, \sigma_l)$, where σ denotes amino acids in the 20-letter representation, and $U_{20}(\sigma_k, \sigma_l)$ is the 20-letter Miyazawa–Jernigan matrix of interaction [24]; $\hat{\sigma}$ denotes amino acids in the 6-letter representation. $N_{\hat{\sigma}}$ is the number of actual amino acids of type $\hat{\sigma}$, e.g. for the aliphatic group, $N_{\hat{\sigma}} = 6$.

Table 1
The values of the correlation coefficient r for the linear regression of $S_{MF}(k)$ versus S_{acr} for Ig, OB, and R, folds and the corresponding optimal values of the temperature $T = T_{sel}$. The last column corresponds to the correlation coefficient for the studied folds at a fixed selective temperature $T = 0.25$

Fold	N_s	Representative protein		Correlation coefficient, $r(T)$		
		PDB code [25,26]	N	$r(T_{sel})$	T_{sel}	$r(T = 0.25)$
Ig	51	1TEN	89	0.63	0.34	0.57
OB	18	1MJC	69	0.69	0.19	0.69
R	166	3CHY	128	0.71	0.25	0.71

Next, we determine the sequence entropy, $S_{acr}(k)$, at each position, k, of structurally aligned protein analogs:

$$S_{acr}(k) = -\sum_{\sigma} P_{acr}(\sigma_k) \ln P_{acr}(\sigma_k) \, . \tag{7}$$

If the stability selection were a factor in evolution of proteins and our model captures it, then we should observe correlation between predicted mean field based sequence entropies, $S_{MF}(k)$, and actual sequence entropies $S_{acr}(k)$ in real proteins. Thus, the question is: "Can we find such T, so that predicted conservatism profile $S_{MF}(k)$ matches the real one $S_{acr}(k)$?"

By varying the values of the temperature, T in the range $0.1 \leqslant T \leqslant 4.0$, we minimize the distance, $D^2 \equiv \sum_{k=1}^{N}(S_{MF}(k) - S_{acr}(k))^2$, between the predicted and observed conservatism profiles. We exclude from this sum such positions in structurally aligned sequences that have more than 50% gaps in the structural (FSSP) alignment. We denote by T_{sel} the temperature that minimizes D.

We study three folds: Immunoglobulin fold (Ig), Oligonucleotide-binding fold (OB), and Rossman fold (R). We compute correlation coefficient [19] between values of $S_{MF}(k)$, obtained at T_{sel}, and $S_{acr}(k)$ for all three folds. The results are summarized in Table 1. The plots of $S_{MF}(k)$ and $S_{acr}(k)$ versus k as well as their scatter plots are shown in Figs. 1–3(a,b).

The correlation between $S_{MF}(k)$ and $S_{acr}(k)$ is remarkable for all three folds and indicates that our mean field model is able to select the conserved amino acids in protein fold families. It is fully expected that the correlation coefficient is smaller than 1. The reason for this is that computation of $S_{MF}(k)$ takes into account evolutionary selection for stability only and it does not take into account possible additional pressure to optimize kinetic or functional properties. Interestingly, the correlation between $S_{MF}(k)$ and participation in the secondary structure is poor (results are not shown), indicating the special role of energetic factors in evolutionary conservatism.

The additional evolutionary pressure due to kinetic or functional importance of amino acids results in pronounced deviations of S_{MF} from S_{acr} for few amino acids that may be kinetically or functionally important. A number of amino acids whose conservatism is much greater than predicted by our model form a group of "outliers", from otherwise
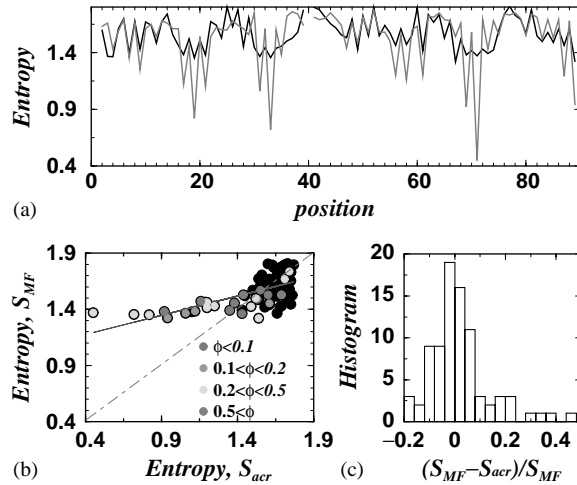
(a)

(b)   (c)

Fig. 1. (a) The values $S_{MF}(k)$ (black line) and $S_{acr}(k)$ (red line) for all positions, k, for the Ig-fold. The lower the values of $S_{MF}(k)$ the more conservative amino acids are at these positions. (b) The scatter plot of predicted $S_{MF}(k)$ versus observed $S_{acr}(k)$. The linear regression correlation coefficients are shown in Table 1. Blue line is the linear regression has the slope different than 1 (red line), corresponding to the $S_{MF}(k) = S_{acr}(k)$ relation. (c) The histogram of the differences between $S_{MF}(k)$ and $S_{acr}(k)$. In (b) we assign colors to data points corresponding to amino acids with the specific range of ϕ-values [21]: red, if $0.5 < \phi < 1$, yellow, if $0.2 < \phi < 0.5$, magenta, if $0.1 < \phi < 0.2$, violet if $\phi < 0.1$, and black if ϕ-values are not determined.
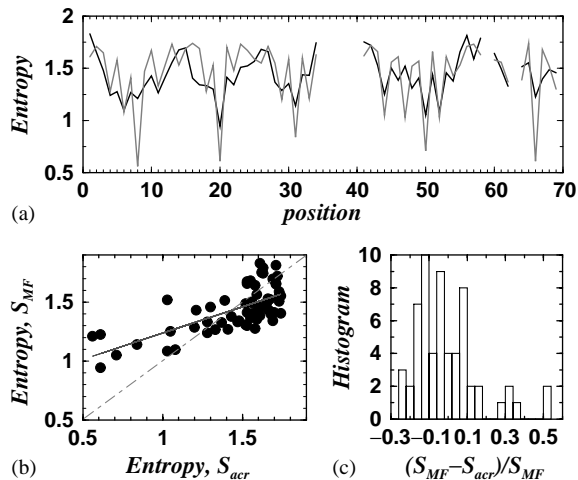

(a)

(b)   (c)

Fig. 2. (a)–(c) The same as in Fig. 1 but for the OB-fold.

very close correspondence between S_{MF} and S_{acr}. To demonstrate that some of those amino acids are important for folding kinetics and as such they can be under additional evolutionary pressure, we color data points on S_{MF} versus S_{acr} scatter plot according to
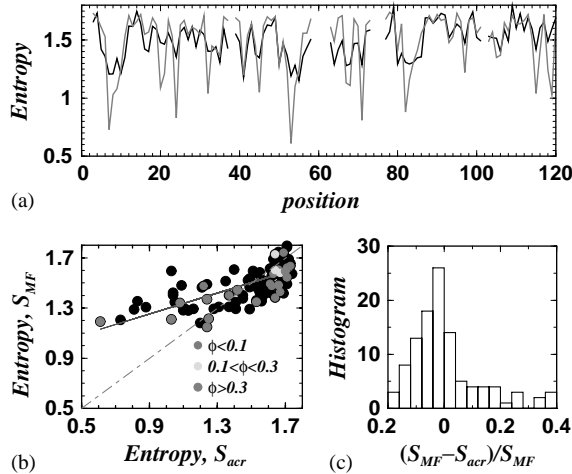
Fig. 3. (a)–(c) The same as in Fig. 1 but for R-fold. In (b) we assign colors to data points corresponding to amino acids with the specific range of ϕ-values [22]: red, if $0.3 < \phi < 1$, yellow, if $0.1 < \phi < 0.3$, violet if $\phi < 0.1$, and black if ϕ-values are not determined.

the range of ϕ-values [20] that corresponding amino acids fall into. Thermodynamics and kinetic role of individual amino acids was studied extensively (i) by Hamill et al. [21] for the TNfn3 (1TEN) protein, and (ii) by López-Hernández and Serrano [22] for the CheY protein. We use the ϕ-values for individual amino acids obtained in [21,22]. We observe that (i) for TNfn3 protein most of the points on Fig. 1(b) that belong to the outlier group have ϕ-values ranging from 0.2 to 1, and (ii) for CheY protein most of the points (for which ϕ-values are known) on Fig. 3(b) that belong to the outlier have ϕ-values ranging from 0.3 to 1. The number of high ϕ-value residues in the "outliers" is statistically significantly greater than expected if those were distributed by chance (calculations not shown).

The proposed model has a dual role. On one side, it allows to understand the selective temperature scale, T_{sel}, which is the measure of evolutionary optimization. On the other side, the correlation coefficient between $S_{MF}(k)$ and $S_{acr}(k)$ does not vary strongly in the range of T_{sel} from 0.19 to 0.34, thus, allowing one to use the effective temperature of $T_{sel} = 0.25$ to predict actual conservatism profiles of proteins (see Table 1).

To conclude, we present a theory that explains sequence conservation caused by the most basic and universal evolutionary pressure in proteins to maintain stability. The theory predicts very well sequence entropy for the majority of amino acids, but not all of them. The amino acids that exhibit considerably higher conservatism than predicted from stability pressure alone are likely to be important for function and/or folding. Comparison of the "base-level" stability conservatism $S_{MF}(k)$ with $S_{acr}(k)$—actual conservatism profile of a protein fold—allows to identify functionally and kinetically important amino acid residues and potentially gain specific insights into folding and function of a protein.

References

[1] C. Branden, J. Tooze, Introduction to Protein Structure, Garland Publishing Inc., New York, 1998.
[2] L.A. Mirny, E.I. Shakhnovich, J. Mol. Biol. 291 (1999) 177.
[3] N.V. Dokholyan, E.I. Shakhnovich, J. Mol. Biol. 312 (2001) 289.
[4] O.B. Ptitsyn, J. Mol. Biol. 278 (1998) 655.
[5] L.A. Mirny, V.I. Abkevich, E.I. Shakhnovich, Proc. Natl. Acad. Sci. USA 95 (1998) 4976.
[6] N.A.J. van Nuland, H. Chiti, N. Taddei, G. Raugei, G. Ramponi, C. Dobson, J. Mol. Biol. 283 (1998) 883.
[7] N.A.J. van Nuland, W. Meijberg, J. Warner, V. Forge, R. Scheek, G. Robbilard, C. Dobson, Biochemistry 37 (1998) 622.
[8] N.V. Dokholyan, S.V. Buldyrev, H.E. Stanley, E.I. Shakhnovich, J. Mol. Biol. 296 (2000) 1183.
[9] A. Sali, E.I. Shakhnovich, M. Karplus, J. Mol. Biol. 235 (1994) 1614.
[10] E.I. Shakhnovich, A.M. Gutin, Proc. Natl. Acad. Sci. USA 90 (1993) 7195.
[11] S. Ramanathan, E.I. Shakhnovich, Phys. Rev. E 50 (1994) 1303.
[12] A.V. Finkelstein, A. Gutin, A. Badretdinov, Proteins 23 (1995) 142.
[13] V.S. Pande, A.Y. Grosberg, T. Tanaka, Phys. Rev. E 51 (1995) 3381.
[14] J.U. Bowie, R. Luthy, D. Eisenberg, Science 253 (1991) 164.
[15] L. Holm, C. Sander, J. Mol. Biol. 233 (1993) 123.
[16] B. Rost, Folding Des. 2 (1997) S19.
[17] G. Tiana, R. Broglia, E.I. Shakhnovich, Proteins: Struct. Funct. Genet. 39 (2000) 244.
[18] C. Dodge, R. Schneider, C. Sander, Nucl. Acid Res. 26 (1998) 313.
[19] W.H. Press, B.P. Flannery, S.A. Teukolsky, W.T. Vetterling, Numerical Recipes, Cambridge University Press, Cambridge, 1989.
[20] L.S. Itzhaki, D.E. Otzen, A.R. Fersht, J. Mol. Biol. 254 (1995) 260.
[21] S.J. Hamill, A. Steward, J. Clarke, J. Mol. Biol. 297 (2000) 165.
[22] E. López-Hernández, L. Serrano, Folding Des. 1 (1996) 43.
[23] R.L. Jernigan, I. Bahar, Curr. Opinion Struct. Biol. 6 (1996) 195.
[24] S. Miyazawa, R.L. Jernigan, Proteins 34 (1999) 49 (Table V.).
[25] F.C. Bernstein, T.F. Koetzle, G.J.B. Williams, E.F.M. Jr., M.D. Brice, J.R. Rodgers, O. Kennard, T. Shimanouchi, M. Tasumi, J. Mol. Biol. 112 (1977) 535.
[26] E.E. Abola, F.C. Bernstein, S.H. Bryant, T.F. Koetzle, J. Weng, in: F.H. Allen, G. Bergerhoff, R. Sievers (Eds.), Crystallographic Databases-Information Content, Software Systems, Scientific Applications, Data Commission of the International Union of Crystallography, Cambridge, 1987, pp. 107–132.