Structural Constraints and Emergence of Sequence Patterns in Protein Evolution

Gustavo Parisi and Julián Echave
Universidad Nacional de Quilmes, Bernal, Argentina

The aim of this work was to study the relationship between structure conservation and sequence divergence in protein evolution. To this end, we developed a model of structurally constrained protein evolution (SCPE) in which trial sequences, generated by random mutations at gene level, are selected against departure from a reference three-dimensional structure. Since at the mutational level SCPE is completely unbiased, any emergent sequence pattern will be due exclusively to structural constraints. In this first report, it is shown that SCPE correctly predicts the characteristic hexapeptide motif of the left-handed parallel β helix (LβH) domain of UDP-N-acetylgalactosamine acetylttransferases (LpxA).

Introduction

Protein sequences diverge due to amino acid replacements with a mostly neutral effect on organism fitness (Kimura 1983; Perutz 1983). To be neutral, sequence variation should have little effect on the protein's function, resulting in small variations in, for example, active sites (Mirny and Shakhnovich 1999), folding transition states (Shakhnovich, Abkevich, and Pitsyn 1996; Mirny and Shakhnovich 1999; Li, Mirny, and Shakhnovich 2000), and native state structures (Bajaj and Blundell 1984; Chothia and Lesk 1986; Flores et al. 1993; Wood and Pearson 1999). Structure conservation would explain why substitution patterns depend on factors such as amino acid physicochemical properties (Xia and Li 1998), local structural environment (Overington et al. 1990), and overall environmental constraints (Tourasse and Li 2000). It is known that considering structural effects on protein evolution can improve phylogenetic inference (Naylor and Brown 1997). The latest models of protein evolution take this into account by using structure-dependent amino acid replacement rates (Koshi and Goldstein 1998; Liò and Goldman 1998). Despite the proven usefulness of such models, to gain further insight into protein evolution, new models should be developed in which structure-dependent substitution patterns are not set in advance, but result naturally as a consequence of restraining structure variation. In this report, a structurally constrained protein evolution (SCPE) model is presented. The model is related to methods used to study the inverse protein folding problem (Babajide et al. 1997; Koehl and Levitt 1999a, 1999b).

Materials and Methods

The Model

In SCPE, trial sequences, generated by random mutations at the gene level, are selected against departure from a reference structure. The model is based on a sequence-structure distance score, \( S_{\text{dist}} \), which depends on a reference native structure, and a parameter, \( S_{\text{div}} \), which measures the degree of structural divergence tolerated by natural selection. The sequence-structure distance measure \( S_{\text{dist}} \) is calculated as follows. First, the trial sequence is forced to adopt the three-dimensional reference structure. Then, mean field energies per position \( E_{\text{trial}}(p) \) and \( E_{\text{ref}}(p) \) are calculated for the trial and reference sequences, respectively. Finally, \( S_{\text{dist}} = \left( \sum_p |E_{\text{trial}}(p) - E_{\text{ref}}(p)|^2 \right)^{1/2} \) is obtained. To calculate the mean-field energies, we used the PROSA II potential (Sippl 1993), which includes additive pair contributions that depend on the amino acid types and the geometric distance between the Cα atoms of the interacting amino acids, as well as a surface term that models the protein-solvent interactions.

An SCPE simulation starts with a reference DNA sequence that codes for a reference protein of known three-dimensional structure. Then, each run involves the repetition of evolutionary time steps, which consist of the application of the following four operations. First, the DNA sequence of the previous time step is mutated by introducing a random nucleotide substitution into a randomly chosen sequence position (Jukes-Cantor model). Second, if the mutation introduces a stop codon, the mutated DNA is rejected; otherwise, the mutated DNA is translated, using the genetic code, to obtain a trial protein sequence. Third, the sequence-structure distance score, \( S_{\text{dist}} \), is computed. Finally, the trial sequence is accepted only if \( S_{\text{dist}} \) is below the specified cut-off, \( S_{\text{div}} \), which represents the degree of structural divergence allowed by natural selection.

Test System

The SCPE model was tested on the left-handed parallel β helix (LβH) domain of UDP-N-acetylgalactosamine acetylttransferases (LpxA), which displays a distinctive sequence pattern that is likely to result from structural constraints. The reference for SCPE simulations was the LpxA of *Escherichia coli* (Raetz and Roderick 1995) (fig. 1A and B; PDB code 1lxq). The sequences of the LβH domain of members of the LpxA family (fig. 1C) consist of the imperfect tandem repetition of hexapeptide units (Vaara 1992; Vuorio et al. 1994; Raetz...
Fig. 1.—LpxA of Escherichia coli and LpxA family (Vaara 1992; Vuorio et al. 1994; Raetz and Roderick 1995). A, Cartoon view of the LβH domain of the LpxA of E. coli. PDB entry 1lxa. The left-handed β helix is formed by nine triangular coils (C1–C9). Each coil is formed by three hexapeptides, colored red, yellow, and blue, respectively. Loops are colored gray. B, Detailed view of coil C2 of A. Amino acids at conserved hexapeptide positions 1 (A18, A24, and C30) and 3 (I20, I26, and V32) are labeled. Panels A and B were prepared with the program MOLMOL (Koradi, Billeter, and Wuthrich 1996). C, Multiple-sequence alignment of the LβH domain of the members of the LpxA family. The alignment was obtained using CLUSTAL W (Thompson, Higgins, and Gibson 1994). Sequences are identified using the SwissProt/TrEMBL codes (Bairoch and Apweiler 2000). Conserved substitutions are shaded in black if the whole column is conserved, and they are shaded in gray if >75% is conserved. The following classes were used to judge conservation: aliphatic (ACILMV), aromatic (FHWY), polar (NQST), charged positive (KR), charged negative (DE), and special (GP). Colors of the first line of C relate the alignment to the structure shown in A and B. In addition, different coils (C1–C9) and hexapeptide third positions (dots) are explicitly indicated.
The hexapeptides are characterized by a high degree of conservation of the third position, which usually displays I, L, or V (a one-letter code is used to designate amino acids). Hexapeptide position 1 is also significantly conserved, although less so than position 3, whereas the other four hexapeptide sites (2, 4, 5, and 6) are not conserved. Figure 1B shows that the residues of conserved sites 1 and 3 point toward the inside of the beta helix, whereas those in variable positions point toward the outside. The LpxA family belongs to a larger superfamily of LβH acyltransferases. All members of this superfamily present the hexapeptide sequence motif, and those members whose structures have been determined display the LβH fold (Raetz and Roderick 1995; Kisker et al. 1996; Beaman et al. 1997; Beaman, Sugantino, and Roderick 1998; Brown et al. 1999). Thus, both the LβH structure and the hexapeptide motif are highly conserved, despite the considerable divergence in sequence and function observed in the LβH superfamily (Parisi, Fornasari, and Echave 2000).

Probability Distributions

In order to compare the outcome of our simulations with the sequence patterns of actual sequences, we used amino acid probability distributions and entropies. The probability distributions for each of the sequence site classes were calculated as follows. First, the sequences to be used to estimate the distribution were aligned. Second, a matrix \( H \) was built, where \( H(p, a) = 1 \) if amino acid \( a \) is found at column \( p \) of the multiple-sequence alignment, and \( H(p, a) = 0 \) otherwise. Finally, \( P(s, a) = \sum_{p \in s} H(p, a) / \sum_{p \in s} H(p, a) \) was calculated, where \( p \in s \) indicates that the sum is limited to sequence positions that belong to the same class.

Let \( P(a; M) \) and \( P(a; D) \) be, respectively, a simulated distribution obtained with model \( M \) and the distribution obtained from experimental data set \( D \). Then, the goodness of fit between the model and the data was measured using \( z_p(M, D) = [e(M, D) - e(M)] / \sigma_p(M) \), where the error was defined as \( e(M, D) = \sum_{a} [(P(a; M) - P(a; D))^2 / (P(a; M) + P(a; D))], \) and \( e(M) \) and \( \sigma_p(M) \) are the average and standard deviations of the errors obtained from comparing pairs of simulated runs. From such simulations, the distribution of \( z_p(M, D) \) was obtained numerically, and it was found that it could be fit by a normal distribution of zero mean and unit standard deviation. To compare the abilities of two models \( M_0 \) and \( M_1 \) to fit the observed amino acid distributions, we used \( z_p(M_0, M_1) = \sqrt{2} \ln \frac{z_p(M_0, D) - z_p(M_1, D)}{z_p(M_0, D) + z_p(M_1, D)} \), which has a normal distribution with zero mean and unit variance.

Entropies

The variability of each site class was characterized using the site class entropy. These entropies were calculated from the amino acid probability distributions in the usual way using \( S(s) = -\sum_{a} P(s, a) \ln P(s, a) \).

The entropies of a model \( M \) and experimental data set \( D \) were compared using \( z_S(M, D) = \sqrt{2} \ln \frac{S(D) - S(M)}{\sigma_S(M)} \), where \( S(D) \) is the entropy of \( D \), \( S(M) \) is the entropy of \( M \) averaged over independent runs, and \( \sigma_S(M) \) is the corresponding standard deviation. The cumulative distribution function was found numerically from simulations to be well fitted by \( P(z_S < z) = 2\Phi(z) - 1 \), where \( \Phi(z) \) is the normal cumulative distribution with zero mean and unit variance. As in the previous section, the abilities of two models \( M_0 \) and \( M_1 \) to fit the same data \( D \) can be compared using \( z_S(M_0, M_1) = \sqrt{2} \ln \frac{S(M_0, D) - S(M_1, D)}{S(M_0, D) + S(M_1, D)} \), whose distribution is approximately normal with zero mean and unit variance.

Results and Discussion

We begin by exploring the relationship between sequence divergence and constraint for structure conservation. While the hexapeptide motif is very well conserved in the LpxA family, sequences can definitely diverge, showing as little as 40% identity (Vuorio et al. 1994). Figure 2 shows that SCPE predicts a sigmoidal relationship between sequence divergence and tolerance to structural divergence (\( S_{div} \)). Even though sequences diverge, structure conservation limits this divergence to an extent that depends on the degree of constraint imposed by the environment. If structure divergence is too constrained (\( S_{div} \approx 0 \)), sequences cannot diverge at all, whereas in the limit of unconstrained evolution (\( S_{div} \approx \infty \)), they lead to effectively random sequences. Since \( S_{div} \) measures the tolerance of the environment to structural divergence, it is expected to depend on the protein’s function. This suggests the interesting possibility of a connection between figure 2 and the recent observation of a sigmoidal dependence of function similarity on sequence similarity (Wilson, Kreychman, and Gerstein 2000).

We further tested to see if the SCPE model can reproduce the characteristic variability pattern of the hexapeptide motif. The site entropy was used as a mea-
Fig. 3.—SCPE accounts for the observed hexapeptide variability pattern. Shown is site entropy versus hexapeptide position for the observed LpxA family (squares), and three SCPE simulations: maximum structural constraint $S_{\text{div}} = 0$ (SCPE0: circles), intermediate constraint $S_{\text{div}} = 6$ (SCPE6: diamonds). For the SCPE6 and SCPE$\propto$ cases, entropies are calculated for 11 SCPE runs, from which averages and standard deviations (error bars) are calculated. The SCPE0 variability plot is that of the LpxA of *Escherichia coli*, since no variation is allowed.

Table 1
Comparison of Variability (entropy) Patterns of Different Models with Experimental Data Sets

<table>
<thead>
<tr>
<th>Model(s)</th>
<th>Site 1</th>
<th>Site 2</th>
<th>Site 3</th>
<th>Site 4</th>
<th>Site 5</th>
<th>Site 6</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$z$</td>
<td>$p^d$</td>
<td>$z$</td>
<td>$p$</td>
<td>$z$</td>
<td>$p$</td>
</tr>
<tr>
<td>SCPE$^a$</td>
<td>5.61</td>
<td>0.000</td>
<td>0.15</td>
<td>0.881</td>
<td>0.45</td>
<td>0.653</td>
</tr>
<tr>
<td>SCPE$^b$</td>
<td>3.05</td>
<td>0.002</td>
<td>6.01</td>
<td>0.000</td>
<td>8.04</td>
<td>0.000</td>
</tr>
<tr>
<td>SCPE$\propto$</td>
<td>13.22</td>
<td>0.000</td>
<td>1.45</td>
<td>0.147</td>
<td>11.12</td>
<td>0.000</td>
</tr>
<tr>
<td>SCPE0 vs. SCPE6$^c$</td>
<td>-1.81</td>
<td>0.965</td>
<td>4.15</td>
<td>0.000</td>
<td>5.37</td>
<td>0.000</td>
</tr>
<tr>
<td>SCPE$^a$ vs. SCPE6$^c$</td>
<td>7.61</td>
<td>0.000</td>
<td>1.30</td>
<td>0.097</td>
<td>10.67</td>
<td>0.000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Site 1</th>
<th>Site 2</th>
<th>Site 3</th>
<th>Site 4</th>
<th>Site 5</th>
<th>Site 6</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$z$</td>
<td>$p$</td>
<td>$z$</td>
<td>$p$</td>
<td>$z$</td>
<td>$p$</td>
</tr>
<tr>
<td>LβH SUPERFAMILY$^b$</td>
<td>5.67</td>
<td>0.500</td>
<td>1.34</td>
<td>0.453</td>
<td>1.53</td>
<td>0.126</td>
</tr>
<tr>
<td></td>
<td>0.07</td>
<td>0.947</td>
<td>12.78</td>
<td>0.000</td>
<td>5.44</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>1.51</td>
<td>0.131</td>
<td>9.51</td>
<td>0.000</td>
<td>1.57</td>
<td>0.059</td>
</tr>
<tr>
<td></td>
<td>6.28</td>
<td>0.000</td>
<td>4.06</td>
<td>0.000</td>
<td>1.44</td>
<td>0.075</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Site 1</th>
<th>Site 2</th>
<th>Site 3</th>
<th>Site 4</th>
<th>Site 5</th>
<th>Site 6</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$z$</td>
<td>$p$</td>
<td>$z$</td>
<td>$p$</td>
<td>$z$</td>
<td>$p$</td>
</tr>
<tr>
<td></td>
<td>3.22</td>
<td>0.001</td>
<td>3.42</td>
<td>0.000</td>
<td>6.01</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>5.44</td>
<td>0.000</td>
<td>5.28</td>
<td>0.000</td>
<td>5.44</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>9.51</td>
<td>0.000</td>
<td>1.57</td>
<td>0.059</td>
<td>6.28</td>
<td>0.000</td>
</tr>
</tbody>
</table>

*a* LpxA family; the six hexapeptide sites are considered separately.

*b* LβH superfamily; only hexapeptide site 1, for which SCPE gives the worst results when compared with the LpxA data, is shown. To obtain the experimental pattern all the LβH families with members of known structure were considered.

*c* For comparison of model $M_0$ with data $D$ (first three rows), $z = z_0(M_0, D)$ (see Materials and Methods). Good models should result in low $z$ values. For comparison of two models $M_0$ and $M_1$, $z = z_1(M_0, M_1)$. For large $z$ values, $M_0$ should be rejected against $M_1$.

*d* For comparison of model $M_0$ with data $D$ (first three rows), $P$ is the probability of rejecting $M_0$ when it is correct. Good models should have low $P$ values. For comparison of two models $M_0$ and $M_1$, $P$ is the minimum significance level for which $M_0$ should be rejected against $M_1$.

*e* SCPE model with $S_{\text{div}} = 6$.

*f* SCPE model with $S_{\text{div}} = 0$. The protein sequence remains constant at its original value, the LpxA of *E. coli*.

*g* SCPE model with $S_{\text{div}} = \propto$. This is just the unconstrained Jukes-Cantor model with selection against stop codons.
Fig. 4.—SCPE predicts the hexapeptide amino acid distributions. The probability (bubble area) distribution for each hexapeptide site of the LpxA family (LpxA: light gray) is compared with three SCPE simulations: maximum structural constraint $S_{div} = 0$ (SCPE0: black), intermediate constraint $S_{div} = 6$ (SCPE6: dark gray), and no constraint $S_{div} = \infty$ (SCPE: white). Amino acids are sorted as follows: aliphatic (ACILMV), aromatic (FHWY), polar (NQST), charged positive (KR), charged negative (DE), and special (GP). SCPE0 is the distribution corresponding to LpxA of Escherichia coli, since no variation is allowed. On the other hand, SCPE is the distribution obtained with the Jukes-Cantor model with no selection, except for the rejection of stop codons. Therefore, the differences between SCPE6 and SCPE are due to the selection pressure against structural divergence imposed by $S_{div} = 6$. Triangles are used to mark differences between SCPE6 and LpxA. Triangles pointing upward denote amino acids that, despite being absent in the LpxA family, were found in other proteins of the $H$ superfamily considered. The SCPE6 and SCPE distributions were obtained by averaging over 11 runs of 550 amino acid replacements, which was enough to converge the probability distributions for the SCPE6 case. The first and last coils, C1 and C9 in figure 1, which are subject to constraints different from those of the other coils of LpxA, were not included in the calculation of probability distributions.

W are confirmed predictions, since they are also present in the LpxA family. Y, which does not appear in the LpxA distribution, is also a confirmed prediction, since we found it in other $H$ proteins. In general, all upward triangles in figure 4 mark amino acids predicted by SCPE that, despite not being found in LpxA, are found in other $H$ families. In contrast, downward triangles indicate differences between $S_{div} = 6$ SCPE and LpxA distributions that could not be found in the other $H$ proteins considered. Note, however, that the probabilities of most downward-triangle amino acids are so small that they are not likely to be found in a sample the size of the $H$ families considered. Moreover, it is interesting to note that even though downward-triangle amino acids may arise during evolution, they are selected against in $S_{div} = 6$ SCPE, as compared with the unconstrained case $S_{div} = \infty$.

In table 2, a quantitative comparison of the amino acid distributions of figure 4 is performed. The fourth row of table 2 shows that SCPE with $S_{div} = 6$ fits the LpxA distributions significantly better than SCPE with $S_{div} = 0$ for most hexapeptide sites. As with entropies, an exception is site 1, for which the LpxA distribution is closer to that of the LpxA of E. coli than to the $S_{div} = 6$ SCPE distributions. As before, the situation is reversed when other members of the $H$ superfamily are considered in the determination of the experimental pattern (last two columns of table 2). The last row of table 2 shows that the $S_{div} = 6$ SCPE gives significantly better results than the unconstrained case ($S_{div} = \infty$) for the conserved hexapeptide sites 1 and 3, but that the unconstrained model cannot be significantly rejected in favor of the constrained one for the variable sites 2, 4, 5, and 6.

Conclusions

This report presented a novel and general model of structurally constrained protein evolution, developed to study the effects of structural constraints on sequence divergence. For the $H$ domain of the LpxA family, with the only information of the sequence and structure of one of its members, the model predicts the sequence patterns characteristic of the whole family with a remarkable accuracy. Clearly, the general applicability of the SCPE model to other protein families remains to be studied, but it will take some time, since the model is computationally demanding. In this report, we aimed to present the model and show its applicability by studying one example case. From a mutational point of view, the present model treats all sites and all nucleotide replacements equivalently. Therefore, the observed biases in amino acid replacement patterns are a genuine outcome of the model, showing that they result naturally from constraining structural divergence.

Three considerations should be taken into account. First, SCPE is a neutral evolution model that cannot account for adaptive amino acid substitutions. However,
this is not a serious drawback, since such replacements are very rare (Perutz 1983; Golding and Dean 1998). Second, SCPE does not explicitly consider the folding pathway, whereas folding constraints are known to result in sequence conservation (Shakhnovich, Abkevich, and Petsyn 1996; Li, Mirny, and Shakhnovich 2000). Nevertheless, this should not be a major shortcoming, since folding seems to be largely determined by the native structure (Baker 2000). Finally, it is important to stress that introducing mutations at the gene level, rather than at the protein level, apart from being more realistic, makes this model potentially useful for studying issues such as the effects of nucleotide substitution biases on amino acid sequence patterns or the effects of selection at protein level on nucleotide substitution patterns.

The SCPE model can easily be improved by using a nucleotide mutation model that is more realistic than the Jukes-Cantor model. Also, different energy functions can be used to calculate the sequence-structure distance score. Finally, in the present case we accepted all sequences with scores \( S_{\text{dist}} < S_{\text{div}} \) and rejected those with \( S_{\text{dist}} > S_{\text{div}} \) but other dependencies of the probability of acceptance on \( S_{\text{dist}} \) could be used.

Regarding the dynamics of the substitutional process under the SCPE model, some of the issues that are currently being addressed in our group are (1) the site-dependent amino acid substitution probabilities under the SCPE model and their comparison with current models of protein evolution, (2) substitutional rate variation among amino acid sites, (3) correlations between the evolution of different amino acid sites, and (4) effects of structural constraints on the patterns of nucleotide substitution.

Even though our aim in building the SCPE model was to gain a better understanding of the process of molecular evolution, this model can also be useful in addressing phylogenetic inference issues. Thus, the model can be used to generate large benchmark data sets for the assessment of current probabilistic models. Furthermore, SCPE may be used to obtain structure-dependent substitution matrices and build structure-based probabilistic models that can be used, in turn, for phylogenetic inference purposes. Both issues are currently being studied in our group.

Acknowledgments

This work was supported by the Universidad Nacional de Quilmes and the Fundación Antorchas. J.E. is a Researcher of CONICET and a Guggenheim Fellow.

LITERATURE CITED


<table>
<thead>
<tr>
<th>Table 2</th>
<th>Comparison of Amino Acid Probability Distributions of Different Models with Experimental Data Sets</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MODEL(S)</strong></td>
<td><strong>LPXA FAMILY</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Site 1</strong></td>
</tr>
<tr>
<td>SCPE6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( z )</td>
</tr>
<tr>
<td>SCPE0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( z )</td>
</tr>
<tr>
<td>SCPE0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( z )</td>
</tr>
<tr>
<td>SCPE0 vs. SCPE6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( z )</td>
</tr>
<tr>
<td>SCPE0 vs. SCPE6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( z )</td>
</tr>
</tbody>
</table>

\( a \) LpxA family; the six hexapeptide sites are considered separately.

\( b \) LβH superfamily; only hexapeptide site 1, for which SCPE6 gives the worst results when compared with the LpxA data, is shown. To obtain the experimental dist, we used the LpxA family with members of known structure were considered.

\( c \) For comparison of model \( M_0 \) with data \( D \) (first three rows), \( z = z_0(M_0, D) \) (see Materials and Methods). Good models should result in low \( z \) values. For comparison of two models \( M_0 \) and \( M_1 \), \( z = z_0(M_0, M_1) \). For large \( z \) values, \( M_0 \) should be rejected against \( M_1 \).

\( d \) For comparison of model \( M_0 \) with data \( D \) (first three rows), \( P \) is the probability of rejecting \( M_0 \) when it is correct. Good models should have low \( P \) values.

\( e \) SCPE model with \( S_{\text{div}} = 6 \).

\( f \) SCPE model with \( S_{\text{div}} = 0 \). The protein sequence remains constant at its original value, the LpxA of ECOLI.

\( g \) SCPE model with \( S_{\text{div}} = \infty \). This is just the unconstrained Jukes-Cantor model with selection against stop codons.

William Taylor, reviewing editor
Accepted November 20, 2000