Protein Conformational Diversity Modulates Sequence Divergence

Ezequiel Juritz,[1] Nicolas Palopoli,[1] Maria Silvina Fornasari,[1] Sebastian Fernandez-Alberti,[1] and Gustavo Parisi*[1]

[1]Departamento de Ciencia y Tecnología, Universidad Nacional de Quilmes, Bernal, Argentina

***Corresponding author:** E-mail: gusparisi@gmail.com.

Associate editor: Jeffrey Thorne

Abstract

It is well established that the conservation of protein structure during evolution constrains sequence divergence. The conservation of certain physicochemical environments to preserve protein folds and then the biological function originates a site-specific structurally constrained substitution pattern. However, protein native structure is not unique. It is known that the native state is better described by an ensemble of conformers in a dynamic equilibrium. In this work, we studied the influence of conformational diversity in sequence divergence and protein evolution. For this purpose, we derived a set of 900 proteins with different degrees of conformational diversity from the PCDB database, a conformer database. With the aid of a structurally constrained protein evolutionary model, we explored the influence of the different conformations on sequence divergence. We found that the presence of conformational diversity strongly modulates the substitution pattern. Although the conformers share several of the structurally constrained sites, 30% of them are conformer specific. Also, we found that in 76% of the proteins studied, a single conformer outperforms the others in the prediction of sequence divergence. It is interesting to note that this conformer is usually the one that binds ligands participating in the biological function of the protein. The existence of a conformer-specific site-substitution pattern indicates that conformational diversity could play a central role in modulating protein evolution. Furthermore, our findings suggest that new evolutionary models and bioinformatics tools should be developed taking into account this substitution bias.

Key words: conformational diversity, sequence divergence, native state, protein evolution.

Introduction

Protein sequences are a source of invaluable information used in many diverse fields, such as molecular evolution, structural bioinformatics, and proteomics. This information is derived from site-specific substitution patterns encoded in an alignment as the differential probability of occurrence of the amino acids in different positions of a protein. The understanding of substitution patterns will provide us with a deeper insight of the underlying evolutionary process. This is a central issue to develop and improve bioinformatics tools that use sequence information. The discovery that the conservation of protein structure is related to sequence divergence has represented a significant improvement in the field (Lesk and Chothia 1980; Chothia and Lesk 1986). As proteins evolve, they are subjected to selective pressures to conserve their structures constraining sequence divergence, evidencing that sequences have structural information encoded in their substitution patterns. One of the first tools that implicitly captured the importance of this information was the use of profiles to search for distant homologous proteins (Gribskov et al. 1987), which soon after was followed by the use of Hidden Markov Models (HMM) for sequential studies (Krogh et al. 1994; Karplus et al. 1997). When structural signatures were considered explicitly, both methodologies increased their performance in distant homologous searches, fold recognition, and protein structure model assessment (Gribskov et al. 1988; Luthy et al. 1991, 1992; Eisenberg et al. 1992). The inclusion of structural information to obtain a better description of substitution patterns was in agreement with the early mechanistic view that certain amino acids occur preferentially in given secondary structure elements (Guzzo 1965; Levitt 1978). This view evolved to the idea that local structural environments could modulate the substitution patterns of amino acids (Overington et al. 1990; Overington 1992). As in the case of profiles and HMMs, evolutionary models gained reliability to describe the substitution pattern when they include explicit structural information. Several of these models have been developed considering a variety of structurally related properties including the protein fold and its thermodynamic stability, evolutionary derived potentials, physicochemical environments, quaternary protein structures, and structural perturbations due to nonsynonymous mutations (Koshi and Goldstein 1995; Bastolla et al. 1999; Dokholyan and Shakhnovich 2001; Parisi and Echave 2001; Fornasari et al. 2007; Kleinman et al. 2010).

Although the relationships between substitution pattern and protein structure are well established and have been used in several bioinformatics analysis, most of the conclusions were obtained describing the protein native state as a single structure. However, increasing experimental evidence supports the notion that the native state is better described by an ensemble of interchangeable conformers moving in a

Article

free-energy landscape (Volkman et al. 2001; James and Tawfik 2003; Henzler-Wildman et al. 2007; Lange et al. 2008). This notion has been previously used to explain the heterogeneity in the binding properties of bovine seroalbumin (Karush 1950). However, a more formal description was included in the Monod–Wyman–Changeux model of allosteric regulation also known as preexisting equilibrium model (Monod et al. 1965). The idea was later included in the concept of folding funnels (Bryngelson and Wolynes 1989; Bryngelson et al. 1995) explaining protein-folding pathways with a rugged bottom representing a collection of conformational isomers. According to this view, the populations of these conformers follow statistical thermodynamic distributions, and the energy barriers separating them define their conformational exchange (Nienhaus et al. 1997; Tsai, Ma, et al. 1999). The extent of conformational diversity is then related to the extent of the ruggedness at the bottom of the funnel including the distribution and heights of barriers between conformers. Rigid proteins, that is, proteins with relatively small structural differences between their conformers, can be related with very rugged landscapes. Similarly, more flexible proteins presenting large conformational diversity can be associated with very smooth landscapes. Moreover, the relative population in the conformational ensemble was shown to be related with the protein fold (Keskin et al. 2000), with the presence of certain mutations (Sinha and Nussinov 2001), and with the evolutionary history of the protein (Maguid et al. 2006). The early idea of the "static" folding funnels was later extended to the notion of dynamic funnels to include the effect of the environment in their overall shape (Kumar et al. 2000). Dynamic landscapes support the conformational binding selection hypothesis according to which the ligand selects the conformation with better affinity as antigens select the highest affinity antibody in the immunological response (Foote and Milstein 1994). Furthermore, despite that the highest ligand affinity conformer may correspond to a conformation with a high relative energy, the conformers belonging to a scarcely populated state in the ensemble could still bind to the ligand and shift the equilibrium toward the bound form to proceed with the reaction (Kumar et al. 2000). The description of dynamic landscapes offers a central view to explain structure–function–dynamics relationships (James and Tawfik 2003; Tokuriki and Tawfik 2009), and it is in agreement with most recent experimental data describing protein behavior (Boehr et al. 2006; Hilser 2010). These concepts provide more satisfactory explanations to several experimental data and have replaced the well-known and long-established "lock-key" (Fischer 1894) or "induced-fit" models to describe protein–ligand interaction (Koshland et al. 1958).

Considering the native state of proteins as an ensemble of conformers as well as the constraints that structure conservation imposes to sequence divergence, a specific characteristic substitution pattern should be expected for each conformer. Following this idea, the sequence information contained in an alignment of homologous proteins could be a complex blend of different structural constraints introduced by the native ensemble. Therefore, in the present work, we studied the effect that protein conformational diversity could have on the sequence substitution pattern originated from structural constraints. To this end, we estimated the conformational diversity using a data set of redundant structures for the same protein retrieved from the Protein Conformational Diversity Database (PCDB) recently developed in our group (Juritz et al. 2011). Our procedure is validated by previous works that have proved the correspondence between structural deformations detected under different crystallographic conditions and conformational changes related to the flexibility of the native state (Zoete et al. 2002; Best et al. 2006). For each protein with more than two different structures, we estimated its substitution pattern using SCPE (from structurally constrained protein evolution), a model of protein evolution developed previously (Parisi and Echave 2001). The main output of the model is a whole set of site-specific substitution matrices (Fornasari et al. 2002). The influence of each conformer in the substitution pattern found in the alignment of homologous proteins was studied using a maximum likelihood (ML) approach. Briefly, we found that conformational diversity strongly modulates the substitution pattern derived from structural constraints. Although each conformer has its own specific structural constraints, we found that, in 76% of the proteins under study, one conformer is associated with the best ML value and statistically outperforms the other members of the structural ensemble. Interestingly, 62% of these conformers are involved in cognate ligand binding, and only 28% of these correspond to the minimum relative energy structure.

Materials and Methods

Data Set Collections

We used the Protein Conformational Database (http://pcdb.unq.edu.ar) (Juritz et al. 2011) to retrieve a data set of 900 randomly chosen proteins with different degrees of conformational diversity. The maximum Root Mean Square Deviation (RMSDmax) between alpha carbon coordinates of the different conformers, calculated with MAMMOTH (Ortiz et al. 2002), is taken as a measure of the conformational diversity of the protein. In the data set studied, the range of RMSDmax is between 0 and 22.65 Å. The average RMSDmax value of multiple conformers for each protein is 7.5 Å. The solvent accessible area for each residue in each structure was obtained using the program Naccess (Hubbard and Thornton, Computer program, Dept of Biochemistry and Molecular Biology, University College London).

SCPE Simulations

Site-specific substitution matrices were obtained using SCPE (from structurally constrained protein evolution model) (Parisi and Echave 2001). SCPE simulates protein evolution by introducing random mutations in a protein with known crystallographic- or NMR-derived structure. Mutations are selected against too much structural perturbation using a score that measures the energetic difference introduced by the mutation. The mutation is accepted using a derived probability, which is a function of the score and a parameter λ that

is a measure of the selective pressure for the acceptance of nonsynonymous substitutions. The parameter λ is related with the selection pressure for structural conservation: high values of λ correspond to stronger structural constraints and then a low acceptance probability. On the other hand, low values of λ correspond with weaker structural constraints and a high acceptance probability. The main output of SCPE is a set of site-specific substitution matrices derived by counting the accepted mutations in the simulation (Fornasari et al. 2002). In this manuscript, SCPE was run using 8,000 independent calculations for each structure, with a default $\lambda = 0.25$ and a divergence time of 20 nonsynonymous substitutions per site in average. These parameter values were previously assessed as the best set of values for different unrelated proteins (Parisi G, unpublished results). Value $\lambda = 0.0001$ was used to explore mutational sites, implying a complete relaxation of the selective pressure against structural constraints. In this way, the obtained substitution matrices are dominated by the underlying mutational process that in SCPE is defined by an empirical codon substitution model (Schneider et al. 2005).

ML Calculations

ML calculations require a model of evolution, a multiple alignment, and the corresponding phylogenetic tree. The program HYPHY (Pond et al. 2005) was used with custom scripts to allow the inclusion of the site-specific substitution matrices from SCPE. For each protein in the data set, an alignment of homologous sequences was derived from the HSSP database (Sander and Schneider 1993). Each alignment contains at least 20 sequences with no less than 35% of sequence identity referred to the sequence of the known structure. The phylogenetic inference was made for each alignment using the PROTPARS maximum parsimony approach (Felsenstein 1989). Jones, Taylor, and Thorton (JTT) model was used as a reference model (Jones et al. 1992). As it is very frequent that crystallographic structures have missing residues, all ML calculations were made for residues shared by all the conformers for each protein in the data set. A model is said to outperform another model if it has a significantly higher ML value. Model comparisons were statistically assayed using Akaike information criteria (AIC) coefficient (Akaike 1974), and a ranking for the estimated models was estimated using ΔAIC (Burnham and Anderson 2003). A difference of $\Delta\text{AIC} < 2$ was taken as the accepted upper limit measure of model support.

For each protein, the SCPE sites were identified comparing SCPE and JTT performances for each site. SCPE sites were estimated for each protein using each of the different structures derived from PCDB representing the conformational ensemble of the native state. As SCPE is a structurally constrained model, each structure originates a different substitution pattern. When SCPE outperforms JTT in a particular site, it is called a SCPEs (SCPE site/s), otherwise, it is a JTTs (JTT site/s). When both models are undistinguishable, the sites are called Ms (Mutational site/s). We then registered common SCPEs to all structures and conformer-specific sites. We have also obtained the total ML (sum of the log ML per site) using SCPE runs and JTT. The statistical performance of the total ML obtained was again assessed using ΔAIC.

Ligand Occurrence and Minimum Relative Energy Data Set

To study the relationships between ligand occurrence, relative energy between conformers, number of interatomic contacts, and likelihood values obtained for the different structures, we used a set of 55 proteins (and their 320 corresponding structures) extracted from the original data set. These proteins fulfill some conditions required for the analysis. For example, the relative energy between the conformers for each protein was estimated using the knowledge-based potential derived by Ferrada and Melo (2009), which requires lengths above 90 residues. The presence of cognate ligands in each structure was made using the Procognate database (Bashton et al. 2008). The cognate ligands deposited in this database are those involved in the biological function of the proteins. The proteins in the data set contain at least a pair of conformer with and without a bound ligand. The number of contacts in each protein conformer were determined using the contact definition of Berrera et al. (2003) that is also used by the SCPE program (Parisi and Echave 2001).

The list of the proteins and structures used in this work can be downloaded as supplementary material (Supplementary Material online). The scripts to perform site-specific ML calculations using HYPHY are available upon request to the author.

Results

Evolution under Structural Constraints

We are interested in the analysis of protein conformational diversity impact on amino acid substitution pattern. To elucidate this effect, it is important to know in what extent the structural information is operating on protein evolution. For this purpose, ML calculations (Felsenstein 1981) were performed using alignments derived from the HSSP database and their estimated phylogenies (see Materials and Methods). We first compared ML estimations obtained using two models of protein evolution, the SCPE (Parisi and Echave 2001) and JTT (Jones et al. 1992) models. The SCPE is a model of evolution that explicitly considers the conservation of protein structure to simulate sequence divergence. For this reason, SCPE is called a "constrained" model. The main result of SCPE runs is a whole set of site-specific substitution matrices. The model was proved to be useful for detecting signatures derived from structural constraints in protein families (Parisi and Echave 2001, 2004, 2005; Fornasari et al. 2007). On the other hand, JTT is a well-established and extensively used molecular evolution model that uses a single substitution matrix for all sites of any protein. This matrix is derived from the accumulation of amino acid replacements from a large number of proteins. Consequently, JTT can be considered an "unconstrained" model, in the sense that it does not consider explicitly specific structural constraints. We have selected JTT as a reference unconstrained model, but others

could have been chosen as well, for example, LG (Le and Gascuel 2008) or WAG (Whelan and Goldman 2001).

ML calculations allow us to compare the performance of both models to describe the substitution pattern found in the alignment. As it has been mentioned before, model comparison was made using AIC (Akaike 1974), and their statistical significance was evaluated following Burnham and Anderson (2003) (see Materials and Methods). In those cases where SCPE outperforms JTT, structural restrictions are expected to be an important component in the sequence alignment. Otherwise, when JTT outperforms SCPE, two different situations could occur. On one hand, the alignments might not contain enough structural information or SCPE is unable to detect it. On the other hand, a given physicochemical trend or other determining factor of the substitution pattern could be better described by JTT. Additionally, a third possible outcome is obtained when both models explain equally well the substitution pattern. We show below that these sites are not evolving under structural neither other physicochemical constrains.

Our data set is composed by a redundant set of 3,896 domains from the CATH database (Greene et al. 2007) that belong to 900 proteins from different structural families extracted from PCDB (Juritz et al. 2011). Using ML calculations for each domain, we found that SCPE outperforms JTT in 89% of the cases. This result suggests that most of the proteins in our data set have structurally constrained positions affecting sequence divergence. This hypothesis could be explored in more detail performing ML estimations and statistical evaluations per position to calculate the percentage of sites where SCPE outperforms JTT. We found an average of 37% of these sites per protein and we called them SCPEs (SCPE sites) (fig. 1). On the other hand, JTT outperforms SCPE in only 12% of the sites and then we called them JTTs (JTT sites). It is important to note that the classification of SCPE, JTT, and mutational sites is operational and probably depends on the reference unconstrained model considered. To explore this influence, we also compared SCPE with the LG model. The estimation of the fraction of SCPE sites using LG as reference model is 35% in average in 100 randomly taken structures of the data set used compared with 38% estimated using JTT as reference in the same set. Although LG model has been proven to outperform JTT model (Le and Gascuel 2008), the structural information these type of models contain seems to have little impact on the characterization of structurally constrained sites distribution.

SCPEs are protein positions that evolve under well-defined tertiary structure constraints. However, it is clear that additional structural constraints could influence protein divergence, like protein–protein interactions (Fornasari et al. 2002), misfolding (Drummond and Wilke 2008), or protein aggregation (Monsellier and Chiti 2007). The structural constraints in SCPE emerge as a function of the number of interatomic contacts between residues. In supplementary figure 1 (Supplementary Material online), the distribution of SCPEs and JTTs as a function of the number of contacts is shown. Most of the SCPEs have between four and six contacts per residue, whereas the JTTs have a mean of one. Furthermore,

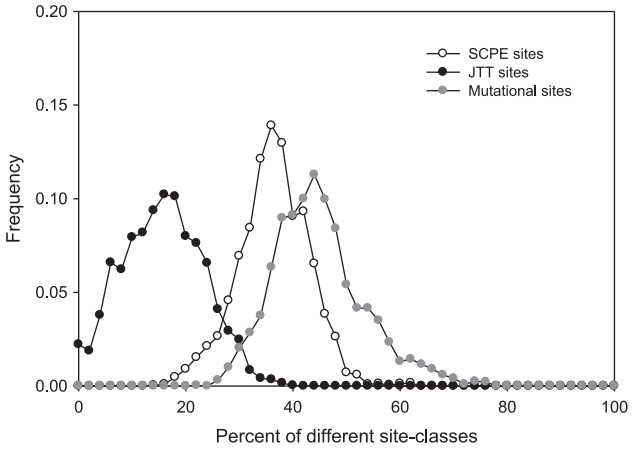

FIG. 1. Protein frequency distribution of percent of site classes in the proteins studied in our data set. Site classes describe different substitution patterns related with the corresponding structural constraints. The classes were derived from the comparison between an evolutionary model which contains structural information (SCPE) and other model without this information (JTT). Those sites where SCPE outperforms JTT are called SCPE sites, and those where JTT outperforms SCPE are JTT sites. When both models are indistinguishable, the sites are called mutational sites (Ms) because, at protein level and under these models, they could not be detected as subjected to structural or other physicochemical selective pressures. Taking only one structure to describe the native state of a protein, most sites evolve without structural or other physicochemical constraints (gray circles) followed by those subjected to structural constraints (white circles).

SCPEs usually correspond to buried residues, whereas JTTs are mostly exposed to solvent (supplementary fig. 2, Supplementary Material online). In figure 2, the different types of sites are shown for an example protein.

It is interesting to note that the majority of the sites (in average 45%) cannot be classified neither as SCPEs nor JTTs (fig. 1 gray circles). These sites are equally well described by JTT and SCPE. To study their properties, an independent set of SCPE simulations were performed but without imposing structural constraints (see Materials and Methods). Under these conditions, the resulting substitution patterns correspond to the empirical codon substitution model used by SCPE in the mutational process (Schneider et al. 2005). This means that the substitution matrices obtained by these simulations contain no specific information about structural constraints. When they were used to calculate the ML per site, we found that 89% of the sites that are equally explained by SCPE and JTT are also equally explained by these mutational substitution matrices. Consequently, these sites are called Ms (mutational site/s), and according to our results, they evolve under no other constraints beyond the mutational process considered.

At this point, it is interesting to analyze how SCPEs, being in average only the 37% of all sites, can explain the overall better performance of SCPE over JTT in the 89% of the 3,896 structures studied. The reason for this outperformance could be found calculating the average ML difference per site ($ML_{SCPE} - ML_{JTT}$). We have found a value of 2.47 for SCPEs

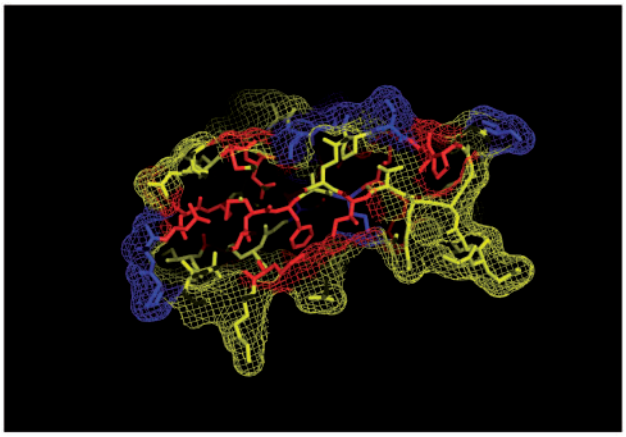

FIG. 2. Example using mesh and stick representation of the distribution of different site classes in a protein evolving under structural constraints (Bone morphogenetic protein receptor type IA; PDB code 1goo, CATH 1gooB00). The figure shows the relative distribution of SCPE sites (in red), JTT sites (in blue), and mutational sites (Ms) (yellow). The structural constraints over the substitution process along evolution explain why SCPE sites usually correspond to buried residues, whereas JTT sites are mostly solvent exposed.

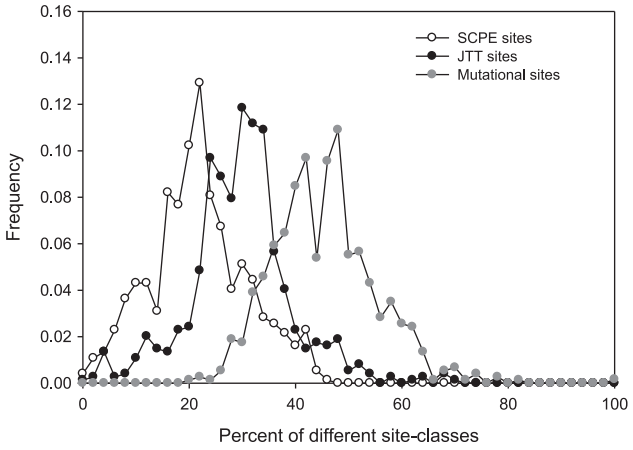

FIG. 3. Protein frequency distribution of percent of site classes in proteins with loose or absent structural constraints. As could be expected, SCPE and JTT sites change their relative positions compared with figure 1 where most proteins have larger proportions of structural constraints. Also, it is interesting to note that the fraction of mutational sites (gray circles) remains almost the same, independently of the level of structural constraints, when the native state of the protein is described with only one structure.

representing a significantly large improvement in the evolutionary description of these sites according to the AIC. Therefore, these improvements gained at the SCPEs are large enough to lead to an average improvement of the whole structure. However, an important proportion of the sites (45% in average) are equally described by the two models meaning that neither structural nor other physicochemical aspects are required to explain these sites. The consideration of conformational diversity will alter this view.

Evolution under Loose or Absent Structural Constraints

As it has been mentioned above, we found that JTT outperforms SCPE in 11% of the CATH domains in our data set. For these cases, structural constraints could be absent or SCPE could fail to adequately describe their substitution pattern. The individual analysis of these structures has shown that at least two main causes could explain the observed model performances. On one hand, we have found that in very few examples, the structures are associated with large ligands involving the occurrence of large cavities. These types of interactions (protein–ligand) are not explicitly considered in SCPE. Thus, the residues evolving under the structural constraints related to the binding of large ligands would be not correctly simulated by this model. On the other hand, we have stated that most of the structures belonging to this set correspond to structures with loose or absent structural constraints. Supplementary figure 3 (Supplementary Material online) shows proteins with different types of structural organization to exemplify these observations. Some of these proteins have very few secondary structural elements, large loops, or folds with low density of interresidue contacts. For these cases, the distribution between SCPEs and JTTs is inverted compared to structurally constrained proteins (fig. 1), whereas the distribution of Ms remains almost the same (fig. 3). Therefore, most of the residues in these proteins are structurally unconstrained (mutational sites). JTT matrix accumulates the substitution probabilities from different proteins and then from different structural and functional environments. This averaged information could explain the better description of the substitution pattern found in sites without specific structural constraints (in average 30%) but still showing a substitution bias with some kind of physicochemical pattern. As mutational sites are equally explained by the three models (SCPE, JTT, and mutational), this fraction of JTT sites explains the outperformance of this model over SCPE.

Evolution under Conformational Diversity

In order to study the effect of conformational diversity on protein sequence divergence, we used the same redundant collection of structures used before (from 3,896 CATH domain structures) to describe the substitution pattern of 900 proteins. We have, then, an average of 4.3 structures per protein to describe the influence of the conformational ensemble on the substitution process evaluated with ML methods. As a measure of conformational diversity, we used the RMSDmax derived from an all-against-all comparison between all collected structures of each protein in the database. The distribution of these values is shown in supplementary figure 4 (Supplementary Material online), and it is in agreement with the distribution reported by previous studies in a larger data set (Burra et al. 2009). More than 40% of the proteins in the data set show RMSDmax above 0.4 Å, which is the value commonly observed between different crystals of a protein obtained under the same crystallization conditions (Berman et al. 2000). Structures with differences above this

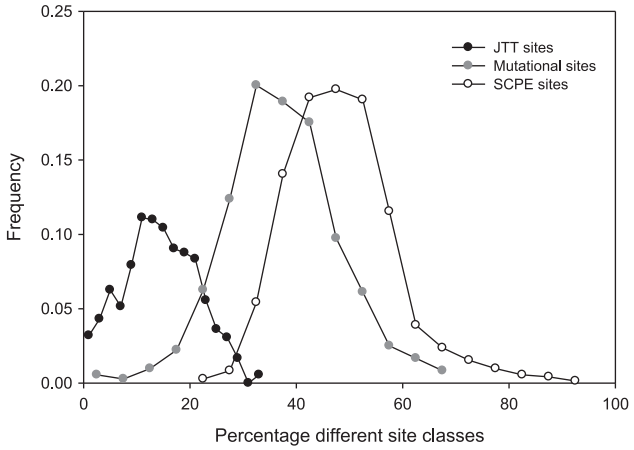

Fig. 4. Protein frequency distribution of site classes when different conformations for each protein are considered (900 proteins and their corresponding putative conformers). The main consequence of considering conformational diversity is the increase of the number of SCPE sites that now accounts for the majority of the site classes studied (please see fig. 1). This increase reflects that some sites could be classified as mutational or SCPE sites according to the selected conformer.

value are expected to be obtained in different crystallization conditions (i.e., presence/absence of ligands, changes in oligomerization state, changes in pH, etc.). The different structures could be taken as different instances of protein conformational diversity or protein dynamism (Tokuriki and Tawfik 2009). This view is supported by the correlation found between collections of crystallographic structures for the same protein and NMR measurements in several systems (Zoete et al. 2002; Best et al. 2006; Kondrashov et al. 2008; Friedland et al. 2009).

Using the same approach explained above to classify the sites, the number of SCPEs, JTTs, and Ms were calculated for each protein as the sum of the number of different sites over the different conformations. Figure 4 shows the distribution obtained for these three classes. It can be seen now that the average of the total number of SCPEs increases from 37% to 48% when all the members in the ensemble are considered. Comparing distributions in figures 1 and 4, it is clear that the main consequence of considering conformational diversity is an increment in the number of SCPEs that now accounts for the majority of the sites of the protein. Moreover, the total SCPEs can be classified as: 1) SCPEs shared by all the conformations (73% of the SCPEs), that is, sites that can be classified as SCPEs in all the conformations of the protein and, 2) SCPEs that are distinctive of a single conformation (27% of SCPEs), that is, sites that can be classified as SCPEs in only one of the conformations of the ensemble. To study the relationship between the percentage of SCPE sites and the extension of conformational diversity, we calculated the Spearman correlation coefficient using a log-log distribution between RMSDmax and the percentage of total SCPEs. A value of 0.36 was obtained using the SCPE sites collected over all the conformations. However, a better correlation (0.48) is obtained when we considered the percent of SCPEs that are characteristic of different single conformations. This last value evidently better captures the relationship between protein mobility and structural constraints in evolution.

It is important to note that in 76% of the proteins, we found that a single conformer outperforms the rest in terms of the total ML using the AIC test. The importance of the existence of this "best conformer" is related with its influence to explain sequence divergence. This feature could be associated with a stronger selective pressure during evolution; therefore, these conformers could be important from a biological perspective. In order to get a deeper understanding of this finding, we studied the relationship between the presence of a best conformer, the relative energies between conformers, and the presence of ligands. We have used a knowledge-based potential (Ferrada and Melo 2009) to estimate the relative energies between conformers and the Procognate database (Bashton et al. 2008) to assign the presence of ligands for each conformer in our set. Using the knowledge-based potential, we found that the best conformer corresponds to the conformer with the minimum energy in only 28% of the cases but in 62% to the conformer with a bound ligand. These findings could indicate that conformers with bound ligands, in spite of being less represented in the conformational ensemble, are under the strongest selective pressure influencing the substitution pattern during evolution.

Discussion and Conclusions

In this study, we have found that the presence of structural conformational diversity modulates protein sequence divergence. Although the conformers share most of their SCPEs, almost 30% of them are specific of each conformer. This means that a sequence alignment of homologous proteins contains conformational information, an observation that agrees with previous findings describing that dynamism is a conserved feature in protein evolution (Maguid et al. 2005, 2006, 2008). However, the different conformations in the ensemble do not impact in the substitution pattern in the same degree. It is interesting to note that among all the conformations, in 76% of the proteins, a single conformation (called here the "best conformer") outperforms the rest considering ML calculations. Searching for a deeper understanding of this result, we related the distribution of conformers to two of the major factors (relative energy and presence of ligands) influencing the equilibrium between the conformers in the ensemble according to the pre-equilibrium hypothesis (Monod et al. 1965; Tsai, Kumar, et al. 1999; Tsai, Ma, et al. 1999). We found that the characterized best conformer is mostly associated (62%) with the conformer that binds the ligand rather than with the conformer with the lowest relative energy. Consequently, the conformer with the lowest energy, that is usually associated with low binding activity (Kantrowitz and Lipscomb 1990; Velyvis et al. 2007), could have a weaker influence on sequence divergence than those conformers, less populated in the preexisting equilibrium, supporting the biological function due to the binding of cognate ligands.

In general, it is established that the so-called "close conformation" is associated with binding capacity and biological

activity (Gutteridge and Thornton 2005). Thus, it is possible that best conformer could be a compact conformer with a higher level of contacts between residues favoring the occurrence of SCPEs. However, in our data set, the best conformer is associated with the conformer with the relatively higher number of contacts in only 25% of the cases. In this way, the reported bias in the substitution patterns of the different conformers could reflect the outcome of differential selective pressures during evolution and may not be related with the conformers' compactness.

Our results are in agreement with previous ideas explaining general trends in protein evolution and protein folding, in particular the "minimal frustration" theory (Bryngelson et al. 1995). According to this view, amino acid replacements are selected for the occurrence of side chain interactions in such a way that favors the acquisition of the folded state of the protein. However, a small fraction of energetically "frustrated" residues could still occur and have been associated with binding residues (Ferreiro et al. 2007). In a similar way, and although the conformers share the majority of the SCPEs, we have found SCPEs specific of less-populated (higher energy) conformers and therefore energetically frustrated. SCPE simulations do not include selective functions to discriminate functional residues, using this term in the classical way to indicate residues associated with catalysis and binding. However, as protein function relies on conformational diversity (Karplus and Kuriyan 2005; Zhuravlev and Papoian 2010), the conformer-specific substitution pattern found in this work could indicate the existence of a broader class of "functional residues" related with the dynamism required to sustain the biological activity. Moreover, in a given protein family with well-conserved residues involved in catalysis and binding, it is expected certain functional diversification that could be related with the presence of substrate promiscuity as well as multispecificity (Tokuriki and Tawfik 2009). This functional divergence could be explained in terms of the conformational sampling of catalytic or binding residues due to local (flexibility of loops and side chains) and global (domain movements or fold transitions) rearrangements found in the conformational ensemble. Different mutations occurring during evolution could affect the relative population of conformers in the ensemble originating the functional divergence (Ma et al. 2002). The information associated to the conformer-specific SCPEs could help in the estimation of the sequence space defined by a given conformational ensemble and then could be an important tool in the characterization of functional clusters or subfamilies (Hannenhalli and Russell 2000; Abhiman and Sonnhammer 2005).

We also found that SCPEs represent as much as 37% of protein sites in average considering a single structure per protein and up to 48% when conformational diversity is taken into account. Additionally, we observed that the distribution of structurally constrained sites is very broad (spanning from 20% to 90%, fig. 3). Although it is beyond the scope of this manuscript, the amount of structural constraints has been recently related with the evolutionary rate of proteins. By far, one of the strongest and consistent correlations between genomic data and rates of evolutionary change is the expression level of genes (Drummond et al. 2005). Previous estimations have established that structural constraints could explain as much as 10% of the variation of evolutionary rate (Bloom et al. 2006). However, recent findings indicate that structure–function features and translation rates could have comparable contributions to explain evolutionary rates (Wolf et al. 2010). The observation that each conformation modulates the substitution pattern and that conformational diversity could involve in average 48% of the residues indicate that structural constraints could be a major force modeling protein divergence and could play a crucial role in the understanding of evolutionary rates.

Our results show the relevance of considering conformational diversity in our understanding of protein evolution mechanisms and consequently reveal the possibility to develop better evolutionary models. Besides that, the study of the substitution pattern and the understanding of the sequence conservation heterogeneity are central issues in several bioinformatics techniques. We suggest that the consideration of conformational diversity and its derived amino acid substitution bias is an essential aspect to be taking into account in the development of new bioinformatics tools.

Supplementary Material

Supplementary figures S1–S5 and material are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We would like to thank to Francisco Melo who kindly provided us with the potential, Matthew Bashton for his help with Procognate database, and Diego Ferreiro for helpful comments during the preparation of the manuscript. This work was funded by PIP CONICET grant 112-200801-02849 and Universidad Nacional de Quilmes. G.P., M.S.F., and S.F.A. are CONICET fellows, and E.J. has a Type II CONICET fellowship.

References

Abhiman S, Sonnhammer EL. 2005. FunShift: a database of function shift analysis on protein subfamilies. *Nucleic Acids Res.* 33:D197–D200.

Akaike H. 1974. A new look at the statistical model identification. *IEEE Trans Automat Contr.* 19:716–723.

Bashton M, Nobeli I, Thornton JM. 2008. PROCOGNATE: a cognate ligand domain mapping for enzymes. *Nucleic Acids Res.* 36: D618–D622.

Bastolla U, Roman HE, Vendruscolo M. 1999. Neutral evolution of model proteins: diffusion in sequence space and overdispersion. *J Theor Biol.* 200:49–64.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. 2000. The Protein Data Bank. *Nucleic Acids Res.* 28:235–242.

Berrera M, Molinari H, Fogolari F. 2003. Amino acid empirical contact energy definitions for fold recognition in the space of contact maps. *BMC Bioinformatics* 4:8.

Best RB, Lindorff-Larsen K, DePristo MA, Vendruscolo M. 2006. Relation between native ensembles and experimental structures of proteins. *Proc Natl Acad Sci U S A.* 103:10901–10906.

Bloom JD, Drummond DA, Arnold FH, Wilke CO. 2006. Structural determinants of the rate of protein evolution in yeast. *Mol Biol Evol.* 23: 1751–1761.

Boehr DD, McElheny D, Dyson HJ, Wright PE. 2006. The dynamic energy landscape of dihydrofolate reductase catalysis. *Science* 313: 1638–1642.

Bryngelson JD, Onuchic JN, Socci ND, Wolynes PG. 1995. Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins* 21:167–195.

Bryngelson JD, Wolynes PG. 1989. Intermediates and barrier crossing in a random energy model (with applications to protein folding). *J Phys Chem.* 93:6902–6915.

Burnham KP, Anderson DR. 2003. Model selection and multimodel inference: a practical information-theoretic approach, 2nd ed. New York: Springer-Verlag.

Burra PV, Zhang Y, Godzik A, Stec B. 2009. Global distribution of conformational states derived from redundant models in the PDB points to non-uniqueness of the protein structure. *Proc Natl Acad Sci U S A.* 106:10505–10510.

Chothia C, Lesk AM. 1986. The relation between the divergence of sequence and structure in proteins. *EMBO J.* 5:823–826.

Dokholyan NV, Shakhnovich EI. 2001. Understanding hierarchical protein evolution from first principles. *J Mol Biol.* 312:289–307.

Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH. 2005. Why highly expressed proteins evolve slowly. *Proc Natl Acad Sci U S A.* 102:14338–14343.

Drummond DA, Wilke CO. 2008. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* 134:341–352.

Eisenberg D, Bowie JU, Luthy R, Choe S. 1992. Three-dimensional profiles for analysing protein sequence-structure relationships. *Faraday Discuss.* 93:25–34.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol.* 17:368–376.

Felsenstein J. 1989. PHYLIP—phylogeny inference package (version 3.2). *Cladistics* 5:164–166.

Ferrada E, Melo F. 2009. Effective knowledge-based potentials. *Protein Sci.* 18:1469–1485.

Ferreiro DU, Hegler JA, Komives EA, Wolynes PG. 2007. Localizing frustration in native proteins and protein assemblies. *Proc Natl Acad Sci U S A.* 104:19819–19824.

Fischer E. 1894. Einfluss der Configuration auf die wirkung der Enzyme. *Ber Dtsch Chem Ges.* 27:2985–2993.

Foote J, Milstein C. 1994. Conformational isomerism and the diversity of antibodies. *Proc Natl Acad Sci U S A.* 91:10370–10374.

Fornasari MS, Parisi G, Echave J. 2002. Site-specific amino acid replacement matrices from structurally constrained protein evolution simulations. *Mol Biol Evol.* 19:352–356.

Fornasari MS, Parisi G, Echave J. 2007. Quaternary structure constraints on evolutionary sequence divergence. *Mol Biol Evol.* 24:349–351.

Friedland GD, Lakomek NA, Griesinger C, Meiler J, Kortemme T. 2009. A correspondence between solution-state dynamics of an individual protein and the sequence and conformational diversity of its family. *PLoS Comput Biol.* 5:e1000393.

Greene LH, Lewis TE, Addou S, et al. (14 co-authors). 2007. The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. *Nucleic Acids Res.* 35:D291–D297.

Gribskov M, Homyak M, Edenfield J, Eisenberg D. 1988. Profile scanning for three-dimensional structural patterns in protein sequences. *Comput Appl Biosci.* 4:61–66.

Gribskov M, McLachlan AD, Eisenberg D. 1987. Profile analysis: detection of distantly related proteins. *Proc Natl Acad Sci U S A.* 84:4355–4358.

Gutteridge A, Thornton J. 2005. Conformational changes observed in enzyme crystal structures upon substrate binding. *J Mol Biol.* 346: 21–28.

Guzzo AV. 1965. The influence of amino-acid sequence on protein structure. *Biophys J.* 5:809–822.

Hannenhalli SS, Russell RB. 2000. Analysis and prediction of functional sub-types from protein sequence alignments. *J Mol Biol.* 303:61–76.

Henzler-Wildman KA, Thai V, Lei M, et al. (12 co-authors). 2007. Intrinsic motions along an enzymatic reaction trajectory. *Nature* 450: 838–844.

Hilser VJ. 2010. Biochemistry. An ensemble view of allostery. *Science* 327: 653–654.

James LC, Tawfik DS. 2003. Conformational diversity and protein evolution—a 60-year-old hypothesis revisited. *Trends Biochem Sci.* 28: 361–368.

Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci.* 8: 275–282.

Juritz J, Fernandez-Alberti S, Parisi G. 2011. PCDB: a database of proteins with conformational diversity. *Nucleic Acids Res.* 39(1 suppl): D475–D479.

Kantrowitz ER, Lipscomb WN. 1990. Escherichia coli aspartate transcarbamoylase: the molecular basis for a concerted allosteric transition. *Trends Biochem Sci.* 15:53–59.

Karplus K, Sjolander K, Barrett C, Cline M, Haussler D, Hughey R, Holm L, Sander C. 1997. Predicting protein structure using hidden Markov models. *Proteins* (Suppl 1), 134–139.

Karplus M, Kuriyan J. 2005. Molecular dynamics and protein function. *Proc Natl Acad Sci U S A.* 102:6679–6685.

Karush F. 1950. Heterogeneity of the binding sites of bovine serum albumin. *J Am Chem Soc.* 72:2705–2713.

Keskin O, Jernigan RL, Bahar I. 2000. Proteins with similar architecture exhibit similar large-scale dynamic behavior. *Biophys J.* 78:2093–2106.

Kleinman CL, Rodrigue N, Lartillot N, Philippe H. 2010. Statistical potentials for improved structurally constrained evolutionary models. *Mol Biol Evol.* 27:1546–1560.

Kondrashov DA, Zhang W, Aranda R, Stec B, Phillips GN Jr. 2008. Sampling of the native conformational ensemble of myoglobin via structures in different crystalline environments. *Proteins* 70: 353–362.

Koshi JM, Goldstein RA. 1995. Context-dependent optimal substitution matrices. *Protein Eng.* 8:641–645.

Koshland DE Jr, Ray WJ Jr, Erwin MJ. 1958. Protein structure and enzyme action. *Fed Proc.* 17:1145–1150.

Krogh A, Brown M, Mian IS, Sjolander K, Haussler D. 1994. Hidden Markov models in computational biology. Applications to protein modeling. *J Mol Biol.* 235:1501–1531.

Kumar S, Ma B, Tsai CJ, Sinha N, Nussinov R. 2000. Folding and binding cascades: dynamic landscapes and population shifts. *Protein Sci.* 9: 10–19.

Lange OF, Lakomek NA, Fares C, Schroder GF, Walter KF, Becker S, Meiler J, Grubmuller H, Griesinger C, de Groot BL. 2008.

Recognition dynamics up to microseconds revealed from an RDC-derived ubiquitin ensemble in solution. *Science* 320:1471–1475.

Le SQ, Gascuel O. 2008. An improved general amino acid replacement matrix. *Mol Biol Evol.* 25:1307–1320.

Lesk AM, Chothia C. 1980. How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. *J Mol Biol.* 136:225–270.

Levitt M. 1978. Conformational preferences of amino acids in globular proteins. *Biochemistry* 17:4277–4285.

Luthy R, Bowie JU, Eisenberg D. 1992. Assessment of protein models with three-dimensional profiles. *Nature* 356:83–85.

Luthy R, McLachlan AD, Eisenberg D. 1991. Secondary structure-based profiles: use of structure-conserving scoring tables in searching protein sequence databases for structural similarities. *Proteins* 10: 229–239.

Ma B, Shatsky M, Wolfson HJ, Nussinov R. 2002. Multiple diverse ligands binding at a single protein site: a matter of pre-existing populations. *Protein Sci.* 11:184–197.

Maguid S, Fernandez-Alberti S, Echave J. 2008. Evolutionary conservation of protein vibrational dynamics. *Gene* 422:7–13.

Maguid S, Fernandez-Alberti S, Ferrelli L, Echave J. 2005. Exploring the common dynamics of homologous proteins. Application to the globin family. *Biophys J.* 89:3–13.

Maguid S, Fernandez-Alberti S, Parisi G, Echave J. 2006. Evolutionary conservation of protein backbone flexibility. *J Mol Evol.* 63:448–457.

Monod J, Wyman J, Changeux JP. 1965. On the nature of allosteric transitions: a plausible model. *J Mol Biol.* 12:88–118.

Monsellier E, Chiti F. 2007. Prevention of amyloid-like aggregation as a driving force of protein evolution. *EMBO Rep.* 8:737–742.

Nienhaus GU, Muller JD, McMahon BH, Frauenfelder H. 1997. Exploring the conformational energy landscape of proteins. *Physica D* 107: 297–311.

Ortiz AR, Strauss CE, Olmea O. 2002. MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. *Protein Sci.* 11:2606–2621.

Overington J. 1992. Structural constraints on residue substitution. *Genet Eng (N Y).* 14:231–249.

Overington J, Johnson MS, Sali A, Blundell TL. 1990. Tertiary structural constraints on protein evolutionary diversity: templates, key residues and structure prediction. *Proc R Soc B Biol Sci.* 241: 132–145.

Parisi G, Echave J. 2001. Structural constraints and emergence of sequence patterns in protein evolution. *Mol Biol Evol.* 18:750–756.

Parisi G, Echave J. 2004. The structurally constrained protein evolution model accounts for sequence patterns of the LbetaH superfamily. *BMC Evol Biol.* 4:41.

Parisi G, Echave J. 2005. Generality of the structurally constrained protein evolution model: assessment on representatives of the four main fold classes. *Gene* 345:45–53.

Pond SL, Frost SD, Muse SV. 2005. HyPhy: hypothesis testing using phylogenies. *Bioinformatics* 21:676–679.

Sander C, Schneider R. 1993. The HSSP database of protein structure-sequence alignments. *Nucleic Acids Res.* 21:3105–3109.

Schneider A, Cannarozzi GM, Gonnet GH. 2005. Empirical codon substitution matrix. *BMC Bioinformatics* 6:134.

Sinha N, Nussinov R. 2001. Point mutations and sequence variability in proteins: redistributions of preexisting populations. *Proc Natl Acad Sci U S A.* 98:3139–3144.

Tokuriki N, Tawfik DS. 2009. Protein dynamism and evolvability. *Science* 324:203–207.

Tsai CJ, Kumar S, Ma B, Nussinov R. 1999. Folding funnels, binding funnels, and protein function. *Protein Sci.* 8:1181–1190.

Tsai CJ, Ma B, Nussinov R. 1999. Folding and binding cascades: shifts in energy landscapes. *Proc Natl Acad Sci U S A.* 96:9970–9972.

Velyvis A, Yang YR, Schachman HK, Kay LE. 2007. A solution NMR study showing that active site ligands and nucleotides directly perturb the allosteric equilibrium in aspartate transcarbamoylase. *Proc Natl Acad Sci U S A.* 104:8815–8820.

Volkman BF, Lipson D, Wemmer DE, Kern D. 2001. Two-state allosteric behavior in a single-domain signaling protein. *Science* 291: 2429–2433.

Whelan S, Goldman N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol Biol Evol.* 18:691–699.

Wolf YI, Gopich IV, Lipman DJ, Koonin EV. 2010. Relative contributions of intrinsic structural-functional constraints and translation rate to the evolution of protein-coding genes. *Genome Biol Evol.* 2:190–199.

Zhuravlev PI, Papoian GA. 2010. Protein functional landscapes, dynamics, allostery: a tortuous path towards a universal theoretical framework. *Q Rev Biophys.* 43:295–332.

Zoete V, Michielin O, Karplus M. 2002. Relation between sequence and structure of HIV-1 protease inhibitor complexes: a model system for the analysis of protein flexibility. *J Mol Biol.* 315:21–52.