ELSEVIER

Biophysical Chemistry 124 (2006) 134–144

Biophysical Chemistry

http://www.elsevier.com/locate/biophyschem

Evaluation of models for the evolution of protein sequences and functions under structural constraint

Shruti Rastogi [a,b], Nathalie Reuter [b], David A. Liberles [a,b,*]

[a] *Department of Molecular Biology, University of Wyoming, Laramie, WY, 82071, USA*
[b] *Computational Biology Unit, BCCS, University of Bergen, 5020 Bergen, Norway*

Received 18 May 2006; received in revised form 13 June 2006; accepted 14 June 2006
Available online 22 June 2006

Abstract

In the field of evolutionary structural genomics, methods are needed to evaluate why genomes evolved to contain the fold distributions that are observed. In order to study the effects of population dynamics in the evolved genomes we need fast and accurate evolutionary models which can analyze the effects of selection, drift and fixation of a protein sequence in a population that are grounded by physical parameters governing the folding and binding properties of the sequence. In this study, various knowledge-based, force field, and statistical methods for protein folding have been evaluated with four different folds: SH2 domains, SH3 domains, Globin-like, and Flavodoxin-like, to evaluate the speed and accuracy of the energy functions. Similarly, knowledge-based and force field methods have been used to predict ligand binding specificity in SH2 domain. To demonstrate the applicability of these methods, the dynamics of evolution of new binding capabilities by an SH2 domain is demonstrated.

Keywords: Coarse-grained models; Folding; Binding; Population dynamics; Structural genomics; Molecular evolution

1. Introduction

Molecular evolution at the structural level plays an important role in structural genomics, which has been concerned with describing the "parts list" of protein structures that are used to construct various genomes [1]. At the same time, comparative genome analysis (and population genetic theory) have led to the hypothesis that many of the differences between genomes can be explained by differences in effective population size during the evolution of organisms [2]. In attempting to build a bridge between these two views of genome evolution, both a theoretical [3–6] and a lattice modeling framework [7–10] have been used for linking the evolution of sequence through structure to function. Three dimensional windows or contact maps have been developed to study the physical interactions governing protein folding [11,12] and more recently, the evolution of individual regions of a protein [13,14]. These can be used to extend lattice modeling studies to the study of real proteins with different folds. To enable such simulation studies, effective coarse-grained methods are needed to evaluate the folding and binding capabilities of real protein structures, ultimately enabling an understanding of the underlying rules that dictate the "parts list" for any genome.

In our previous work we have studied the mechanism of the evolution of proteins through the passage of gene duplication followed by mutation and selection with the help of three dimensional protein lattices [10]. In addition to functional selective pressures we evolved the protein lattices under structural selective pressure on the basis of folding energy, calculated with the use of defined interactions at the lattice points and contact potentials derived from a contact energy matrix [12,15]. Lattice models have previously been used in this context to make important predictions about the behavior of proteins in evolutionary contexts, including their metastability [16].

Since proteins are robust to site mutations and plastic in nature in that they accept mutations without destroying the fold [16–18], the development of an evolutionary model based upon population genetics theory together with the evolution of lattice

* Corresponding author. Department of Molecular Biology, University of Wyoming, Laramie, WY, 82071, USA. Tel.: +1 307 766 5206; fax: +1 307 766 5098.
E-mail address: liberles@uwyo.edu (D.A. Liberles).

0301-4622/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.bpc.2006.06.008

(or real protein-encoding) genes based upon either statistical or physical (force field) energy constraints is an other challenge. The evolutionary constraints for proteins include that they should perform a function and they must be stable enough to perform that function reliably while resisting unfolding, aggregation, and proteolysis.

In fact, a new field is emerging, as large scale gene and genome sequencing is enabling not only the comparison of closely related species, but increasingly of populations within a species. Simultaneously, the field of molecular evolution is increasingly interested in models of sequence evolution that incorporate structure. This combination is bringing together traditional population genetics with structural biology, where population-level variation can be examined not only at the sequence level, but also at the protein structural level in analyses characterizing variation in protein function.

One of the central problems in developing evolutionary models for proteins is developing an empirical energy function whose global minimum occurs when the protein is folded into the native state. Also, this can help us in analyzing the effect of mutations on evolving protein molecules to test if the global minimum has shifted. There are two classes of methods that are presently available for designing empirical energy functions. Knowledge-based methods depend upon contact interaction matrices derived from known protein folds in PDB and are widely used in molecular evolutionary studies. Alternatively, force field methods are parameterized with the forces governing interactions between atoms in proteins.

Some recent work has focused on deriving knowledge-based energy matrices both for long range and short range interactions between amino acids in protein folds based on RMSD between α carbons, the torsion and bond angle changes of virtual Cα–Cα bonds, and the coupling between them [19,20]. Another method involves deriving energy parameters for simplified models of folding based on the maximization of the thermodynamic average of the overlap between protein native structures and a Boltzman ensemble of alternative structures [21]. Lastly, a third approach uses an all atom model and a physical energy function [22].

In addition to developing an energy function, another challenge facing us is fast and accurate side-chain conformation prediction. An efficient approach uses results from graph theory to solve the combinatorial problem encountered in the side-chain prediction problem [23]. In this method, side chains are represented as vertices in an undirected graph. Any two residues that have rotamers with nonzero interaction energies are considered to have an edge in the graph. The resulting graph can be partitioned into connected subgraphs with no edges between them. These subgraphs can in turn be broken into biconnected components, which are graphs that cannot be disconnected by removal of a single vertex. The combinatorial problem is reduced to finding the minimum energy of these small biconnected components and combining the results to identify the global minimum energy conformation.

In this study we compare various computationally fast methods to evaluate which methods are best able to characterize the folding and binding of real proteins for use in population genomic studies where large numbers of sequences, folds, and mutations need to be evaluated. We further evaluate how these methods can be used to model the evolution of new binding functionalities in the absence of gene duplication. Ultimately, such methods can be used towards developing a better understanding of the structural "parts list" found differentially in various genomes.

2. Methods

In order to evaluate various methods for folding and binding, we have analyzed three protein folds from the categories, only α (only containing alpha helices), only β (only containing beta sheets), α+β (mainly antiparallel beta sheets (segregated alpha and beta regions)) and α/β (mainly parallel beta sheets (beta–alpha–beta units)): Globin-like, SH3 domain, SH2 domain and Flavodoxin-like, respectively. We have downloaded all coordinate files for these structures from Protein Data Bank (PDB) [24] and analyzed them with the methods described below. Fig. 1 shows how diverse the sequences are in each fold using a sequence logo based on a multiple sequence alignment generated using program CLUSTAL W with default parameters [25] and logos generated by WebLogo [26].

In all the methods, we have calculated fold energy parameters for different proteins belonging to the above-mentioned four domains and also threaded 1000 random sequences with the same amino acid frequencies as the real sequences through the protein backbones with the adjustment of the side-chain conformations using SCRWL3 [23]. SCRWL3 uses a few seconds with small proteins on a single desktop computer, but would require more time for big folds like Flavodoxin-like proteins, due to large number of side-chain clashes.

2.1. Protein folding models

In Method 1 a coarse-grained representation of protein structures is being used, as a binary contact matrix C_{ij} with elements equal to one if residues i and j are in contact (at least one pair of heavy atoms, one for each amino acid, are less than 4.5 Å apart), and zero otherwise. The effective free energy for a sequence A in configuration C is approximated by an effective contact free energy function E(A, C) given as,

$$\frac{E(\mathrm{A},\ \mathrm{C})}{k_{\mathrm{B}}T} = \sum_{i<j} C_{ij} U\left(A_i, A_j\right),$$

where U is a 20×20 symmetric matrix with U(a, b) representing the effective interaction, in units of $k_{\mathrm{B}}T$, of amino acids a and b when they are in contact. The interaction matrix used is derived from [21] based on thermodynamic average of the overlap between protein native structures and a Boltzmann ensemble of alternative structures.

Model 2 is based on [27], which presents a simple force field for evaluating proper folding. Another important aspect of the model is the incorporation of the helix propensity rule into the

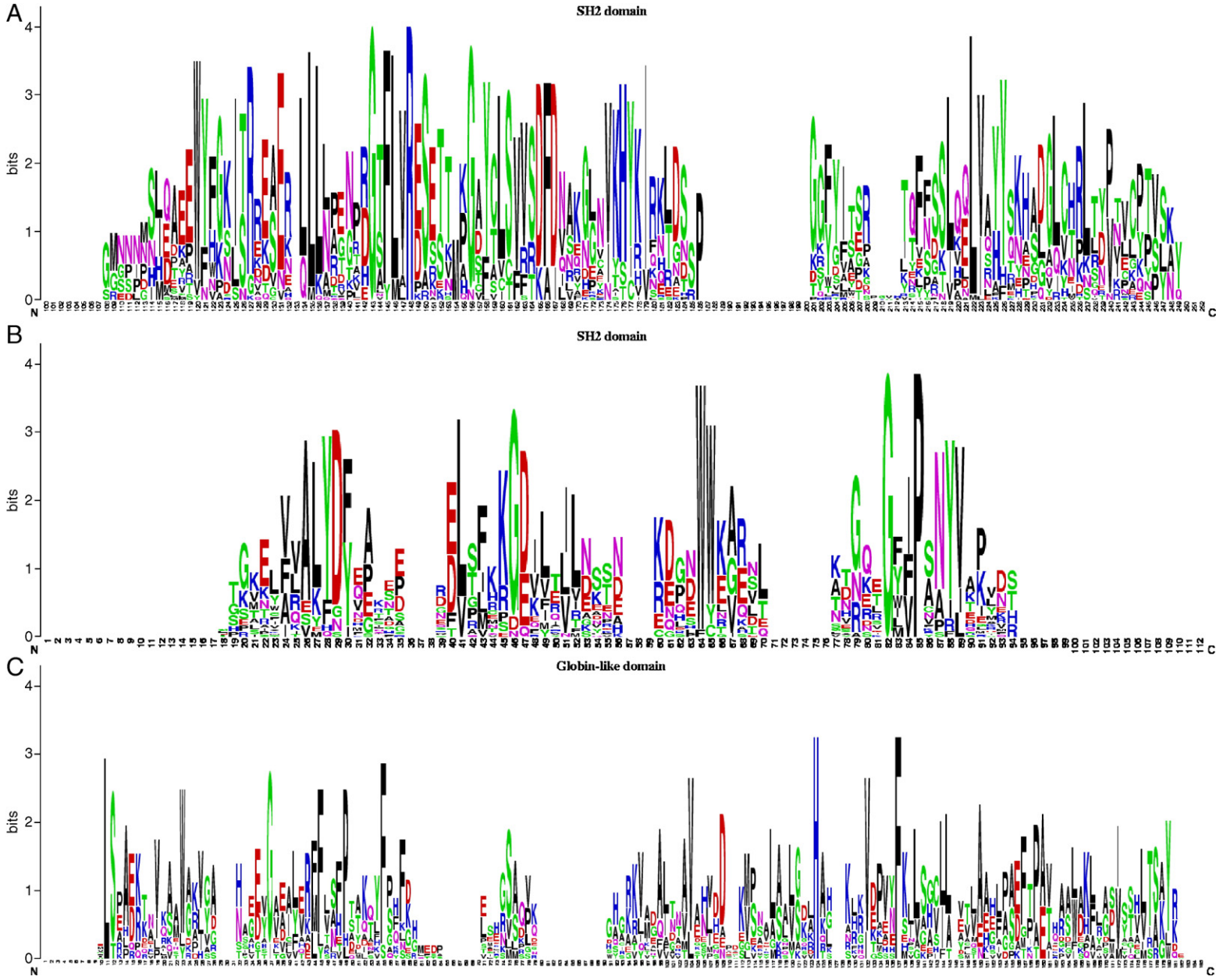

Fig. 1. The sequence logos demonstrating the sequence diversity of threaded structures (of those indicated in Table 1) for the four folds are shown for (A) SH2, (B) SH3, (C) Globin-like, and (D) Flavodoxin-like folds.

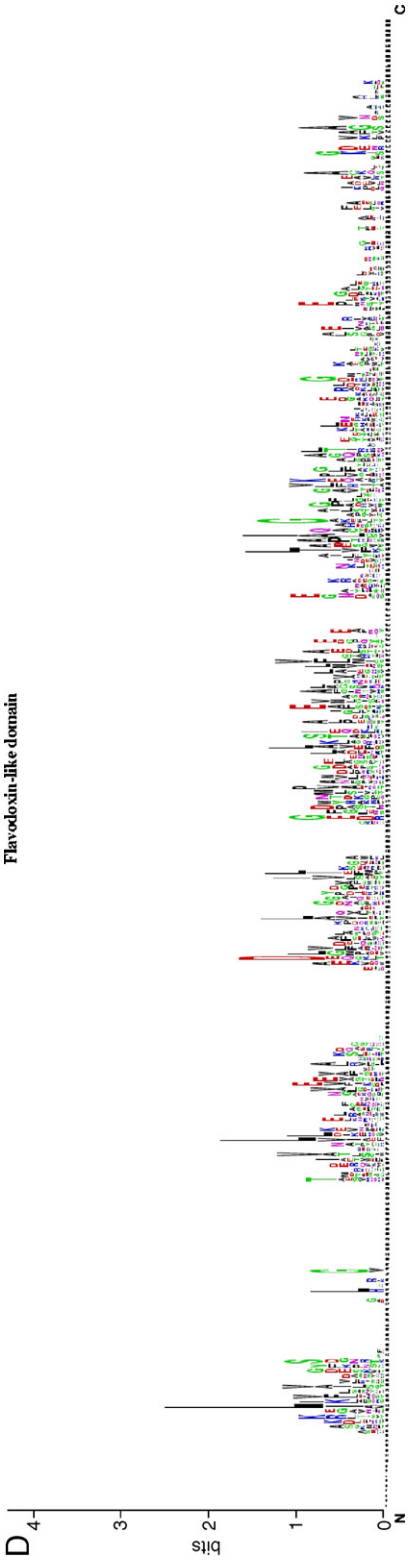

Fig. 1 (*continued*).

intramolecular potential. An α helix plays an important role in the early stages of protein folding and is the most prevalent type of secondary structure found in proteins. The folding potential for a protein sequence A in a configuration C is calculated according to

$$V_{\text{total}} = V_{\text{bond}} + V_{\text{theta}} + V_{\text{torsion}} + V_{\text{LJ}} + V_{\text{helix}}$$

where V_{bond} is the bonding potential, the sum of the bonding energy between the Cα atoms (backbone atoms) and the side-chain residues, V_{theta} is the bending potential around a central atom, the sum of three bending potential terms involving two other backbone atoms and one side-chain residue, V_{torsion} is the torsional potential based on four torsional angles per residue between the two Cα atoms excluding the terminal bonds which contain only two torsional angles, V_{LJ} is the nonbonding potential, the sum of the pair of Lennard–Jones interactions (dispersion and repulsion energies), and V_{helix} is helix propensity (taken from [27]). The four torsional angles are the dihedral angle ω for rotation about the peptide bond, namely $C\alpha_1$–{C–N}–$C\alpha_2$, dihedral angle φ for rotation about the bond between N and Cα, dihedral angle ψ for rotation about the bond between Cα and the carbonyl carbon, and dihedral angle χ for rotation about bond between Cα and Cβ.

Model 3 is a modification of Model 2 in which we have included Beta Sheet propensity as V_{sheet} in the energy function. The propensity values for different amino acids were taken from [28].

In Model 4, the folding energy for a protein sequence A in configuration C is calculated according to a contact map, where two α carbon atoms of amino acids α_i and α_j in a protein are considered to be in contact with each other if their distance $\delta(\alpha_i, \alpha_j)$ is at some threshold t where t varies from 2 Å to 12.8 Å (and the sum of all interactions over the range of t is used) and the sequence separation, given as $|i-j| \geq 2$. The energy of the protein is calculated on the basis of the long range interaction energies [19]. Model 4 is similar to Model 1 but does not consider side-chain interactions.

In Model 5 the backbone fold energy was calculated according to the statistical potentials, as based on the virtual bond model for the bond angle and bond torsions [20,29], as

$$E_{\text{SR}}(\Phi) = \sum_{i=2}^{N-1} E(\theta_i) + \sum_{i=3}^{N-1} [E(\phi_i^-)/2 + E(\phi_i^+)/2 - \Delta E(\phi_i^-, \phi_i^+)] + \sum_{i=3}^{N-1} [\Delta E(\theta_i, \phi_i^-) + \Delta E(\theta_i, \phi_i^+)].$$

where the first summation is the energy due to bending of the backbone bond angles, the second is the energy due to torsion of bonds Φ^- and Φ^+ referring to the rotational angles of the virtual bonds preceding and succeeding the ith α-carbon, respectively. The last summation is for the pairwise inter-dependence of the torsion and or bond angle bending. The values of these energies are obtained from short range energy interaction matrices [20].

Model 6 is based upon the scoring function proposed in [30]. According to Sippl, a proper normalization of energies suitable for comparison of different sequences in a given conformation is obtained as the z-score given as:

$$z(s, \text{c}) = \frac{E(s, \text{c}) - \bar{E}(s)}{\sigma(s)}$$

where $E(s, \text{c})$ is the knowledge-based energy of sequence s in conformation c, $E(\text{s})$ is the average energy of s in all conformations of a database and $\sigma(\text{s})$ is the standard deviation of the corresponding distribution.

We have calculated z-scores for all four folds. In each case, the database was the collection of PDB files for the structures belonging to that particular fold instead of using the entire PDB database. The knowledge-based energy potentials used were given in [21].

Model 7 is based upon the scoring function proposed in [31]. This scoring function is based on the distinction between a set of inter-atomic distances within a structure $\{d_{\text{ab}}^{ij}\}$, where d is the distance between atoms i and j of type a and b. We have considered distance bins as 1.5 Å, 2.5 Å,..., 20.5 Å and counted the number of contacts in each bin for a particular real protein structure. The scoring function S is defined as,

$$S\left(\left\{d_{\text{ab}}^{ij}\right\}\right) = -\sum_{ij} \ln \frac{P(d_{\text{ab}}^{ij}|C)}{P(d_{\text{ab}}^{ij})} \propto -\ln P\left(C|\left\{d_{\text{ab}}^{ij}\right\}\right)$$

and

$$P(d_{\text{ab}}|C) = f(d_{\text{ab}}) = \frac{N(d_{\text{ab}})}{\sum_d N(d_{\text{ab}})}$$

where $N(d_{\text{ab}})$ is the number of contacts of amino acid types a and b in a particular distance bin d, $\sum_d N(d_{\text{ab}})$ is the total number of a–b contacts observed for all distance bins and $P(d_{\text{ab}})$ is the probability of finding amino acid types a and b in a distance bin d in any compact conformation, native or otherwise. And $P(d_{\text{ab}})$ is given as,

$$P(d_{\text{ab}}) = P(d) = f(d) = \frac{\sum_{\text{ab}} N(d_{\text{ab}})}{\sum_d \sum_{\text{ab}} N(d_{\text{ab}})}$$

where $\sum_{\text{ab}} N(d_{\text{ab}})$ is the total number of contacts between all pairs of amino acid types in a particular distance bin d and $\sum_{\text{d}} \sum_{\text{ab}} N(d_{\text{ab}})$ is the total number of contacts between all pairs of atom types summed over the d distance bins.

2.2. Ligand binding models for SH2 domain

SH2 domains [32] are motifs of approximately 100 amino acids, and are prevalent in proteins that play a crucial role in intracellular signal transduction. All SH2 domain proteins have the ability to bind pTyr-containing ligands specifically, dependent upon the neighboring amino acid residues. In addition to the folding energy calculations for SH2 domain proteins, we have analyzed their binding specificity with three methods. In each method we have considered a threshold value as the original binding energy of original ligand and

considered binding possibility of a random ligand if the binding energy is less than or equal to the threshold value.

In Method 1, we have used knowledge-based matrices [19] for amino acid interactions between protein and ligand based on a contact map with threshold RMSD values varying from 2.0 Å to 12.8 Å. In Method 2, we have used a different knowledge-based method, based upon contact between the heavy atoms of the ligand and protein at an RMSD less than 4.5 Å. The contact energy values were obtained from an interaction matrix derived from [21]. The matrices in Method 1 and Method 2 differ in both the contacts that they consider significant and in the original set of protein structures examined. In Method 3, we have used protein force field calculations for Van der Waals potentials, H-bonding and electrostatic potentials between protein and ligand. The energy functions calculated were based on energy parameters used in the AUTODOCK [33] and DelPhi [34] programs. DelPhi calculates electrostatic potentials in and around macromolecules. This program was used to calculate the electrostatic component of the binding energy. The Van der Waals component of the binding energy was calculated through the AUTODOCK program.

2.3. Protein evolutionary model

A population of 1000 haploid unicellular organisms was assumed to have an SH2 domin protein 1d4t, which binds to a peptide ligand KSLTIYAQVQK. A new ligand QQA-MAASGGPL was picked from a set of 1000 random ligands with maximum binding energy (the most difficult to evolve towards a new binding functionality) at a binding interface of the protein. This ligand was assumed to be present in the cell which does not bind to the protein at the start of the simulation, but conferred an advantage to the cell to be able to bind the new ligand (biological examples are numerous and include new metabolites, toxins, and many other types of ligands). Inclusion of the new ligand in the cell was required to model new functionality in protein during evolution. The protein molecule was evolved in a constant population of 1000 cells, where those cells that bound the second ligand were 5% more likely to appear in the next generation (a 5% selective advantage). The fitness function required molecules to fold and to bind to the original ligand (those that bound no ligand resulted in organismal death and were eliminated from the population as a simplification, given that the model ignores insertion/deletion events and inactivating mutations that knock out transcription). During the evolutionary process, the molecules evolved on the basis of a Poisson distribution with an average of 1 DNA mutation per generation and transitions twice as probable as transversions (chosen according to a random number generator). The evolutionary model implemented the standard nuclear vertebrate genetic code in translating DNA into protein, with no insertions or deletions. In every generation the energy calculations for each molecule in the population were performed using Model 1. Cells (gene sequences in our simplified model) in every generation were selected with 5% advantage if they contained molecules that bound to both the original and the new ligand rather than simply the original ligand.

2.4. Computation

All the calculations were done with dual Intel Xeon EM64T 2.8 GHz processor chips on a CentOS Linux platform computer with 2 GB RAM. Absolute times will vary with the computational setup and are given for relative comparison between methods.

3. Results

The folding energies of a set of PDB files (Table 1) were analyzed with various methods by threading both the native sequence and a set of random sequences through the established folds. It is assumed that only a very small fraction of random

Table 1
A list of PDB files used for SH2, SH3, Globin-like and Flavodoxin-like folds in this study

SH2 (α+β)	SH3 (only β)	Globin-like (only α)	Flavodoxin-like (α/β)
2ple.pdb	1shf.ent	3LHB.pdb	1AG9.pdb
2pld.pdb	1sem.ent	2HHB.pdb	1B1C.pdb
1sps.pdb	1qwf.ent	1VRF.pdb	1BDJ.pdb
1spr.pdb	1qwe.ent	1V5H.pdb	1C4W.pdb
1shd.pdb	1qly.ent	1UX9.pdb	1CHN.pdb
1shb.pdb	1qkx.ent	1UT0.pdb	1D4A.pdb
1sha.pdb	1qkw.ent	1UMO.pdb	1D4Z.pdb
1qad.pdb	1oeb.ent	1SDL.pdb	1DCF.pdb
1pic.pdb	1m8m.ent	1SDK.pdb	1DJM.pdb
1oo4.pdb	1kik.ent	1S61.pdb	1EAY.pdb
1oo3.pdb	1k1z.pdb	1R1Y.pdb	1EHC.pdb
1nzv.pdb	1jqq.pdb	1QXE.pdb	1F4V.pdb
1nzl.pdb	1jo8.pdb	1QSI.pdb	1FFG.pdb
1lun.pdb	1jeg.pdb	1QI8.pdb	1FFS.pdb
1lum.pdb	1i0c.pdb	1OJ6.pdb	1FFW.pdb
1luk.pdb	1i07.pdb	1O1O.pdb	1FQW.pdb
1lui.pdb	1hjd.pdb	1NS9.pdb	1FYV.pdb
1lkl.pdb	1h3h.pdb	1NQP.pdb	1FYW.pdb
1lkk.pdb	1gl5.pdb	1KR7.pdb	1H05.pdb
1kc2.pdb	1gcq.pdb	1KFR.pdb	1IZY.pdb
1ka7.pdb	1gcp.pdb	1J7Y.pdb	1IZZ.pdb
1ka6.pdb	1gbq.pdb	1J41.pdb	1JBE.pdb
1jwo.pdb	1g2b.pdb	1IRD.pdb	1JRL.pdb
1ju5.pdb	1fyn.pdb	1IDR.pdb	1M2F.pdb
1is0.pdb	1efn.pdb	1H97.pdb	1MIH.pdb
1ijr.pdb	1e6h.pdb	1G9V.pdb	1P5F.pdb
1i3z.pdb	1e6g.pdb	1DXU.pdb	1PDV.pdb
1h9o.pdb	1csk.pdb	1A3O.pdb	1QCZ.pdb
1fu6.pdb	1ckb.pdb		1QR2.pdb
1fu5.pdb	1cka.pdb		1QVW.pdb
1fbz.pdb	1bu1.pdb		1RW7.pdb
1f2f.pdb	1bbz.pdb		1T0I.pdb
1f1w.pdb	1b07.pdb		1TIK.pdb
1d4t.pdb	1aoj.pdb		1VHQ.pdb
1cwe.pdb	1abo.pdb		2DHQ.pdb
1ayc.pdb			3CHY.pdb
1ayb.pdb			6CHY.pdb
1aya.pdb			
1a1e.pdb			
1a1c.pdb			
1a1b.pdb			
1a1a.pdb			
1a09.pdb			
1a08.pdb			
1a07.pdb			

sequences will preferentially fold into any particular fold and that a method that differentiates sequences that are known to fold into a particular structure over random sequences therefore performs better. In every case the conformation of side chains was predicted using SCRWL3 [21] in combination with the model being tested. Table 2 shows the percentage of random proteins that have folding energies greater than the native protein sequence for the various methods.

It has been shown that no energy function of simple form can ensure thermodynamic stability to the native state of all proteins simultaneously, so individually optimized energy functions are needed [35]. In Model 1 the interaction energy matrix being used is derived from an optimized energy function based on the maximization of the thermodynamic average of the overlap between protein native structures and a Boltzmann ensemble of alternative structures and shows a large difference in fold energies between known and random protein sequences. This difference is depicted in Table 2 and in Fig. 1 of supplementary materials (available at http://www.wyomingbioinformatics.org/LiberlesGroup/Rastogi/Biophys_Chem_06) for the four folds analyzed (SH2 domains, SH3 domains, Globin-like and Flavodoxin-like). Model 2 and Model 3 are based on protein force field calculations and also show a large difference in fold energies between known and random protein sequences, as shown in Table 2 and in Figs. 2 and 3 of supplementary materials (available at http://www.wyomingbioinformatics.org/LiberlesGroup/Rastogi/Biophys_Chem_06) for the four folds. The only difference between Model 2 and Model 3 is the additional calculation of beta sheet propensity in Model 3 and their performances are extremely similar. Model 1 and Model 2 differ in computational time required in ratio 1:3 for threading and calculating fold energies for 1000 random sequences.

Model 4 and Model 5 use knowledge-based statistically derived energy matrices and show very small differences in fold energies between known and random protein sequences, as shown in Table 2 and in Figs. 4 and 5 of supplementary materials (available at http://www.wyomingbioinformatics.org/LiberlesGroup/Rastogi/Biophys_Chem_06). Model 6 provides a scoring function based on a structure specific *z*-score as described in Methods. Its performance is shown in Table 2 and in Fig. 6 of supplementary materials (available at http://www.wyomingbioinformatics.org/LiberlesGroup/Rastogi/Biophys_Chem_06). Model 7 is a scoring function based upon the set of inter-atomic distances within a structure and its performance is shown in Table 2 and Fig. 7 of supplementary materials (available at http://www.wyomingbioinformatics.org/LiberlesGroup/Rastogi/Biophys_Chem_06).

Table 3
Percentage of random ligands rejected over the known ligands for SH2 domain proteins, as analyzed using both knowledge-based and protein force field based methods

SH2 domain proteins	% of random ligand energy > known ligand energy (knowledge-based Model 1)	% of random ligand energy > known ligand energy (knowledge-based Model 2)	% of random ligand energy > known ligand energy (force field model)
1cwe	61.5	90.9	99.0
1d4t	49.1	83.3	99.3
1fu5	13.7	49.8	99.9
1h9o	80.1	64.3	–
1i3z	95.7	74.7	100.0
1ka6	83.3	82.4	100.0
1kc2	99.8	78.7	98.0
1lkl	33.7	65.7	100.0
1pic	90.0	81.2	100.0
2pld	88.1	78.6	100.0

We also analyzed the ligand binding specificity for SH2 domain proteins between known and random ligand sequences. We have seen that force field methods work very well in differentiating between a native ligand sequence and a random ligand sequence. Table 3 shows the percentage of 1000 random ligands that showed higher binding energies than the known SH2 protein ligands. Knowledge-based Model 1 shows a preference for known sequences over more than 70% of random ligands and knowledge-based Model 2 shows this preference over more than 80% of random ligand sequences. Model 3 based on protein force fields prefers known ligand sequences almost 100% of the time. The difference in computational time between these methods is in the ratio 1:1.5:4.5 h in analyzing 1000 random ligands binding to an SH2 domain protein.

We have also developed an evolutionary model for an SH2 domain protein under population genetics criteria for selection and drift in a constant population of 1000 individuals. In this model we have used the empirical energy function of Model 1 in analyzing the effect of mutation on protein folding and binding function in every generation. We have seen fixation of new function in 10 separate populations together with the old function in 200 generations without gene duplication

Table 2
Percentage of random sequences that showed folding energies greater than the proteins known to fold into that structure, as calculated using six models based on protein force field and knowledge-based methods

Folds	Model 1		Model 2		Model 3		Model 4		Model 5		Model 6 (score)		Model 7 (score)	
	P_g	P_l	P_g	P_l	P_g	P_l	P_g	P_l	P_g	P_l	P_g	P_l	P_g	P_l
SH2 (α+β)	99.8	0.1	52.7	0.0	52.8	0.0	18.6	22.7	19.2	13.6	99.0	0.0	61.9	3.6
SH3 (only β)	82.6	0.9	86.9	0.0	86.8	0.2	20.9	7.8	19.5	5.0	56.5	5.1	64.3	9.9
Globin-like (only α)	77.3	0.0	98.1	0.0	92.5	7.5	7.6	13.0	6.1	3.5	91.0	2.8	89.5	1.7
Flavodoxin-like (α/β)	76.2	0.2	82.3	5.4	82.2	6.3	24.4	21.4	39.0	4.9	37.7	29.7	54.9	9.2

P_g = percentage of random sequences with energies/scores worse than sequences known to fold into that structure.
P_l = percentage of random sequences with energies/scores better than sequences known to fold into that structure.

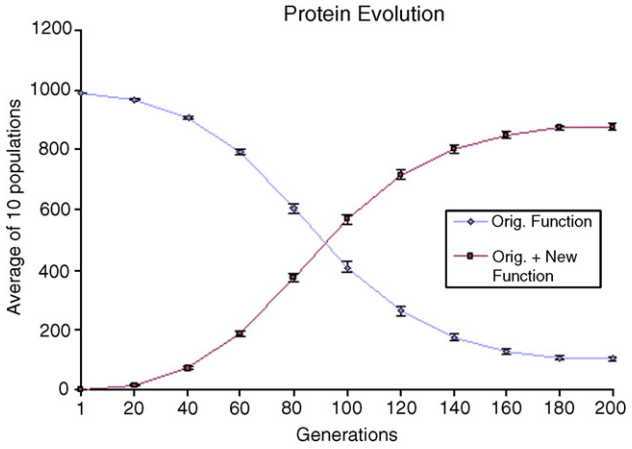

Fig. 2. Evolutionary process of an SH2 domain protein 1d4t.pdb showing fixation of a new binding functionality in the fixed population of 1000 individuals over 200 generations. Error bars represent standard error of the mean. Details of this simulation can be found in Methods.

(orthologous gene neofunctionalization) as shown in Fig. 2. The total running time for this process was approx. 30 h with the fastest knowledge-based method for a small population of 1000 individuals. But for large populations and genomic evolutionary studies and with an increase in parameters this running time will increase dramatically.

4. Discussion

Accurate models (for example, all-atom models that incorporate van der Waals effects, electrostatic interactions, amino acid rotamer information and other important physical principles) provide precise and realistic energies for a single protein structure. However, the computational time spent calculating each variant in a population of similar proteins would make population structural genomic studies impossible, even with the largest supercomputers. Therefore, computationally fast, more approximate methods have been evaluated for the purpose of pursuing such population structural genomic studies.

To evaluate protein folding, knowledge-based Model 1 and force-field based Models 2 and 3 performed best in capturing known sequences in the fold they are known to fold into over random sequences (most of which are not expected to fold stably into the fold they are threaded through). Model 1 uses a knowledge-based approach using native protein structures and a Boltzmann ensemble of alternative structures, which makes it more powerful in distinguishing real proteins from random sequences. On the other hand, knowledge-based Models 4 and 5 performed poorly, but contained limited information in the models and lacked descriptions of key folding forces. Models 6 and 7 differentiated sequences known to fold into a given structure from random sequences, but with a smaller separation than Model 1. Model 6 performed well on 3 of 4 folds, but was essentially random with regard to the Flavodoxin-like folds. Both models were fast, with Model 6 performing approximately equally fast as Model 1, while Model 7 was approximately 10% faster. Models 2 and 3 are force field models which perform well with all the folds but due to a long computational time, they are not suitable for large scale modeling. Ultimately, Model 1 seems to be the most appropriate for large scale modeling.

We have seen knowledge-based Model 1 perform better for small folds like SH2 and SH3 proteins whereas, force field methods work well for analyzing large folds like Globin-like and Flavodoxin-like. This is probably because most of the observed features in knowledge-based potentials are known to be biased to capture hydrophobic interactions better than other interactions [19]. For example, charged or polar residues are driven to the protein surface by the non-polar attractions of other amino acid residues, rather than their favorable interaction with the solvent. As a consequence, the performance of the contact potentials seems to be strongly dependent on the size and composition of the proteins, on the surface to volume ratio, and particularly, on the extent of burial of hydrophobic residues in the protein interior. On the other hand, the force fields used contained only a van der Waals force term to characterize hydrophobic interactions and this may have led to under-performance for folds where hydrophobic interactions are particularly important.

One caveat to the reliance of knowledge-based methods on hydrophobic interactions is their tendency to attribute a greater stability to hydrophobic residues on the surface based upon the interactions of these residues with buried shell residues they may be in contact with, giving the structure a greater apparent folding energy than it really has. As evidence for this, a comparison of protein structures taken from PDB for the four folds analyzed here with the 10% most stable random sequences calculated according to Models 1 and 2 showed that the percentage of surface hydrophobic residues was not significantly different between solved structures and random stable sequences as calculated by Model 2, while Model 1 showed a small but significant increase in the percentage of surface hydrophobic residues in stable random sequences as compared with solved structures (data not shown). Therefore, while knowledge-based methods can reliably be used for threading, care should be exercised when using knowledge-based methods to extract energetic properties of proteins, especially when features like the percentage of hydrophobic residues on the surface are changing.

To evaluate binding, both knowledge-based methods and force field based methods performed well, with the force field based method performing better. Even though the number of interactions being evaluated was smaller, the difference in computational time was more than double between knowledge-based and force field methods. Still, evaluating binding requires much less computational time than evaluating folding and the enhanced performance of the force field-based method would leave this as the preferred method for future large scale studies.

Lastly, these methods were tested on the ability to evolve folded SH2 domains with new binding capabilities under strong positive selective pressure. As seen in Fig. 3, the sequence evolves from a starting point, through a population of mixed sequences and functions after 200 generations, a population that has largely fixed a new binding function. At this point, as seen in Fig. 3C, there is still a large degree of sequence variation that

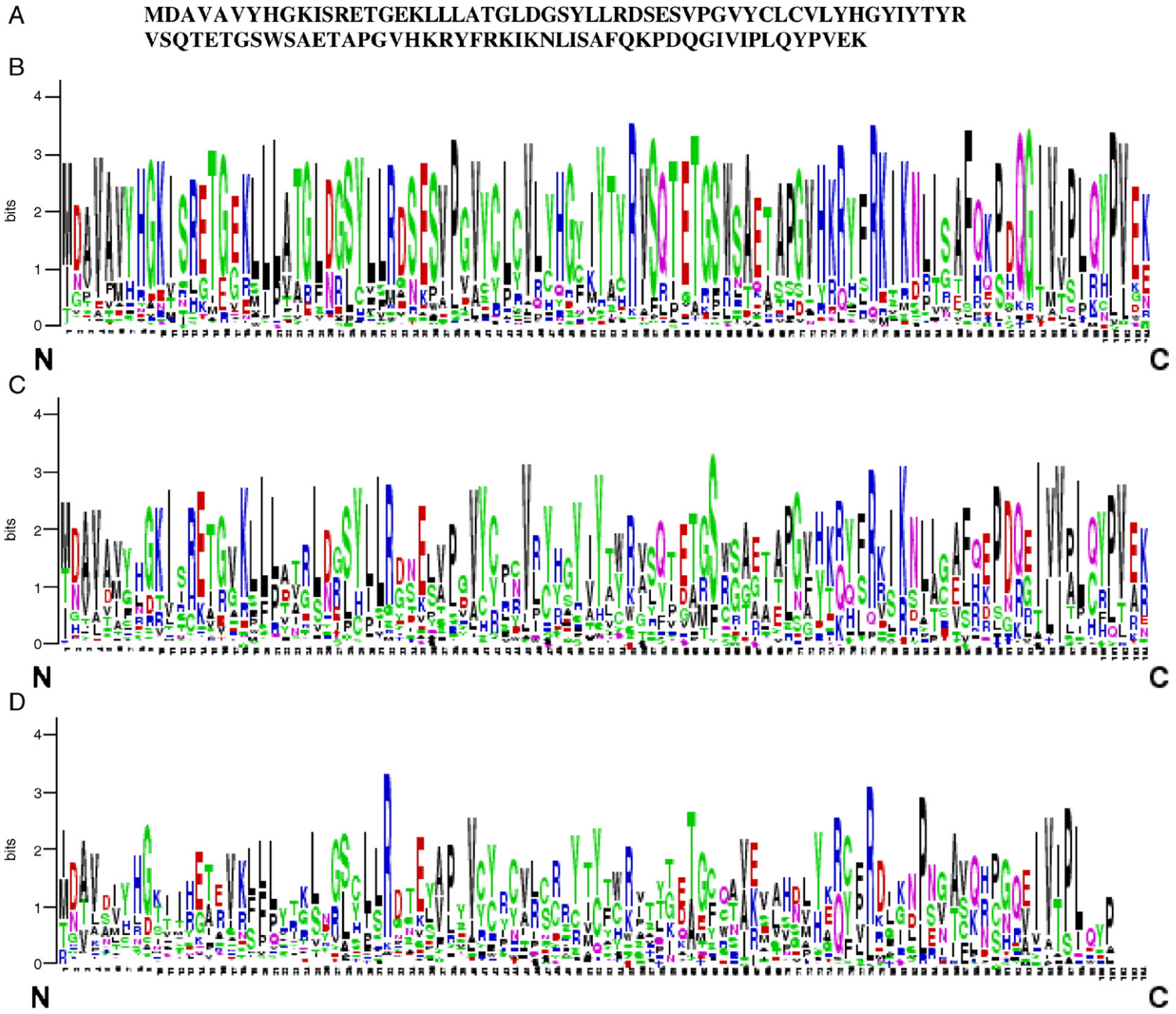

Fig. 3. Sequence logos are used to show the evolution of the population from the starting SH2 sequence (A) through populations at generations 60 (B) and 120 (C) with mixtures of binding functions through the final sequence population after 200 generations, showing largely a novel binding function, but still with large degrees of sequence variability, but all predicted to fold into SH2 domains.

can fold into an SH2 domain and bind to 2 different ligands, where one is a native ligand. This correlates with the sequence diversity that is known for existing SH2 domains [32]. Further, at a more general level, the sequence evolution of individual positions can be traced, including the co-evolutionary dependencies of individual positions on the amino acids at other positions. Additionally, individual residues that are linked to new functions can be discerned (for example the M to I mutation at position 1 in some sequence backgrounds).

Ultimately, these types of models can be used to test which folds are retained in genomes after duplication dependent upon their ability to evolve new functions that have been observed naturally. It is known that selective pressures are much stronger in organisms like Eubacteria with much larger effective population sizes and are weaker in organisms like mammals with smaller effective population sizes. Have 3.5 billion years of divergent evolution under different strengths of selection (the ancestral effective population sizes over this long time along the two lineages are, of course, not known) shaped proteins (and proteomes) that are fundamentally different in their abilities to evolve new functions neutrally versus under strong positive selection? That remains to be tested, but current models have now been evaluated that will enable a systematic testing of this hypothesis.

5. Conclusion

In this study we have compared various knowledge-based methods, and force field methods for protein folding and ligand binding specificity for four different folds, SH2, SH3, Globin-like, and Flavodoxin-like. One knowledge-based energy function (Model 1) showed the best results in differentiating native protein sequences from random sequences in short computational times. On the other hand, protein force field methods showed the best results in characterizing binding specificity for SH2 domain proteins. These combinations of methods can now be used for large scale population structural genomic studies to understand the “parts lists” of various genomes.

Acknowledgments

We are grateful to Arne Elofsson, Knut Teigen, and Jessica Liberles for helpful discussions. Funding for this work was provided by FUGE, the Norwegian functional genomics research platform.

References

[1] M. Gerstein, A structural census of genomes: comparing bacterial, eukaryotic, and archaeal genomes in terms of protein structure, J. Mol. Biol. 274 (1997) 562–576.
[2] M. Lynch, J.S. Conery, On the origins of genome complexity, Science 302 (2003) 1401–1404.
[3] F.N. Braun, D.A. Liberles, Retention of enzyme gene duplicates by subfunctionalization, Int. J. Biol. Macromol. 33 (2003) 19–22.
[4] F.N. Braun, D.A. Liberles, Repeat modulated population genetic effects in fungal proteins, J. Mol. Evol. 59 (2004) 97–102.
[5] U. Bastolla, M. Porto, H.E. Roman, M. Vendruscolo, Statistical properties of neutral evolution, J. Mol. Evol. 57S1 (2003) 103–119.
[6] M. Lynch, Simple evolutionary pathways to complex proteins, Protein Sci. 14 (2005) 2217–2225.
[7] S. Govindarajan, R.A. Goldstein, Evolution of model proteins on a foldability landscape, Proteins Struct. Funct. Genet. 29 (1997) 461–466.
[8] P.D. Williams, D.D. Pollock, R.A. Goldstein, Evolution of functionality in lattice proteins, J. Mol. Graph. Model. 19 (2001) 150–156.
[9] G. Tiana, N.V. Dokholyan, R.A. Broglia, E.I. Shakhnovich, The evolution dynamics of model proteins, J. Chem. Phys. 121 (2004) 2381–2389.
[10] S. Rastogi, D.A. Liberles, Subfunctionalization of duplicated genes as a transition state to neofunctionalization, BMC Evol. Biol. 5 (2005) 28.
[11] H.S. Chan, K.A. Dill, Origins of structure in globular proteins, Proc. Natl. Acad. Sci. U. S. A. 87 (1990) 6388–6392.
[12] M. Vendruscolo, E. Domany, Pairwise contact potentials are unsuitable for protein folding, J. Chem. Phys. 109 (1998) 11101–11108.
[13] Y. Suzuki, Three dimensional window analysis for detecting positive selection at structural regions of proteins, Mol. Biol. Evol. 21 (2004) 2352–2359.
[14] A.C. Berglund, B. Wallner, A. Elofsson, D.A. Liberles, Tertiary windowing to detect positive diversifying selection, J. Mol. Evol. 60 (2005) 499–504.
[15] S. Miyazawa, R.L. Jernigan, Estimation of effective inter-residue contact energies from protein crystal structures—quasi-chemical approximation, Macromolecules 18 (1985) 534–552.
[16] D.M. Taverna, R.A. Goldstein, Why are proteins marginally stable? Proteins Struct. Funct. Genet. 46 (2002) 105–109.
[17] D.M. Taverna, R.A. Goldstein, Why are proteins so robust to site mutations? J. Mol. Biol. 315 (3) (2002) 479–484.
[18] B.E. Shakhnovich, E. Deeds, C. Delisi, E. Shakhnovich, Protein structure and evolutionary history determine sequence space topology, Genome Res. 15 (2005) 385–392.
[19] I. Bahar, R.L. Jernigan, Inter-residue potentials in globular proteins and the dominance of highly specific hydrophilic interactions at close separation, J. Mol. Biol. 266 (1997) 195–214.
[20] I. Bahar, M. Kaplan, R.L. Jernigan, Short-range conformational energies, secondary structure propensities and recognition of correct sequence-structure matches, Proteins Struct. Funct. Genet. 29 (1997) 292–308.
[21] U. Bastolla, J. Farwer, E.W. Knapp, M. Vendruscolo, How to guarantee optimal stability for most representative structures in the protein data bank, Proteins Struct. Funct. Genet. 44 (2001) 79–96.
[22] P. Koehl, M. Levitt, Protein topology and stability define the space of allowed sequences, Proc. Natl. Acad. Sci. U. S. A. 99 (2002) 1280–1285.
[23] A.A. Canutescu, A.A. Shelenkov, R.L. Dunbrack, A graph-theory algorithm for rapid protein side-chain prediction, Protein Sci. 12 (2003) 2001–2014.
[24] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shinyalov, P.E. Bourne, The protein data bank, Nucleic Acids Res. 28 (2000) 235–242.
[25] D. Higgins, J. Thompson, T. Gibson, J.D. Thompson, D.G. Higgins, T.J. Gibson, CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, Nucleic Acids Res. 22 (1994) 4673–4680.
[26] G.E. Crooks, G. Hon, J.M. Chandonia, S.E. Brenner, WebLogo: a sequence logo generator, Genome Res. 6 (2004) 1188–1190.
[27] A. Mukherjee, B. Bagchi, Correlation between rate of folding, energy landscape, and topology in the folding of a model protein HP-36, J. Chem. Phys. 118 (2003) 4733–4747.
[28] A. Kim, M.J. Berg, Thermodynamic β-sheet propensities measured using a zinc-finger host peptide, Nature 362 (1993) 267–270.
[29] N. Kurt, T. Haliloglu, C.A. Schiffer, Structure-based prediction of potential binding and nonbinding peptides to HIV-1 protease, Biophys. J. 85 (2003) 853–863.
[30] M. Wiederstein, M.J. Sippl, Protein sequence randomization: efficient estimation of protein stability using knowledge-based potentials, J. Mol. Biol. 345 (2005) 1199–1212.

[31] R. Samudrala, J. Moult, An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction, J. Mol. Biol. 275 (1998) 895–916.
[32] T. Pawson, G.D. Gish, P. Nash, SH2 domains, interaction modules, and cellular wiring, Trends Cell Biol. 12 (2001) 504–510.
[33] G.M. Morris, D.S. Goodsell, R.S. Halliday, R. Huey, W.E. Hart, R.K. Belew, A.J. Olson, Automated docking using a Lamarckian genetic algorithm and empirical binding free energy function, J. Comput. Chem. 19 (1998) 1639–1662.
[34] W. Rocchia, S. Sridharan, A. Nicholls, E. Alexov, A. Chiabrera, B. Honig, Rapid grid-based construction of the molecular surface for both molecules and geometric objects: applications to the finite difference Poisson–Boltzmann method, J. Comput. Chem. 23 (2002) 128–137.
[35] M. Vendruscolo, R. Najmanovich, E. Domany, Can a pairwise contact potential stabilize native protein folds against decoys obtained by threading? Proteins Struct. Funct. Bioinform. 38 (2000) 134–147.