PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Evolutionary information hidden in a single protein structure

Chien-Hua Shih, Chih-Min Chang, Yeong-Shin Lin, Wei-Cheng Lo, and Jenn-Kang Hwang*

Institute of Bioinformatics, National Chiao Tung University, HsinChu 30050, Taiwan, Republic of China

ABSTRACT

The knowledge of conserved sequences in proteins is valuable in identifying functionally or structurally important residues. Generating the conservation profile of a sequence requires aligning families of homologous sequences and having knowledge of their evolutionary relationships. Here, we report that the conservation profile at the residue level can be quantitatively derived from a single protein structure with only backbone information. We found that the reciprocal packing density profiles of protein structures closely resemble their sequence conservation profiles. For a set of 554 non-homologous enzymes, 74% (408/554) of the proteins have a correlation coefficient > 0.5 between these two profiles. Our results indicate that the three-dimensional structure, instead of being a mere scaffold for positioning amino acid residues, exerts such strong evolutionary constraints on the residues of the protein that its profile of sequence conservation essentially reflects that of its structural characteristics.

Proteins 2012; 80:1647–1657.

Key words: protein structure; sequence conservation; contact number; evolution; *B*-factors.

INTRODUCTION

Functionally and structurally important amino acids can be deduced from their level of conservation in families of homologous proteins. The conserved amino acids are usually involved in enzyme activity, ligand binding or protein–protein interactions, or are buried in the protein cores.[1] A single sequence is unable to convey the wealth of evolutionary information regarding conservation. Determining the level of conservation[2–5] at each amino acid site requires aligning families of homologous sequences and considering factors like amino acid occurrence frequency, stereochemical, or physicochemical properties, substitution matrices, phylogenetic trees, and the probabilistic models underlying the evolution.

It is well observed that families of homologous sequences usually share common three-dimensional folds.[1] Indeed, many successful homology modeling methods[6–8] are based on this observation. Therefore, it is expected that protein structures should contain common evolutionary information shared by their homologous sequences. Recent studies show that the protein structure is more than a mere scaffold for positioning residues. It has been shown that *B*-factors (or atomic mean-square displacements),[9–12] motional correlations in structure,[9,11,12] and the locations of catalytic residues[13–15] can be derived directly from the atomic coordinates of protein backbones without any additional assumptions about the protein models.

Here we report that evolutionary information regarding conservation at the residue level can be quantitatively extracted from a single structure. We show that generally, the sequence conservation profiles closely resemble those of packing density of the structures. Our results indicate that protein structure exerts such strong constraints on the evolvability of each residue that the profile of sequence conservation essentially reflects that of the structure.

RESULTS

Comparison of the weighted contact number and the conservation profiles

The weighted contact number (WCN) is the number of contact atoms at an amino acid site, weighted by the inverse square separation between residues represented (see METHODS). The WCN basically describes the packing density of a protein structure. The larger the WCN of a residue is, the more packed its environment. The conservation score of a protein is based on the evolutionary rate of each residue, computed using the evolutionary relations among homologous sequences and the

Additional Supporting Information may be found in the online version of this article.
Grant sponsors: National Science Council, The MoE ATU Program, Taiwan, R.O.C.
Chien-Hua Shih and Chih-Min Chang contribute equally to this paper.
*Correspondence to: Dr. Jenn-Kang Hwang, Institute of Bioinformatics, National Chiao Tung University, Hsin Chu 30050, Taiwan, R.O.C. E-mail: jkhwang@faculty.nctu.edu.tw
Received 22 August 2011; Revised 7 February 2012; Accepted 12 February 2012
Published online 20 February 2012 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/prot.24058

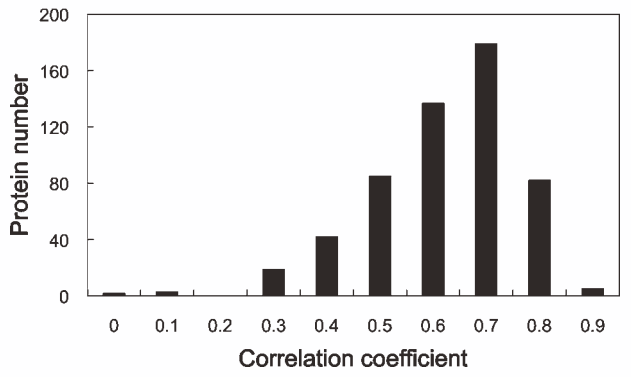

Figure 1

The distribution of the Pearson's correlation coefficients between the sequence conservation and the rWCN profile for the data set of 554 enzymes.

probability of residue replacement (see Methods). The conservation score is defined as such that the smaller the conservation score is, the more conserved the residue is. We compared the reciprocal weighted contact number (rWCN) profile of a structure with the corresponding conservation profile for the data set of 554 nonhomologous enzymes (Supporting Information Table S1; also see DATA). Enzymes were selected because they are among the most well studied proteins. The Pearson's correlation coefficient was used to quantify the correlation between the rWCN and the conservation profiles. The distribution of the Pearson's correlations is shown in Figure 1. Qualitatively, the results show that the residues of larger WCNs (or in more packed environments) are more conserved than those of smaller contact WCNs (or in less packed environments). The results are not surprising, since it is well known that the core residues of a protein are generally more conserved and the surface residues tend to be more variable.[1,16] The average correlation between these two profiles is 0.57. There are 408 out of 554 proteins (i.e., 74%) that have a correlation coefficient > 0.5.

Some examples are shown in Figure 2: the rWCN and the corresponding conservation profiles of three enzymes: transaldolase B (PDB ID: 1ONR:A), phosphofructokinase (PDB ID: 1KZH:A) and aconitase (PDB ID: 1FGH:A). The excellent agreements between the profiles indicate that the conservation profiles of the sequences essentially reflect their structural characteristics. The minima of the profiles, i.e., the more conserved regions, overlap extremely well. For reference, we marked the catalytic residues in the figures, all of which are located at the bottoms of the wells of both profiles. Recent studies show that the catalytic site residues tend to be located in regions of high packing density[15] or lie in close proximity to the structure centroids.[13,14] The protein centroids are known to be usually close to the regions of high packing density.[12,17] Therefore, our results provide a simple interpretation for these observations: since the conserved residues are usually located in the more packed environments, the catalytic residues, which are among the most conserved residues, are also most likely found in those environments.

We expect that our results are quite general and should be applicable to proteins other than enzymes, since, in terms of structure and sequence, enzymes do not differ much from other proteins. However, as examples, we compare the conservation and the rWCN profiles of several non-enzyme proteins. Their results are shown in Figure S1 in the Supporting Information. Their correlation coefficients range from 0.78 to 0.59.

Protein flexibility and sequence conservation

The *B*-factors, or the atomic mean-square displacements, describe protein flexibility. The atomic thermal fluctuations depend in a certain degree on their environments. When in more packed environments, the atoms are expected to have smaller thermal fluctuations (or smaller *B*-factors). On the other hand, when the atoms are in less packed environments, they are expected to have larger thermal fluctuations (or larger *B*-factors). Indeed, studies show that atomic contact numbers and *B*-factors are quantitatively related to each other.[10,12] Furthermore, it is reported that the *B*-factors of the protein backbone are conserved both at family and superfamily levels.[18] Therefore, it is reasonable to assume that the *B*-factors of a structure should have a certain degree of correlation with its sequence conservation profile. We computed the correlations between the *B*-factors and the conservation scores for the 544-protein data set. The distribution of the correlations is shown in Figure 3(a). The *B*-factors do correlate with the corresponding conservation scores, although their correlations are much weaker than those of the WCN's. Two examples are shown in Figure 3(b,c). The conserved residues tend to be more rigid (i.e., have smaller *B*-factors), while the variable residues tend to be more flexible (i.e., have larger *B*-factors). Since the contact number and *B*-factors have been shown to be closely correlated with each other,[10,12] the conservation of catalytic residues is also expected to be associated with their contact number and their *B*-factors.

Evolutionary coupling between subunits

There are 316 multichain structures in the 544-protein data set. For a multi-chain structure, the rWCN profile of a specific chain (or subunit) can be calculated in two ways: the WCNs are either computed using only one particular subunit, or the WCNs are computed using all the subunits of the structure. If the conservation profile of a given subunit agrees better with the rWCN profile computed using only that particular subunit, it is reasonable

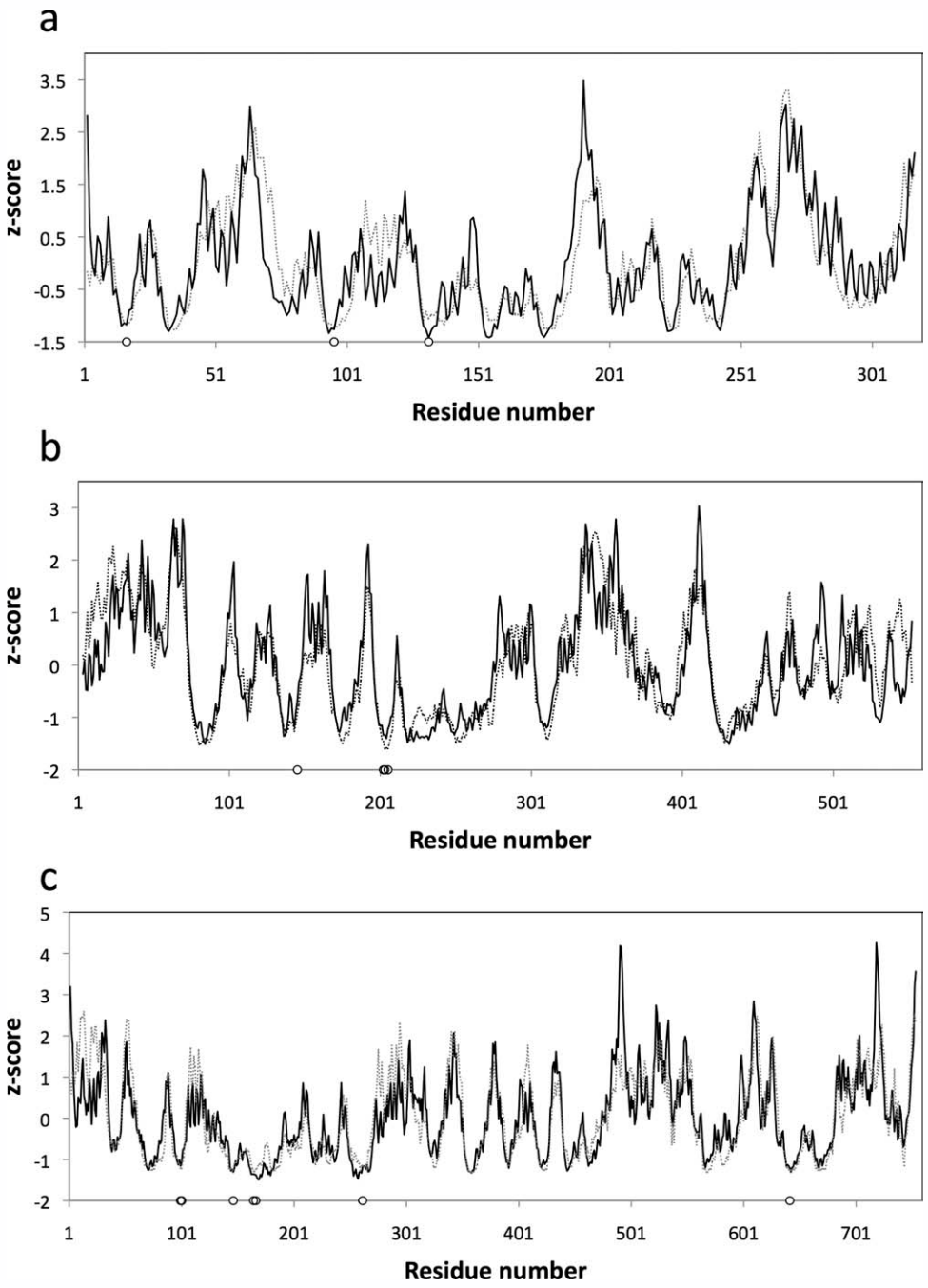

Figure 2

Comparison of the rWCN profile (black line) and the conservation profile (dotted line) of **a,** 1ONR:A, **b,** 1KZH:A, and **c**,1FGH:A. The empty circles mark the catalytic residues. Both the rWCNs and the conservation scores are normalized to their respective *z*-scores.

to assume that the evolution of the subunit is only weakly coupled with or is independent of the other subunits. On the other hand, if the conservation profile of a specific subunit agrees better with the rWCN profile computed using all the subunits of the structure, one will expect that the subunit may be evolutionarily coupled with the other subunits.

In our study, we observed both cases: 128 structures in the former and 188 in the latter. Two examples are presented. The structure of dihydrodipicolinate reductase (1ARZ) is composed of four identical subunits [Fig. 4(a)]. The correlation coefficient of 1ARZ:A is 0.18 if only chain A is used to compute the WCNs [Fig. 4(b)], but becomes 0.72 if all four chains are used [Fig. 4(c)].

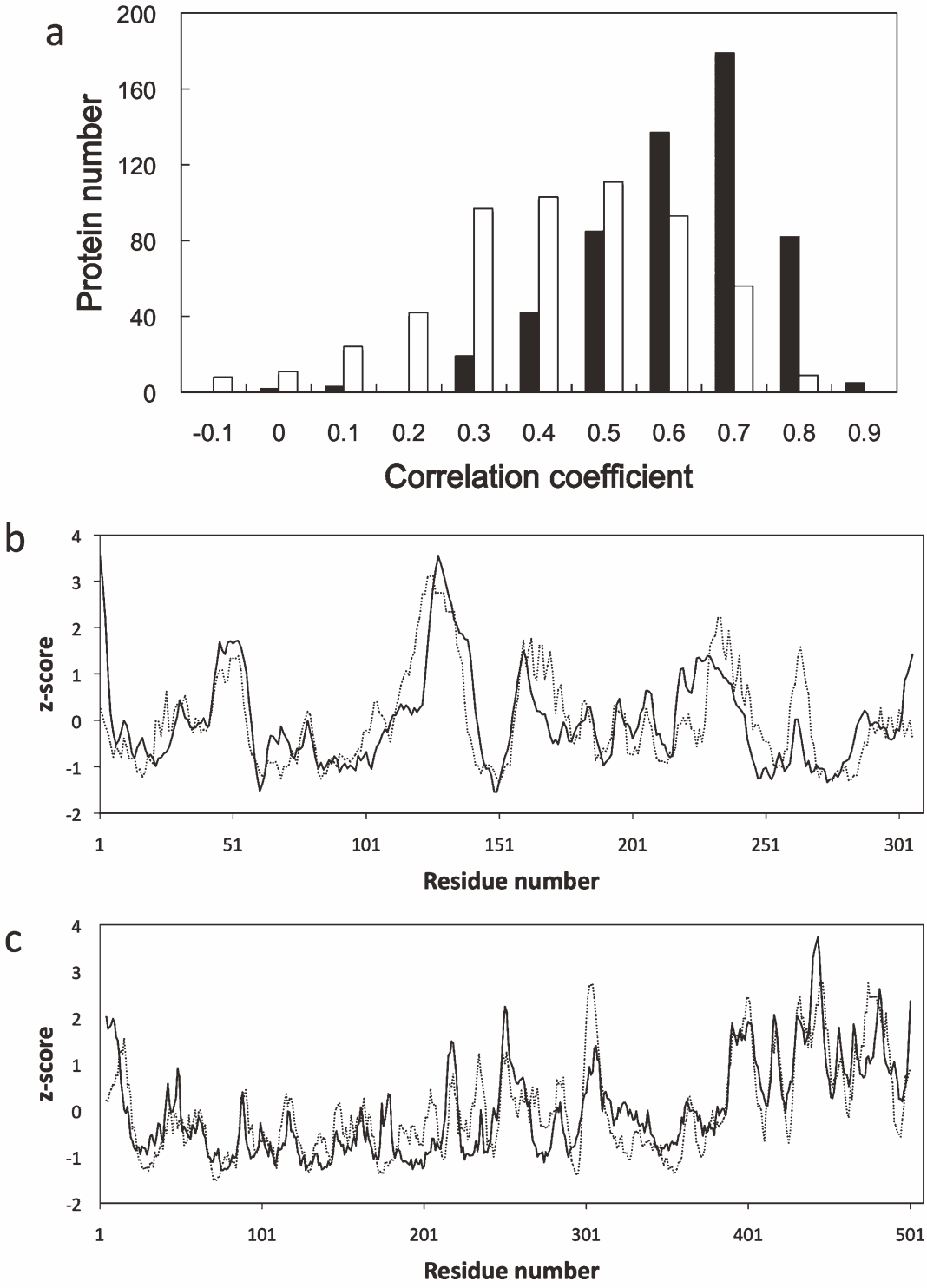

Figure 3

(**a**) The distribution of the Pearson's correlation coefficients between the sequence conservation and the *B*-factor profiles (empty bar) for the data set of 554 enzymes; also shown for comparison is the distribution of the correlation coefficients between the conservation and the rWCN profiles (solid bar). Comparison of the *B*-factor profile (black line) and the conservation profile (dotted line) for **b**, 2DLN:A and **c,** 1PZ3:A. The *B*-factors are normalized to their corresponding *z*-scores.

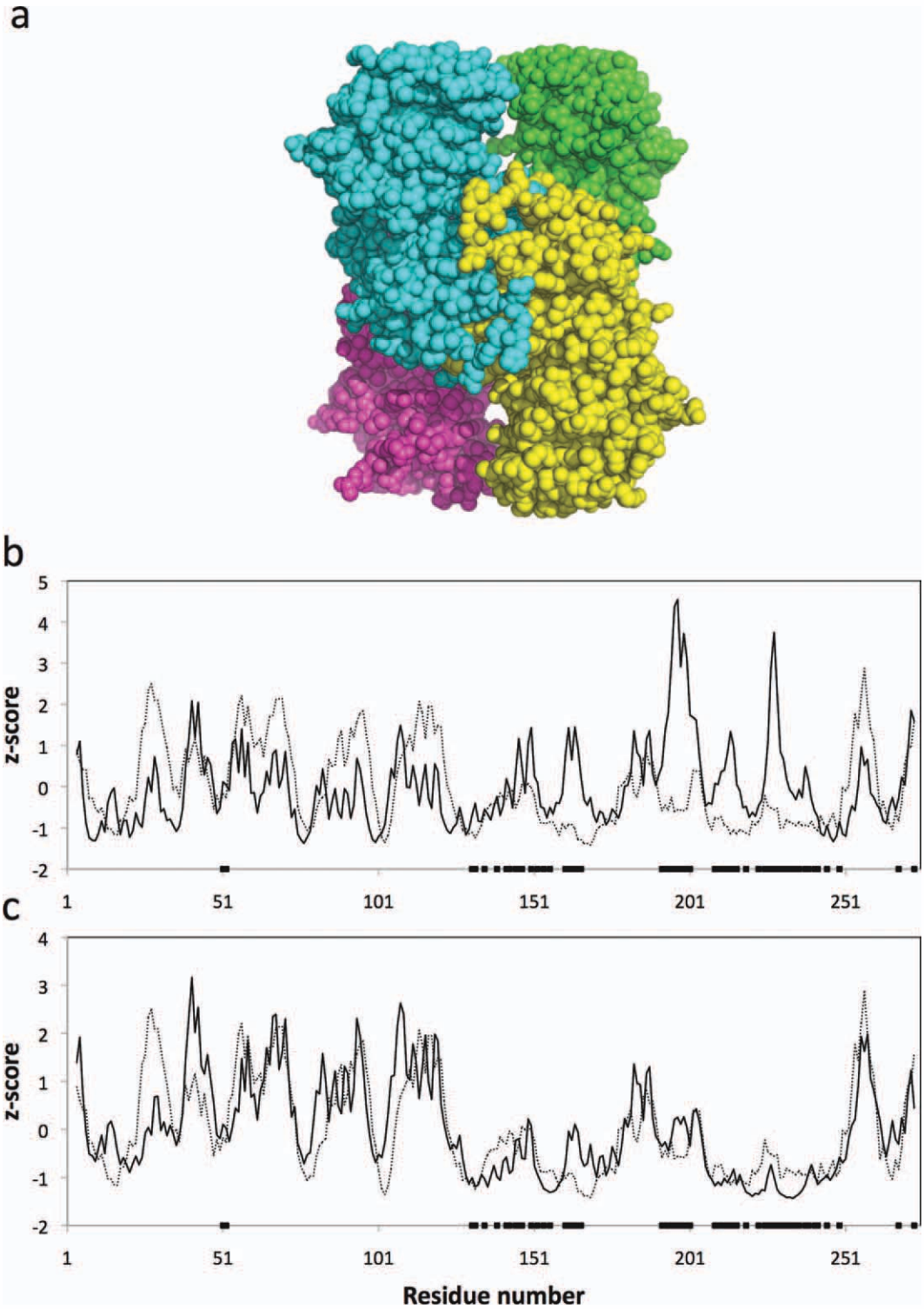

Figure 4

(**a**) The structure of dihydrodipicolinate reductase (PDB ID: 1ARZ) consists of four identical chains—each chain is rendered in different color. Comparison of the conservation profile (dotted line) and the rWCN profile (solid line) of 1ARZ:A computed using (**b**) a single chain and (**c**) 4 chains. The interface residues are marked by the black solid squares. The interface residues are determined following the approach of Porollo and Meller.[36]

The X-ray structure of 1ARZ shows that there are four subunits interacting with each other extensively to form a 16-stranded β-barrel.[19] These results suggest that the subunits may be evolutionarily coupled to the chains. Note that each subunit has two catalytic residues, and they are close to the protein interfaces. Both catalytic residues at the interface come from one subunit. They are within 3.6 Å from each other and one of them (K163) is located at the interface between two subunits.

Another example is sialyltransferase CstII (PDB ID: 1RO7), which consists of four identical chains [Fig. 5(a)]. The correlation coefficient of 1RO7:A is 0.58 when computed using only chain A [Fig. 5(b)], but it drops to 0.17 when all four chains are used [Fig. 5(c)]. The X-ray structure of 1RO7 shows that its active site (each monomer has four catalytic residues) is not at the interface between monomers of the tetramer, and that oligomerization does not play a direct role in catalysis.[20] These results suggest that the evolutionary coupling among the chains of 1RO7 may be relatively weak.

A recent study by Weng and Mintseris[21] on transient and obligate protein–protein interactions showed that the interface residues of obligate complexes tends to evolve at slower rate (i.e., more conserved), allowing them to coevolve with their interaction partners. But in transient complexes, the interface residues are less conserved, thus leaving little room for correlated mutations across the interface (i.e., there is less evolutionary coupling). These are consistent with our results—as shown in Figures 4 and 5, the interface residues of 1ARZ (marked by black solid circles) are among the more conserved residues of the protein. In contrast, the interface residues of 1RO7 (marked by black solid circles) are among the less conserved residues of the protein.

When the size of the interface between the subunits is small, we expect that the correlation coefficients will be similar regardless how we compute the rWCN profiles, i.e., with a single subunit or with all subunits. This is indeed the case, as shown in Figure 6(a). However, when the size of the interface is large, the discrepancies between correlation coefficients become large [Fig. 6(b)].

In the case of single chain protein with multiple domains, the rWCN profile of a domain can be computed in two ways: with one domain or with all of the domains (i.e., the complete chain). We found that the correlation coefficients are in general better if the rWCN profiles are computed with all domains (Fig. 7)—there are 388/478 (81%) of the domains having a higher correlation coefficient. This result suggests that the domains of a single chain protein in most cases evolutionarily coevolve with each other in our dataset, though further study is required. Two examples are shown in Figure 8. Galactose oxidase 1GOG:A has three domains according to the SCOP classification. These domains are closely packed together. The correlation coefficient of its C-terminal domain (SCOP ID: d1goga1) computed using all domains is 0.86 [Fig. 8(a)], but its correlation coefficient becomes almost 0 when using only one domain [Fig. 8(b)]. The most significant difference in correlation coefficients occurs in the loop 574–591. Inspecting the structure of 1GOG:A, we found that this loop deeply inserts into the center of a 7-bladed beta propeller of the domain d1goga3. Another example is hexokinase 1DGK:N, which has four domains according to the SCOP classification. The correlation coefficient of its mammalian type I hexokinase domain (SCOP ID: d1dgkn3) with WCN computed using all domains is 0.28 [Fig. 8(c)], but its correlation coefficient increases to 0.58 when using only its hexokinase domain [Fig. 8(d)]. The most significant difference in correlation coefficients occurs in the region from residue K549 to G593. This region has little contact with other domains—there are only three residues (L734, D895, and K899) from the other domain (SCOP ID: d1dgkn4) within 4 Å of this region.

In summary, the protein's WCN profile, together its conservation profile, do help shed some interesting new light on the evolutionary relationships among the subunits or domains. However, further study is required.

DISCUSSION

Protein evolution is under functional and structural constraints. It is known that protein function depends on a folded conformation and that homologous sequences have similar structures, i.e., they share a common three-dimensional fold. Therefore, it is expected that the three-dimensional fold will exert significant constraints on the evolvability of the residues of a protein. Our results show that the conservation profiles of protein sequences closely resemble the corresponding rWCN profiles. This is remarkable: on one hand, the WCN profile, computed directly from atomic coordinates of Cα atoms, does not use any explicit information of amino acid sequence, while, on the other hand, the sequence conservation profile is computed based only on amino acid sequences. Hence, our results suggest that the protein structure exerts such strong evolutionary constraints on the residues of the protein that its sequence conservation profile of a protein essentially reflects its structural characteristics, i.e., the rWCN profile.

Furthermore, our results provide a straightforward explanation to the recent observations that active sites residues are usually found at the regions of high packing density[15] or close to the structural centroid,[14] and that they usually have lower *B*-factors.[22] These can be understood through the quantitative relationship between the structure packing and sequence conservation.

Recently, Jernigan coworkers[23] studied on the relation between protein's sequence entropy and its contact number. It is remarkable that, for a dataset comprising 130 proteins, Jernigan-Lustig got a near perfect correlation

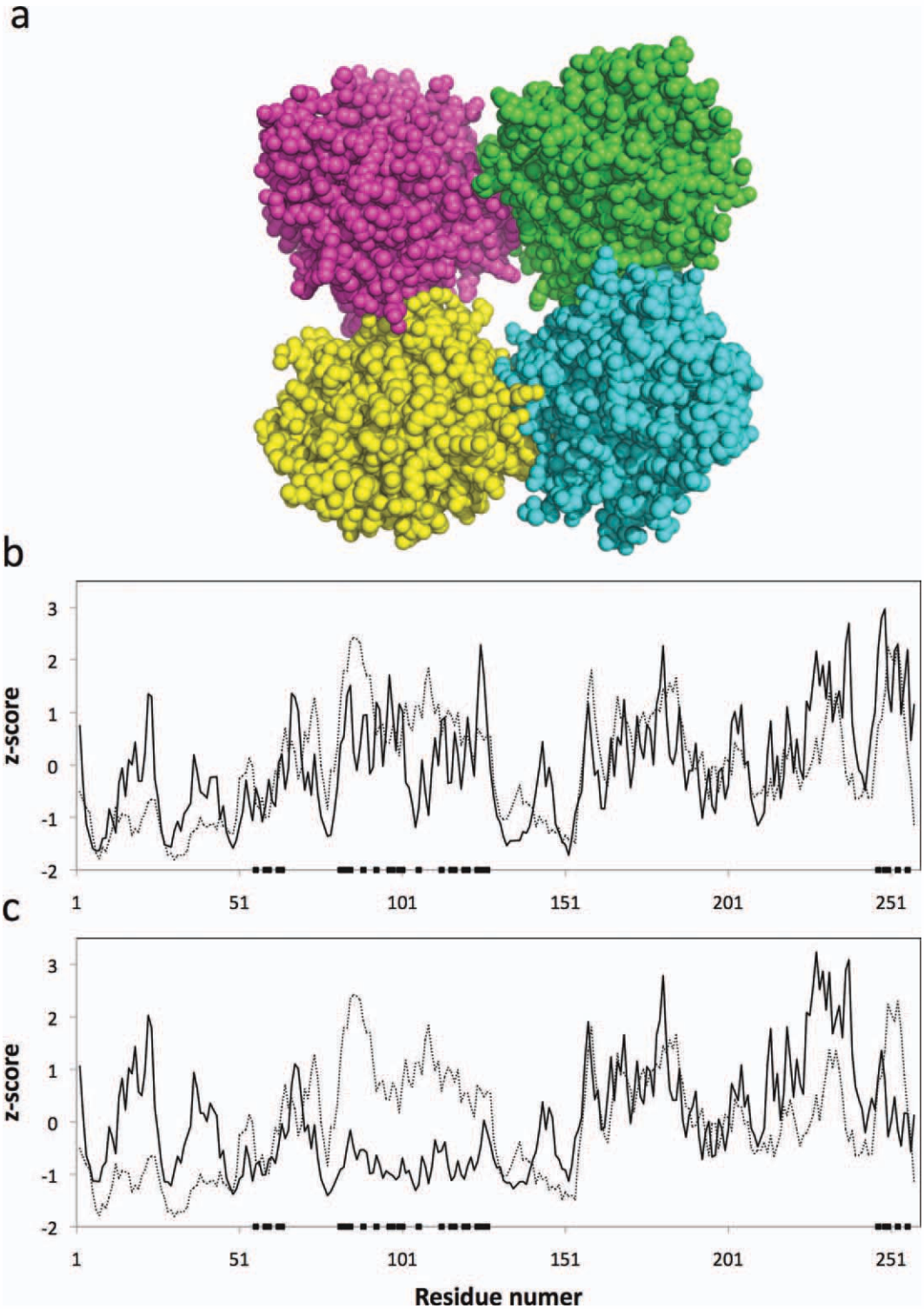

Figure 5

(**a**) The structure of sialyltransferase CstII (PDB ID: 1RO7) consists of four identical chains—each chain is rendered in different color. Comparison of the conservation profile (dotted line) and the rWCN profile (solid line) of 1RO7:A computed using (**b**) a single chain and (**c**) 4 chains. The interface residues are marked by the black solid squares.

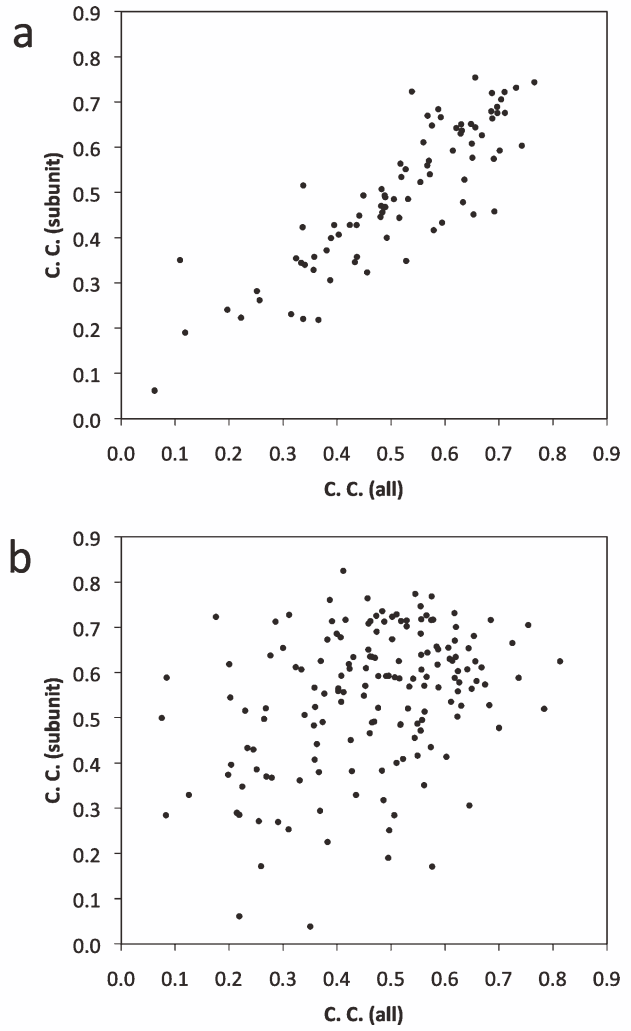

Figure 6

The label "c. c. (all)" on the *X*-axis denotes the correlation coefficient with the rWCN profile computed using all subunits, while "c. c. (subunit)" denotes those computed using the given subunit. **a,** comparison of the correlation coefficients when the ratio between the interface area and the exposed area of the complex is less than 5%; **b,** comparison of the correlation coefficients when the ration between the interface area and the exposed area of the complex is larger than 10%. The definition the interface area follows that of Porollo and Meller.[36]

coefficient between contact number and sequence entropy. However, this was not done for individual residues of a single protein but done with sequence entropy averaged over all residues of all proteins in the dataset within each density bin and with outlying values removed. They emphasized that the average procedures are necessary because "contact number for single residues of a single protein does not necessarily correlate well with the sequence conservation at the site".

We applied Jernigan–Lustig's sequence entropy method to their dataset and our dataset to compute the correlation between the sequence entropy and the contact number profiles for individual residues. In Jernigan–Lustig's work, sequence entropy is calculated from a sequence alignment set generated by BLASTP from the query sequence and packing density is the number of residue's CA atoms within a 9 Å. For the larger dataset comprising 554 proteins, Jernigan–Lustig's method yielded 1% (7/554) of proteins with a correlation coefficient > 0.5, while ours 74% (408/554). Jernigan–Lustig's method gave an average correlation coefficient 0.29, significantly lower than ours 0.57. For the smaller dataset comprising 130 proteins, Jernigan-Lustig's method yielded 2% (3/130) of proteins with a correlation coefficient > 0.5, while our method yielded 69% (90/130). Jernigan–Lustig's method gave an average correlation coefficient 0.31, compared with ours 0.55. These results show that Jernigan–Lustig's method could not establish a quantitative correlation between the protein's conservation profile and its packing density profile on the residue level.

One of the problems with Jernigan-Lustig's approach is that their sequence entropy does not properly describe the level of residue conservation. Jernigan-Lustig's sequence entropy is computed using the pairwise sequence alignment generated by BLASTP from the query sequence. Though there is no standard way to compute residue conservation, to properly address the level of residue conservation one needs to start from multiple sequence alignment.[2,3,24–26] Another problem with Jernigan-Lustig's method is the use of CN to represent packing density. As shown in recent studies,[12,27] WCN provides a more realistic description of packing density than CN. We repeated Jernigan-Lustig's calculations with WCN replacing CN for the 554 protein dataset. We

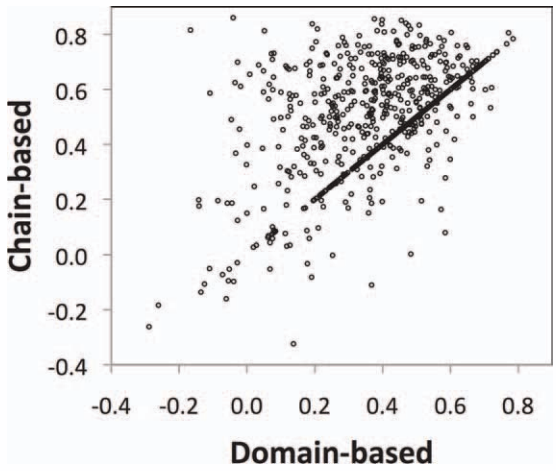

Figure 7

Comparison of the correlation coefficients of single domains with the rWCN profile computed in two different ways, i.e., given each point (x, y), x is the correlation coefficient with rWCN computed using only one domain (hence, labeled as Domain-based), while y is the that computed using the whole single chain (hence, labeled as Chain-based).

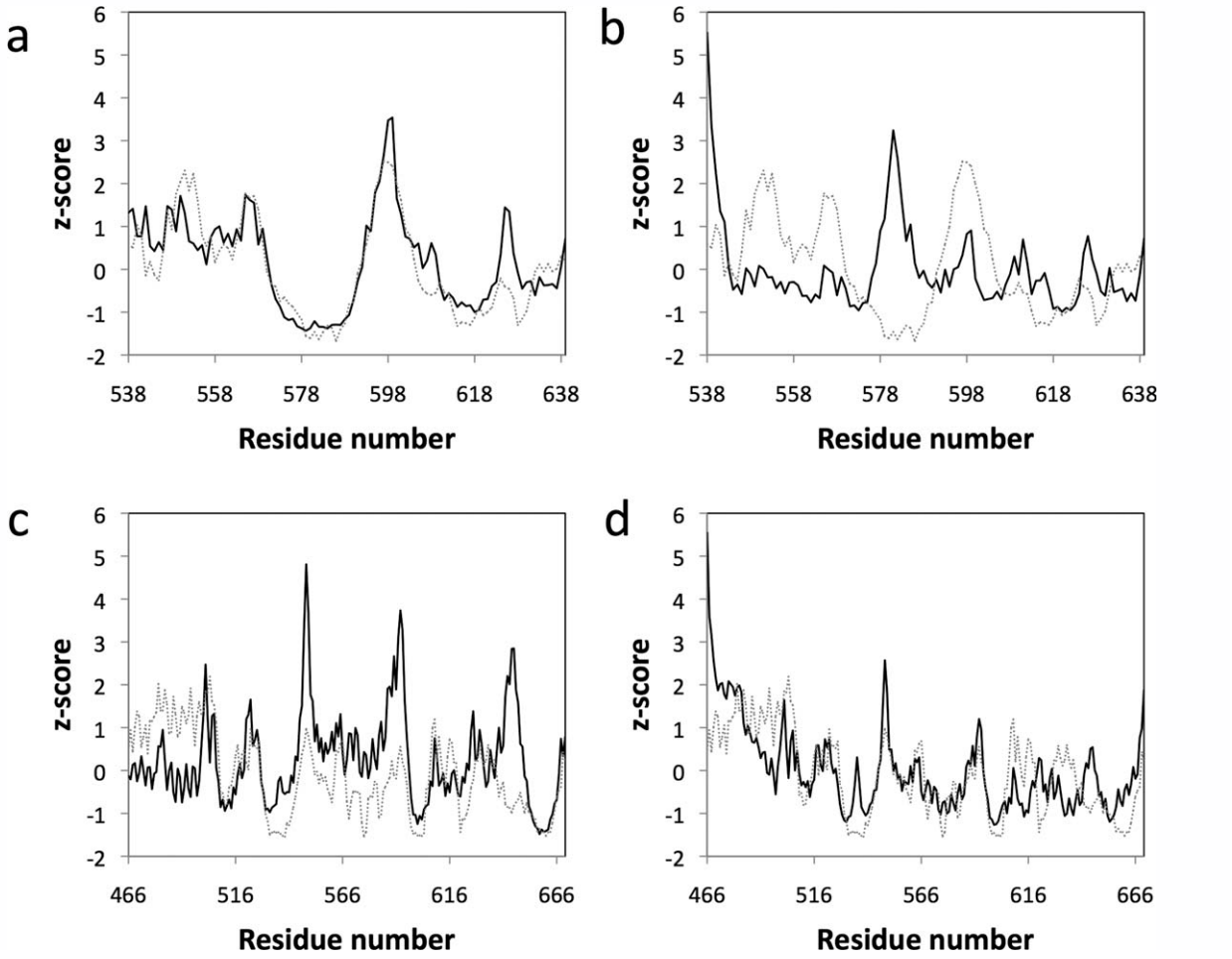

Figure 8

The galactose oxidase 1GOG:A has three domains according to the SCOP classification. Comparison of the rWCN profile (solid line) and the conservation profile (dotted line) of its C-terminal domain with WCN computed using **a**, all domains (correlation coefficient $\sim$ 0.86) and **b**, only that domain (correlation coefficient $\sim$ 0). The hexokinase 1DGK:N has 4 domains according to the SCOP classification. Comparison of the profiles of its mammalian type I hexokinase domain with WCN computed using **c**, all domains (correlation coefficient $\sim$ 0.28) and **d**, only that domain (correlation coefficient $\sim$ 0.58).

obtained a result of 147/554 of proteins having a correlation coefficient $>$ 0.5. This result is much better than Jernigan-Lustig's original result; however, it is still significantly lower than our result: 408/554.

An interesting question is whether sequence diversity will affect the correlation coefficients. To check the effect on sequence diversity on the correlation, we compare two examples: phosphofructokinase (1KZH:A) and deoxyhypusine synthase (1ROZ:A), each computed with two sets of alignment sequences. The first set is generated from our standard procedure, while the second set consists of highly homologous sequences with sequence identity $>$ 90%. In both examples, the correlation coefficients decrease: 1KZH:A decreases from 0.81 to 0.38 and 1ROZ:A from 0.79 to 0.28.

In conclusion, our main result is the establishment of a close correlation between a protein's conservation profile and its packing density profile on the level of individual proteins. We believe our findings will inspire further study on issues like why a packing density profile can be so similar to a sequence conservation profile residue by residue or how much evolutionary information is hidden in a single structure, and will open a new way to study the evolutionary coupling among the subunits or domains of proteins.

METHODS

Protein packing is described using weighted contact number.[12] The weighted contact number (WCN) of an atom is the sum of the inverse square of the distances between it and other atoms. It is different from the usual contact number (CN), which is simply the number of

neighboring atoms within a certain cutoff distance.[10] Formally, we define the WCN as

$$w_i = \sum_{j \neq i}^{N} \frac{1}{r_{ij}^2} \quad (1)$$

where r_{ij} is the distance between Cα atoms of residue i and j and N is the number of residues. The WCN profile is normalized to the corresponding z-scores as: $(w_i - \overline{w})/\sigma_w$, where $\overline{w}$ and σ_w are the mean and the standard deviation of WCN of the sequence, respectively. The reciprocal of the WCN profile is used to compare with the sequence conservation profile. Note that the WCN is computed using only Cα atoms.

Sequence conservation profiles

The sequence-specific conservation scores are computed following the protocol of CONSURF.[28] The advantage of this method is that it takes into account the stochastic processes underlying the evolution process and that it relies on the phylogency of the sequences. This method goes as follows: first, a given sequence's homologous sequences are retrieved using PSI-BLAST[29] from the SwissProt database.[30] Then, the redundant and related sequences are removed using CD-HIT.[31] A multiple sequence alignment (MSA) of the homologous sequences is then performed using MUSCLE.[32] Next, the MSA is then used to build a phylogenetic tree using Rate4Site[33] based on the neighbor joining algorithm.[34] Finally, the position-specific conservation scores are computed using the empirical Bayesian method[33] and are smoothed over a 5-resdiue window. The conservation scores are normalized to their corresponding z-score so that the average is zero and the standard deviation is one.

Dataset

We selected 554 enzymes from Catalytic Site Atlas 2.2.11[35] under the criteria that they had less than 25% pairwise sequence identity and their X-ray structures had less than five missing residues. These 554 proteins are listed in the Supporting Information Table S1.

ACKNOWLEDGMENTS

The authors are also grateful to the Center for Bioinformatics Research at National Chiao Tung University for both hardware and software supports.

REFERENCES

1. Creighton TE. Proteins: structures and molecular properties. New York: W. H. Freeman and Company; 1993.
2. Yang Z. Computational molecular evolution. New York: Oxford University Press; 2006.
3. Valdar WS. Scoring residue conservation. Proteins 2002;48:227–241.
4. Pupko T, Bell RE, Mayrose I, Glaser F, Ben-Tal N. Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. Bioinformatics 2002;18 (Suppl 1):S71–S77.
5. Johansson F, Toh H. A comparative study of conservation and variation scores. BMC Bioinformatics 2010;11:388.
6. Chivian D, Baker D. Homology modeling using parametric alignment ensemble generation with consensus and energy-based model selection. Nucleic Acids Res 2006;34:e112.
7. Fiser A, Sali A. Modeller: generation and refinement of homology-based protein structure models. Methods Enzymol 2003;374:461–491.
8. Chen CC, Hwang JK, Yang JM. (PS)2: protein structure prediction server. Nucleic Acids Res 2006;34(Web Server issue):W152–W157.
9. Shih CH, Huang SW, Yen SC, Lai YL, Yu SH, Hwang JK. A simple way to compute protein dynamics without a mechanical model. Proteins 2007;68:34–38.
10. Halle B. Flexibility and packing in proteins. Proc Natl Acad Sci USA 2002;99:1274–1279.
11. Lu CH, Huang SW, Lai YL, Lin CP, Shih CH, Huang CC, Hsu WL, Hwang JK. On the relationship between the protein structure and protein dynamics. Proteins 2008;72:625–634.
12. Lin CP, Huang SW, Lai YL, Yen SC, Shih CH, Lu CH, Huang CC, Hwang JK. Deriving protein dynamical properties from weighted protein contact number. Proteins 2008;72:929–935.
13. Amitai G, Shemesh A, Sitbon E, Shklar M, Netanely D, Venger I, Pietrokovski S. Network analysis of protein structures identifies functional residues. J Mol Biol 2004;344:1135–1146.
14. Ben-Shimon A, Eisenstein M. Looking at enzymes from the inside out: the proximity of catalytic residues to the molecular centroid can be used for detection of active sites and enzyme-ligand interfaces. J Mol Biol 2005;351:309–326.
15. Huang SW, Yu SH, Shih CH, Guan HW, Huang TT, Hwang JK. On the relationship between catalytic residues and their protein contact number. Curr Protein Pept Sci 2011;12:574.
16. Branden C, Tooze J. Introduction to protein structure. New York: Garlnd Science; 1999.
17. Schultz GE, Schirmer RH. Pinciples of protein structure. New York: Springer; 1979.
18. Maguid S, Fernandez-Alberti S, Pariso G, Echave J. Evolutionary conservation of protein backbone flexibility. J Mol Evol 2006;63:448–457.
19. Scapin G, Reddy SG, Zheng R, Blanchard JS. Three-dimensional structure of *Escherichia coli* dihydrodipicolinate reductase in complex with NADH and the inhibitor 2,6-pyridinedicarboxylate. Biochemistry 1997;36:15081–15088.
20. Chiu CP, Watts AG, Lairson LL, Gilbert M, Lim D, Wakarchuk WW, Withers SG, Strynadka NC. Structural analysis of the sialyltransferase CstII from Campylobacter jejuni in complex with a substrate analog. Nat Struct Mol Biol 2004;11:163–170.
21. Mintseris J, Weng Z. Structure, function, and evolution of transient and obligate protein-protein interactions. Proc Natl Acad Sci USA 2005;102:10930–10935.
22. Yuan Z, Zhao J, Wang ZX. Flexibility analysis of enzyme active sites by crystallographic temperature factors. Protein Eng 2003;16:109–114.
23. Liao H, Yeh W, Chiang D, Jernigan RL, Lustig B. Protein sequence entropy is closely related to packing density and hydrophobicity. Protein Eng Des Sel 2005;18:59–64.
24. Park J, Karplus K, Barrett C, Hughey R, Haussler D, Hubbard T, Chothia C. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. J Mol Biol 1998;284:1201–1210.
25. Armon A, Graur D, Ben-Tal N. ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. J Mol Biol 2001;307:447–463.

26. Capra JA, Singh M. Predicting functionally important residues from sequence conservation. Bioinformatics 2007;23:1875–1882.
27. Yang L, Song G, Jernigan RL. Protein elastic network models and the ranges of cooperativity. Proc Natl Acad Sci USA 2009;106:12347–12352.
28. Ashkenazy H, Erez E, Martz E, Pupko T, Ben-Tal N. ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. Nucleic Acids Res;38(Web Server issue):W529–533.
29. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.
30. Boutet E, Lieberherr D, Tognolli M, Schneider M, Bairoch A. UniProtKB/Swiss-Prot. Methods Mol Biol 2007;406:89–112.
31. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 2006;22:1658–1659.
32. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 2004;32:1792–1797.
33. Mayrose I, Graur D, Ben-Tal N, Pupko T. Comparison of site-specific rate-inference methods for protein sequences: empirical Bayesian methods are superior. Mol Biol Evol 2004;21:1781–1791.
34. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 1987;4:406–425.
35. Porter CT, Bartlett GJ, Thornton JM. The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. Nucleic Acids Res 2004;32(Database issue):D129–D133.
36. Porollo A, Meller J. Prediction-based fingerprints of protein-protein interactions. Proteins 2007;66:630–645.