On the relationship between the sequence conservation and the packing density profiles of the protein complexes

ARTICLE *in* PROTEINS STRUCTURE FUNCTION AND BIOINFORMATICS · JULY 2013

Impact Factor: 2.63 · DOI: 10.1002/prot.24268 · Source: PubMed

Jimmy Chang
National Chiao Tung University

Yu-Wen Huang
National Chiao Tung University

PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

On the relationship between the sequence conservation and the packing density profiles of the protein complexes

Chih-Min Chang, Yu-Wen Huang, Chien-Hua Shih, and Jenn-Kang Hwang*

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

ABSTRACT

We have recently showed that the weighted contact number profiles (or the packing density profiles) of proteins are well correlated with those of the corresponding sequence conservation profiles. The results suggest that a protein structure may contain sufficient information about sequence conservation comparable to that derived from multiple homologous sequences. However, there are ambiguities concerning how to compute the packing density of the subunit of a protein complex. For the subunits of a complex, there are different ways to compute its packing density – one including the packing contributions of the other subunits and the other one excluding their contributions. Here we selected two sets of enzyme complexes. Set A contains complexes with the active sites comprising residues from multiple subunits, while set B contains those with the active sites residing on single subunits. In Set A, if the packing density profile of a subunit is computed considering the contributions of the other subunits of the complex, it will agree better with the sequence conservation profile. But in Set B the situations are reversed. The results may be due to the stronger functional and structural constraints on the evolution processes on the complexes of Set A than those of Set B to maintain the enzymatic functions of the complexes. The comparison of the packing density and the sequence conservation profiles may provide a simple yet potentially useful way to understanding the structural and evolutionary couplings between the subunits of protein complexes.

Proteins 2013; 00:000–000.

Key words: protein structure; sequence conservation; weighted contact number.

INTRODUCTION

Protein evolution is under functional and structural constraints. In addition, folding stability constraint, and possibly folding kinetic accessibility may also play a role in the evolutionary history of proteins. On the sequence level, the degree of amino acid conservation reflects such constraints. Since protein function requires properly folded conformations, it is expected that such evolutionary constraints will also be reflected on the structural level. Structural biologists have long observed that the residues in the protein core regions are usually more conserved than the residues on the surface.[1,2] A recent study took up statistical analysis of this essentially qualitative proposal for a dataset of 130 proteins.[3] It reported an excellent correlation (0.997) between protein's sequence entropies and their contact numbers. However, this high correlation comes from averaging the sequence entropy over the residues within each contact-number bin. Without averaging the sequence entropy over all residues, the average correlation coefficient drops to 0.31.[4] The statistical analysis confirms the relationship but hardly makes the originally descriptive proposal any more useful in practice.

To better describe protein packing, Hwang *et al.* developed the weighted contact number model[4] (see Methods), which takes into account the distance dependence of the packing contributions of the contact atoms, providing a more realistic description of packing density than the usual contact number model. This weighted contact number model has been implemented to elastic network models[5] to give a better prediction of the atomic thermal fluctuations. It has also been successfully

Additional Supporting Information may be found in the online version of this article.
Grant sponsor: Academic Summit Program of National Science; Grant number: 100-2745-B-009-001-ASP; Grant sponsor: "Center of Bioinformatics Research of Aiming for the Top University Program" of the National Chiao Tung University and Ministry of Education, Taiwan, ROC
*Correspondence to: Jenn-Kang Hwang, Institute of Bioinformatics and Systems Biology, National Chiao Tung University, HsinChu 30050, Taiwan, ROC. E-mail: jkhwang@faculty.nctu.edu.tw
Received 18 November 2012; Revised 29 January 2013; Accepted 1 February 2013
Published online 14 February 2013 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/prot.24268

applied to the prediction of the catalytic residues[6] and functionalities.[7] Recently, Hwang *et al.*[4] compared the profiles of the weighted contact numbers and those of the rates of evolution of amino acid sites for a much larger dataset of 554 proteins, and obtained an average correlation coefficient of 0.57 between them. It is worth mentioning that this correlation was obtained on the protein basis (i.e., without averaging over the residues) – 74% (408/554) of the proteins have a correlation coefficient $\geq$ 0.5. To appreciate the similarities between the weighted contact number and sequence profiles, we present two examples in Figure 1. The implications that one can extract information concerning sequence conservation from protein structures are intriguing. First, they promise interesting applications such as to identify the conserved residues directly from protein structures,[8] which is especially important when the proteins are of novel structures and do not have any known homologous sequences. They also stimulate further questions concerning the relationship between protein structure and sequence conservation. For example, are there other structural properties besides packing density that exhibit a similar quantitative correlation with sequence conservation? What are the factors underlying the correlation between packing density and sequence conservation?

An unsettled question arising from Hwang *et al.*'s study[4] is the ambiguous ways of computing the packing density of the subunits of a protein complex. When computing the packing density of a subunit, one can either consider or ignore the packing contributions of the other subunits of a complex. Hwang *et al.*[4] found that either one will work for some cases, but it is not clear why they are so. In this report, we selected two sets of complexes. Set A contains enzyme complexes whose active sites comprise catalytic residues from multiple subunits; Set B contains enzyme complexes whose active sites are located in individual subunits. We found that only when the subunit of Set A is treated as an integral part of the complex will its packing density profile better agree with the corresponding sequence conservation profile. This situation is revered in Set B. Our results indicate that couplings between the subunits may be responsible for the degree of correlation between the packing density and the corresponding sequence conservation profiles of the complexes. In addition, our results suggest a novel way of looking at the couplings between subunits of a complex through comparison of the structural and sequence conservation profiles.

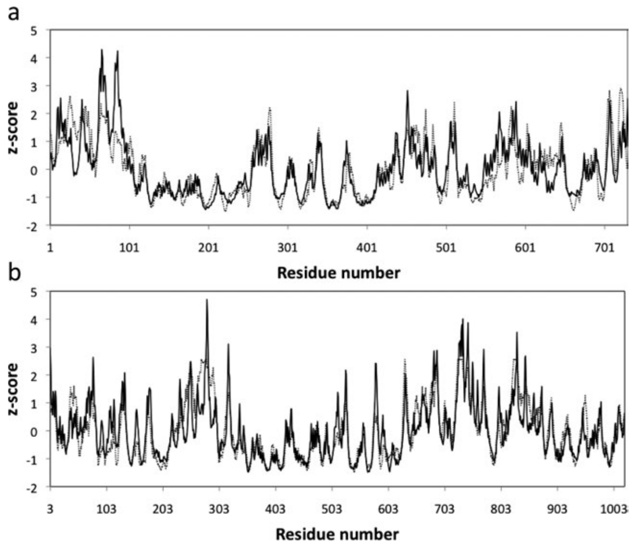

Figure 1

Two examples of the weighted contact number and the sequence conservation profiles: (**a**) ornithine decarboxylase (PDB ID: 1ORD:A) and (**b**) beta-galactosidase (PDB ID: 1BGL:A). The weighted contact number profile is in solid line and the sequence conservation profile in dotted line. Both profiles are normalized to z-sores. The correlation coefficients of (**a**) and (**b**) are 0.77 and 0.81, respectively.

METHODS

Protein packing density profiles

The weighted contact number (WCN) model[9] is a coarse-grained description of the packing density of a protein, considering only Cα atoms of the amino acid residues. This is similar to the usual contact number (CN) model.[10] However, the WCN takes into account of the distance-dependence of the contributions of the neighboring residues, while the CN does not. The WCN model, being a more realistic description of the protein's packing density than the CN model, has since found many interesting applications.[4–7]

Formally, the WCN of residue i is defined as $w_i = \sum_{j\neq i}^{N} 1/r_{ij}^2$, where r_{ij} is the distance between Cα atoms of residue i and j and N is the number of residues of the protein. For a protein, a series of the $w_i^{'}$ s, i.e. $(w_1,w_2,\ldots,w_N)$, is referred to as its WCN profile. However, for convenience, the reciprocal WCN (rWCN) will be used in the text. Define $\omega_i = w_i^{-1}$, the series $(\omega_1,\omega_2,\ldots,\omega_N)$ is referred to as the rWCN profile. We will use the packing density profile and the rWCN profile interchangeably. For easy comparison, the rWCN profile is normalized to the corresponding z-scores: $z_i = (\omega_i - \bar{\omega})/\sigma_\omega$, where $\bar{\omega}$ and σ_ω are the mean and the standard deviation of rWCN, respectively.

Sequence conservation profiles

The sequence-specific conservation scores are computed using CONSURF,[11] which is based on the phylogeny of the sequences. This method has the advantage of taking into account the stochastic nature of the evolutionary process. The basic steps of the method are : (1) The query sequence's homologous sequences are retrieved from the SwissProt database[12] using PSI-BLAST.[13] (2) The redundant are removed through CD-HIT[14] with a sequence identity cutoff of 95%, following the default

Table I
The Protein Complexes of Set A and Set B

Set A									
1OHH: A-H	1CA0: A-D, F-I	1DQR: A-B	1CC1: L-S	1NDO: A-F	1C3C: A-D	1JEN: A-D	1DW9: A-J	1FUQ: A-D	1A30: A-C
1AUW: A-D	1APYA-D	1HIV: A-B	1LYA: A-D	1Q6L: A-B	1DQA: A-D	1FUG: A-B	1DJO: A-B	1DTW: A-D	1MKA: A-B
1W27: A-D	1MHL:A-D	1E7L: A-B	1A05: A-B	1XS1: A-C	1JNR: A-D	1GET: A-B	1BD0: A-B	1JXA: B-C	1N2C: A-H
1T4C: A-B	1UJN:A-B	1XNY: A-F	2AG0: A-B	1B66: A-F	1F80: A-F	1PYA: A-F	1BJP: A-F	2FQQ: A-D	2AHJ: A-D
1D8D: ABP	1ECM:A-B	1NI4: A-D	1C4T: A-C	1EHK: A-C					
Set B									
1NIR: A-B	1E6E: A-D	1B65: A-F	1DIZ: A-B	2BBK: HJML	2DW7: A-P	1RO7: A-D	1OK4: A-J	1QHF: A-B	1HFE: LSTM
1OFG: A-F	1ABR: A-B	1KYW: ACF	1IMA: A-B	1BGL: A-H	1AFR: A-F	1R44: A-F	1E2T: A-H	1Z9H: A-D	13PK: A-D
1EZ1: A-B	1ODT: CH	1ITQ: A-B	1OJ4: A-B	1I9A: A-B	1JKM: A-B	1BXR: A-H	1ALK: A-B	1M54: A-F	1FWK: A-D
1JOF: A-H	1DDJ: A-D	1HPL: A-B	1XTC: A-H	1AVQ: A-C	1PZ3: A-B	1F8R: A-D	1R30: A-B	1AQL: A-B	1KDG: A-B
1APX: A-D	1OS7: A-D	1A4L: A-D	1PYM: A-B	1B5T: A-C	1B5Q: A-C	1D1Q: A-B	1YVE: I-L	1AQ0: A-B	1KFU: LS
1NDI: A-B	1EEJ: A-B	1L1L: A-D	1BU7: A-B	1CNS: A-B	2PFL: A-B	1SMN: A-B	1DHF: A-B	2F61: A-B	1D4C: A-D

value of CONSURF. (3) The program MUSCLE[15] is used for multiple sequence alignment (MSA) of the homologous sequences. (4) A phylogenetic tree is built from the MSA using Rate4Site.[16] (5) The position-specific conservation scores are computed using the empirical Bayesian method.[16] (6) The conservation scores are smoothed through averaging over a 5-resdiue window. (7) For the sake of the comparison, the conservation scores are normalized to the corresponding *z*-scores in such a way that the average conservation score is zero and the standard deviation is one. The series of the conservation scores of a sequence is referred to as its conservation profile. It is noted that, in the conservation profile of a protein, the residue of a lower conservation score is more conserved than that of a higher conservation score.

Interface regions between subunits of complexes

The residue of a complex is defined as an interface residue if the change in its accessible surface area (ASA) is greater than 5 Å^2 and the change in its relative solvent accessibility (RSA) is greater than 4% when the complex structure is formed from its isolated subunits.[17] The ASA of a residue is calculated with DSSP.[18] The RSA of a residue of a protein structure is defined as the ratio of its ASA to the maximal ASA of the isolated residue of an identical amino acid type.

DATASETS

Set A

Set A consists of 45 enzyme complexes, among which are 27 homomers and 18 heteromers. There are a total of 61 distinct subunits. The active sites of the enzyme complexes comprise catalytic residues from multiple subunits. The annotations of the catalytic residues of the enzymes are taken from Catalytic Site Atlas 2.2.11.[19] The pair-wise sequence identities of the subunits are < 30%. There may be multiple active sites in a single enzyme complex. Table I lists the PDB IDs of the complexes together with their respective subunits. The detailed information about the subunits of the complexes of Set A is in Supporting Information Table S1.

Set B

Set B consists of 60 enzyme complexes, among which are 52 homomers and 8 heteromers. There are 60 distinct subunits in Set B. The active sites of the enzymes of Set B, unlike that of Set A, are not located at the interfaces and comprise catalytic residues from single subunits. The annotations of the catalytic residues of the enzymes are taken from Catalytic Site Atlas 2.2.11. The pair-wise sequence identities of the subunits are < 30%. Table I lists the PDB IDs of the complexes together with their respective subunits. The pair-wise sequence identities of the subunits are < 30%. The detailed information about the subunits of the complexes of Set B is in Supporting Information Table S2.

RESULTS

Comparison between the sequence conservation and the packing density profiles of the subunits of Set A

Set A contains 45 enzyme complexes whose active sites are located at the interfaces and comprise catalytic residues from multiple subunits of the complexes. The enzymes are functional in their complex form since only the complex contains a complete active site. Two types of rWCN profiles (or the weighted contact number profile, see Methods) of the subunit are computed: one ignores and the other one considers the packing contributions of the other subunits of the complex. For convenience, we will refer to the first one as the rWCN type I profile (or simply the rWCN I profile) and the second one the rWCN type II profile (or the rWCN II profile). We compare the sequence conservation profiles with both types of rWCN profiles, respectively. We compare in Figure 2 the distributions of the Pearson's correlation coefficients

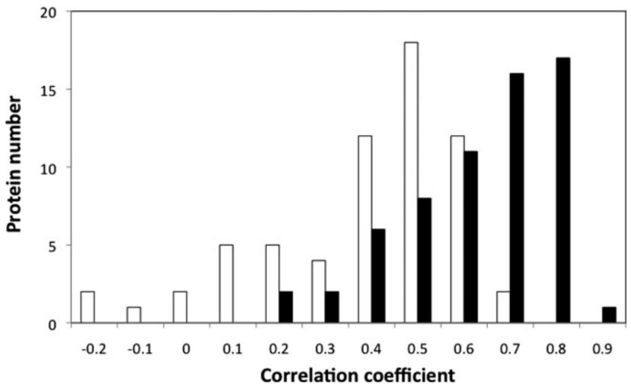

Figure 2

The distributions of the correlation coefficients between the sequence conservation profiles and the rWCN I (empty bars) and the rWCN II (solid bars) profiles, respectively, for the subunits of Set A. The sequence conservation profiles have an average Pearson's correlation of 0.34 with the rWCN I profile and 0.59 with rWCN II profile, respectively.

between the sequence conservation profile and the two types of the rWCN profiles. The average correlation coefficient of the rWCN I profile is 0.34, and that of the rWCN II profile is 0.59. In addition, 61 out of 61 subunits have better correlation coefficients when their rWCN profiles are computed including the packing contributions of the other subunits of the complexes (see Fig. 3). We will discuss two specific cases of Set A in more details in the following section.

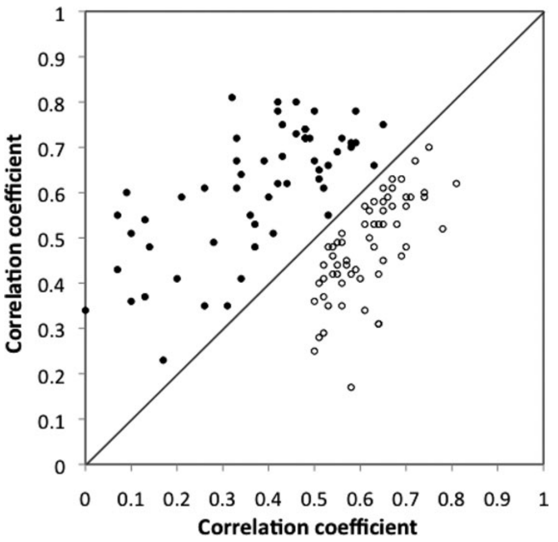

Figure 3

The scatter plot of the distributions of the correlation coefficients between the sequence conservation and the rWCN profiles for the subunits of Set A (filled circles) and the Set B (empty circles). Each subunit is represented by a point whose *x* coordinate is the correlation coefficient of an isolated subunit and *y* coordinate is that of a subunit in the presence of other subunits of a complex.

Benzaldehyde lyase

Benzaldehyde lyase[20] is a homotetramer, each subunit having 563 amino acid residues. The X-ray structure of the complex is shown in Figure 4(a). Four identical subunits form two tight dimers, each dimer containing two identical active centers at its interface. Each active site comprises H29, E50, and Q113 from one subunit and G419 from the other subunit. We compared the sequence profile of one of the subunits with the corresponding rWCN I profile [Fig. 4(b)] and the rWCN II profile [Fig. 4(c)]. The correlation coefficient is 0.48 in Figure 4(b) and 0.76 in Figure 4(c). Note that both the sequence conservation profile and the rWCN profile are normalized to the *z*-scores (see Methods). For the sequence conservation profile, the lower the score the more conserved

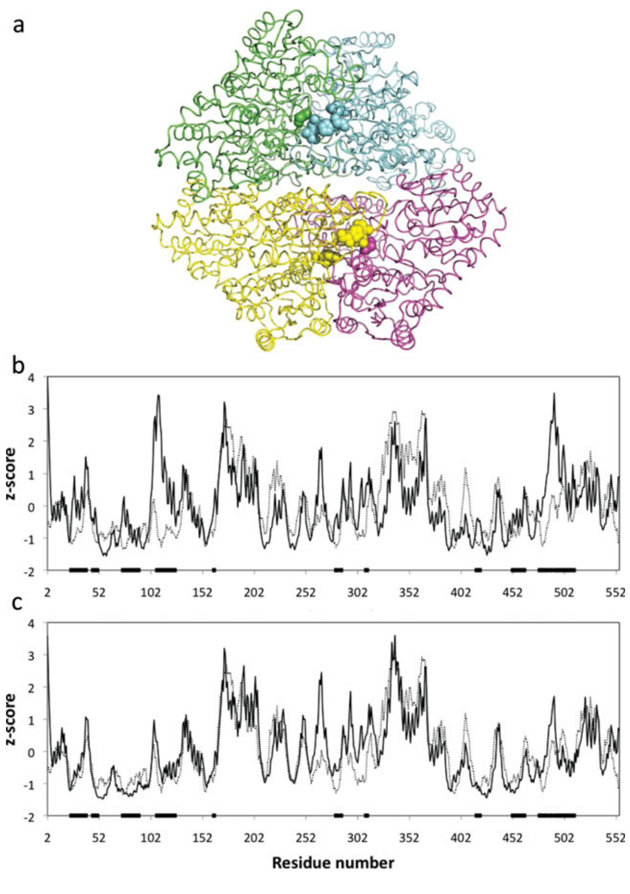

Figure 4

(**a**) The X-ray structure of benzaldehyde lyase (PDB ID: 2AG0) consists of four identical subunits forming two tight dimers (colored cyan and green, and magenta and yellow, respectively). Each tight dimer contains two active sites in the dimer interface. The active site residues are drawn in CPK model. The sequence conservation profile of the subunit is compared with its (**b**) its rWCN I profile and (**c**) rWCN II profile, respectively. The sequence conservation profile and the rWCN profile are drawn in dotted lines and solid lines, respectively. The correlation coefficients of (**b**) and (**c**) are 0.48 and 0.76, respectively. The residues in the interface regions are indicated by black thick line on the horizontal axis.

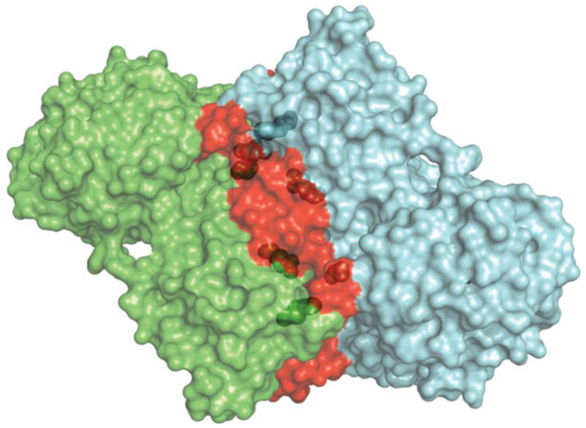

Figure 5

The surface model of the dimeric structure of benzaldehyde lyase. Its subunits are colored cyan and green, respectively. The interface residues, i.e., residues 107–125 and 464–512, are colored red. The catalytic residues are drawn in CPK model. The molecular surface is rendered in a semitransparent way so that the buried catalytic residues can be visible.

the amino acid is, and for the rWCN profile, the lower the score the more packed the amino acid residue is. We noted that in Figure 4(b) the large deviations between the profiles in Figure 4(b) usually (though not always) occur in the interface regions (for example, residues 107-125 and 464-512). Figure 5 shows that active-site residues of benzaldehyde lyase, like other complexes of Set A, are close to the interface regions.

Alanine racemase

Alanine racemase is a homodimer,[21] whose active site comprises Y265 and C311 from subunit A, and K39 and R136 from subunit B. Its dimeric structure is shown in Figure 6(a). The sequence conservation profile of the subunit is compared with the corresponding rWCN I profile [Fig. 6(b)] and the rWCN II profile [Fig. 6(c)], respectively. The average correlation coefficients in Figure 6(b,c) are 0.50 and 0.78, respectively. In Figure 6(b) the most pronounced differences between the profiles occur in the 251–294 and 310–316 regions, at which the catalytic residues Y265 and C311 are located.

Comparison between the sequence conservation and the packing density profiles of the subunits of Set B

Set B contains 60 enzyme complexes whose active sites comprise catalytic residues from single subunits and are far way from the interfaces of the complexes. The subunits of Set B contain complete active sites and are different from those of Set A, which contain only partial active sites.

We compare in Figure 7 the distributions of the Pearson's correlation coefficients between the sequence conservation profile and the rWCN I and II profiles,

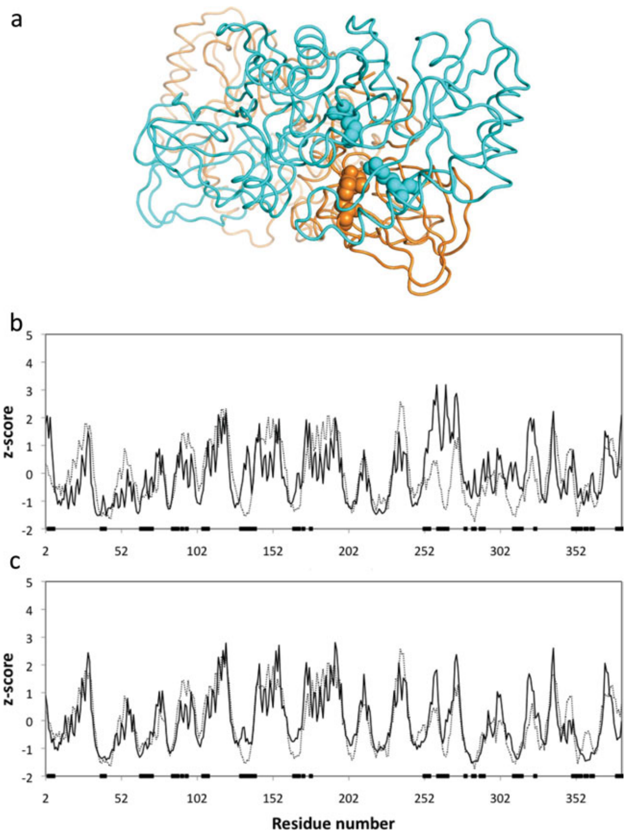

Figure 6

(**a**) The homomeric structure of alanine racemase (PDB ID: 1BD0) consists of two identical subunits (colored in orange and cyan). The active site residues are drawn in CPK model. The sequence conservation profile of the subunit is compared with its (**b**) its rWCN I profile and (**c**) rWCN II profile, respectively. The correlation coefficients of (**b**) and (**c**) are 0.50 and 0.78, respectively.

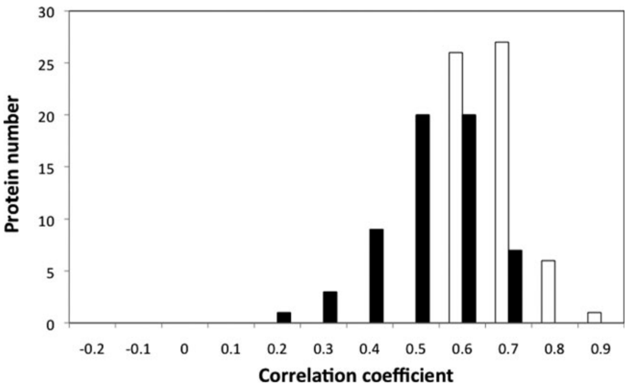

Figure 7

The distributions of the correlation coefficients between the sequence conservation profiles and the rWCN I (empty bars) and the rWCN II (solid bars) profiles, respectively, for the subunits of Set B. The sequence conservation profiles have an average Pearson's correlation of 0.62 with the rWCN I profile and 0.48 with rWCN II profile, respectively.

respectively. The average correlation coefficient of the rWCN I profiles is 0.62, and that of the rWCN II profiles is 0.48. These are opposite to what we got from Set A. The results are in fact much stronger than what the average correlation coefficients suggest. As shown in Figure 3, all subunits of Set A have higher correlation coefficients for the rWCN II profiles and all subunits of Set B have higher correlation coefficients for the rWCN I profiles.

These results may be rationalized as follows. The degree of amino acid conservation reflects the structural or functional constraints on the evolutionary process by which a protein evolves into a functional unit. For an enzyme complex (such as the complex of Set A) whose active site comprises residues of multiple subunits with each subunit containing only part of the active site, the sequence conservation profile of the subunit is expected to reflect to certain degree the couplings, structural or functional, with the other subunits of the complex. Therefore, the sequence conservation profile of the subunit of Set A has a poorer correlation with the rWCN I profile, which does not carry structural information (i.e., in terms of packing contributions) of the other subunits of the complex, than the rWCN II profile, which does. In Set B, the subunit is in principle a functional unit on its own, each containing a complete active site, and is expected to be less functionally or structurally coupled with other subunits than that of Set A. For Set A, the average correlation between the rWCN profiles computed with and without incorporating other subunits is 0.66, and for Set B the average correlation is 0.76. Both are higher than the correlation coefficients between the conservation and the rWCN profiles.

Two examples will be presented in the next sections in more details.

Phosphoglycerate kinase

The X-ray structure phosphoglycerate kinase[22] (PGK) is a homotetramer, each subunit composed of two domains connected by a hinge region. The subunit contains an active site comprising four catalytic residues: R39, K219, G376, and G399. For easy viewing, the dimeric form of PGK is shown in Figure 8(a). It has been reported that the X-ray tetrameric structure of PGK is composed of four independent monomers residing in an asymmetric unit, and that each subunit functions as a monomeric enzyme.[22] Therefore, the subunits of PGK are less evolutionarily coupled with other subunits than the subunits of Set A do. The sequence conservation profile of the subunit is expected to reflect the structural characteristics of only that particular subunit. This is indeed the case: the sequence conservation profile of the PGK monomer has a correlation of 0.70 with the rWCN I profile [Fig. 8(b)], and 0.48 with the rWCN II profile [Fig. 8(c)].

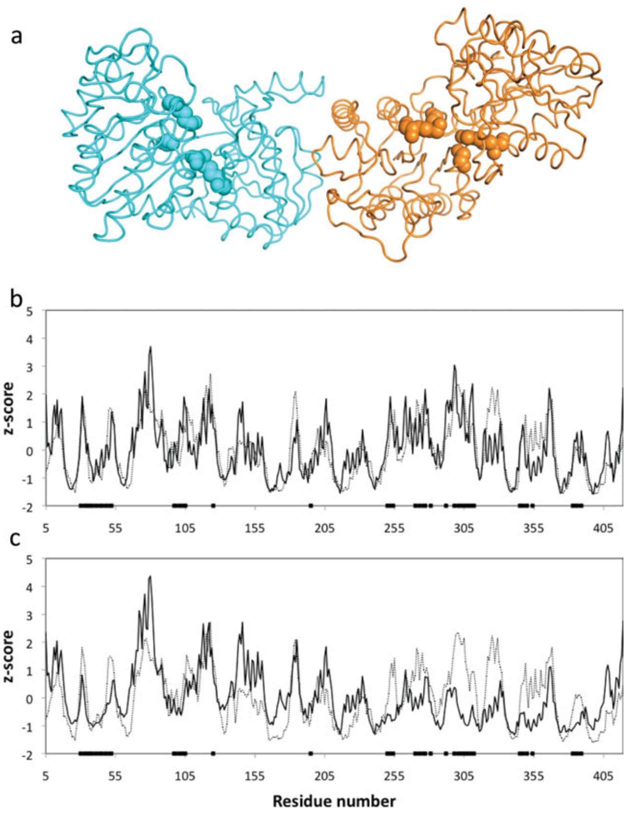

Figure 8
(**a**) The dimeric structure of phosphoglycerate kinase (PDB ID: 13PK). The subunits are colored in cyan and orange, respectively. The catalytic residues are shown in CPK model. The sequence conservation profile of the subunit is compared with (**b**) the rWCN I profile and (**c**) rWCN II profile, respectively. The correlation coefficients of (**b**) and (**c**) are 0.70 and 0.48, respectively.

Muconate lactonizing enzyme

The X-ray structure of muconate lactonizing enzyme (MLE) contains two tetramers, each having a D2-symmetry. Each subunit has an active site comprising four catalytic residues: H148, R196, E212, and R274. A tetramer of MLE is shown in Figure 9(a). The sequence conservation profile of the subunit has a correlation of 0.64 with the rWCN I profile [Fig. 9(b)], and 0.31 with the rWCN II profile [Fig. 9(c)]. The shape of the rWCN I profile is significantly different from that of the rWCN II profile due to the exclusion of the packing contribution of other subunits of MLE. Large changes occur at interface regions like 221–237, 281–310, and 325–349.

DISCUSSION

We compared the sequence conservation profiles and the corresponding rWCN profiles for two sets of proteins, one set composed enzyme complexes with the active sites comprising residues from multiple subunits

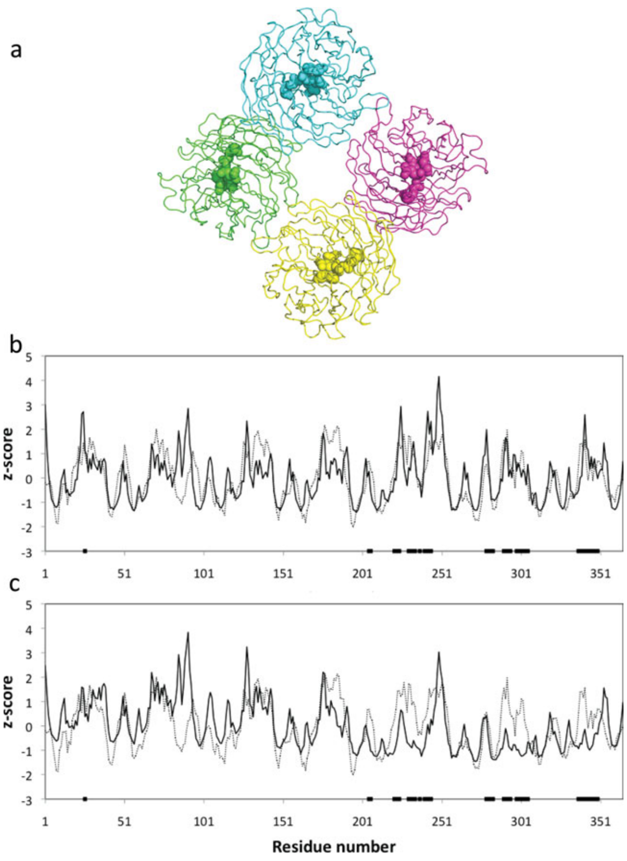

Figure 9

(**a**) The tetrameric structure of carboxy-cis,cis-muconate cyclase (PDB ID: 1JOF). The subunits are rendered in different colors and the catalytic residues shown in CPK model. The sequence conservation profile of the subunit is compared with its (**b**) its rWCN I profile and (**c**) rWCN II profile, respectively. The correlation coefficients of (**b**) and (**c**) are 0.64 and 0.31, respectively.

and the other set with the active sites residing on single subunits. While the sequence conservation profile of a subunit will certainly reflect the influences of other subunits, its rWCN profile can be computed excluding or including the packing contributions of the other subunits of the complex, i.e., the rWCN I and II profiles. It turns out that a simple comparison of the sequence conservation profile of the subunit and the either types of rWCN profile can completely determine the types of the complexes. Our results also clarify the issues concerning under what situations the sequence conservation profile of the subunits of a complex will agree with either the rWCN I or II profiles. For a protein complex, as that of Set A, that has strong evolutionary couplings between its subunits, the sequence conservation profile of its subunit will agree better with the rWCN II profile, which includes the contributions of other subunits. For a complex with weak couplings between the subunits, as that of Set B, the sequence conservation profile will match better with the rWCN I profile, which is computed excluding the contributions of other subunits.

It will be informative to examine the Weng's dataset[23,24] comprising manually classified obligate and transient complexes. The obligate complexes are the complexes whose subunits often fold and bind simultaneously, while the transient complexes, which have the enzyme-substrate (or inhibitor), antibody-antigen or signaling-effector types of interactions, are often involved in momentary contacts between subunits.[24] For the sake of simplicity, we will examine only the obligate and transient dimers. For the obligate dimers, 26% of them are of rWCN I type, and 74% of rWCN II type. The higher percentage of the rWCN II type suggests a higher degree of evolutionary couplings between the subunits of the obligate dimers. For the transient dimers, 48% of them are rWCN I type and 52% of rWCN II type. The relatively lower percentage of the rWCN II type of the set of the transient dimers than that of the obligate dimers suggests a lower degree of evolutionary couplings between their subunits. These results are consistent with the Mintseris and Weng's results[24] that the interface residues of obligate complexes appear to have a lower evolutionary rate, allowing them to have a stronger evolutionary coupling with their interacting partners, while the transient complexes show relatively less correlated mutations across the interfaces.

It is a challenge to select general protein complexes that have possible evolutionary couplings between the subunits. The enzyme complexes whose active sites comparing residues from different subunits provides clear-cut examples of stronger co-evolution across the interfaces as compared with the enzyme complexes with the active sites residing on single subunits. However, at the present stage, we do not yet have a measure, besides the correlation between the density and conservation profiles, sensitive enough to identify subtle coevolutionary signal between subunits.

Many efforts[24,25] have been devoted to understanding the interactions between the subunits of a complex by meticulously examining the structural properties or the evolutionary states of the interface residues. Our approach is appealing both for its simplicity and for its seamlessly combining structural and evolutionary considerations in a natural way. Further study will be required for understanding the nature of and, if possible, quantifying the couplings between the subunits of the complex.

REFERENCES

1. Creighton TE. Proteins: structures and molecular properties. New York: W. H. Freeman and Company; 1993.
2. Branden C, Tooze J. Introduction to protein structure. New York: Garlnd Science; 1999.
3. 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.

4. Shih CH, Chang CM, Lin YS, Lo WC, Hwang JK. Evolutionary information hidden in a single protein structure. Proteins-Struct Funct Bioinformatics 2012;80:1647–1657.
5. Yang L, Song G, Jernigan RL. Protein elastic network models and the ranges of cooperativity. Proc Natl Acad Sci USA 2009;106:12347–12352.
6. Huang SW, Yu SH, Shih CH, Guan HW, Huang TT, Hwang JK. On the relationship between catalytic residues and their protein contact number. Current Protein Peptide Sci 2011;12:574–579.
7. Nosrati GR, Houk KN. Using catalytic atom maps to predict the catalytic functions present in enzyme active sites. Biochemistry 2012;51:7321–7329.
8. 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;12:574–579.
9. 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.
10. Halle B. Flexibility and packing in proteins. Proc Natl Acad Sci USA 2002;99:1274–1279.
11. 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–W533.
12. Boutet E, Lieberherr D, Tognolli M, Schneider M, Bairoch A. UniProtKB/Swiss-Prot. Methods Mol Biol 2007;406:89–112.
13. 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.
14. 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.
15. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 2004;32:1792–1797.
16. 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.
17. Porollo A, Meller J. Prediction-based fingerprints of protein-protein interactions. Proteins 2007;66:630–645.
18. Kabsch W, Sander C. Dictionary of protein secondary structure - pattern-recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.
19. 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:D129–D133.
20. Mosbacher TG, Mueller M, Schulz GE. Structure and mechanism of the ThDP-dependent benzaldehyde lyase from Pseudomonas fluorescens. FEBS J 2005;272:6067–6076.
21. Stamper GF, Morollo AA, Ringe D. Reaction of alanine racemase with 1-aminoethylphosphonic acid forms a stable external aldimine. Biochemistry 1999;38:6714.
22. Bernstein BE, Michels PA, Hol WG. Synergistic effects of substrate-induced conformational changes in phosphoglycerate kinase activation. Nature 1997;385:275–278.
23. Mintseris J, Weng Z. Atomic contact vectors in protein-protein recognition. Proteins 2003;53:629–639.
24. Mintseris J, Weng Z. Structure, function, and evolution of transient and obligate protein-protein interactions. Proc Natl Acad Sci USA 2005;102:10930–10935.
25. Dey S, Pal A, Chakrabarti P, Janin J. The subunit interfaces of weakly associated homodimeric proteins. J Mol Biol 2010;398: 146–160.