Theor Chem Acc (2010) 125:643–650
DOI 10.1007/s00214-009-0656-3

REGULAR ARTICLE

# Why similar protein sequences encode similar three-dimensional structures?

**Szymon Kaczanowski · Piotr Zielenkiewicz**

Received: 3 April 2009 / Accepted: 5 October 2009 / Published online: 23 October 2009


**Abstract** Evolutionarily related proteins have similar sequences. Such similarity is called homology and can be described using substitution matrices such as Blosum 60. Naturally occurring homologous proteins usually have similar stable tertiary structures and this fact is used in so-called homology modeling. In contrast, the artificial protein designed by the Regan group has 50% identical sequence to the B1 domain of Streptococcal IgG-binding protein and a structure similar to the protein Rop. In this study, we asked the question whether artificial similar protein sequences (pseudohomologs) tend to encode similar protein structures, such as proteins existing in nature. To answer this question, we designed sets of protein sequences (pseudohomologs) homologous to sequences having known three-dimensional structures (template structures), same number of identities, same composition and equal level of homology, according to Blosum 60 substitution matrix as the known natural homolog. We compared the structural features of homologs and pseudohomologs by fitting them to the template structure. The quality of such structures was evaluated by threading potentials. The packing quality was measured using three-dimensional homology models. The packing quality of the models was worse for the "pseudohomologs" than for real homologs. The native homologs have better threading potentials (indicating better sequence-structure fit) in the native structure than the designed sequences. Therefore, we have shown that threading potentials and proper packing are evolutionarily more strongly conserved than sequence homology measured using the Blosum 60 matrix. Our results indicate that three-dimensional protein structure is evolutionarily more conserved than expected due to sequence conservation.

**Keywords** Protein · Homology · Modeling · Threading · Evolution

## 1 Introduction

Proteins existing in nature were created by the process of evolution. One can detect evolutionary relationships (homology) because related proteins have similar sequences. Substitution matrices are simple mathematical descriptions of evolutionary relationships [1–4]. They are based on the assumption that residues from different sequence positions evolve independently. Evolutionary related proteins almost always have similar three-dimensional structure [5]. This classical observation of Chothia and Lesk was later confirmed many times (e.g., in the CASP experiments [6]). Even mutations that have a strong impact on protein function usually do not change the protein three-dimensional structure significantly (for example, the mutation causing hemoglobin aggregation and sickle cell anemia has minimal impact on the three-dimensional structure of hemoglobin monomers [7]).

It is often assumed that the structural similarity of homologous proteins is caused only by sequence similarity. This hypothesis is supported by the fact that the probability of substitution of one residue by another during evolution

Dedicated to Professor Sandor Suhai on the occasion of his 65th birthday and published as part of the Suhai Festschrift Issue.

S. Kaczanowski · P. Zielenkiewicz (✉)
Bioinformatics Department, Institute of Biochemistry and Biophysics, Polish Academy of Sciences, Warsaw, Poland
e-mail: piotr@ibb.waw.pl

P. Zielenkiewicz
Plant Molecular Biology Department, Warsaw University, Warsaw, Poland

is higher if these residues have similar volumes and hydrophobicities [1, 8, 9].

In contrast, Regan et al. designed a polypeptide with 50% identical sequence to the $\beta$-sheet B1 domain of Streptococcal IgG-binding protein and also identical to the alpha-helical protein Rop in another 41% of its sequence. According to the CD spectrum, this polypeptide had the structure of the Rop protein [10]. The result was very surprising because natural sequence pairs with 50% identity usually have almost identical structures. The aim of the Regan group was to design proteins having similar sequences and different protein structures (so-called Paracelsus challenge). The design methodology was based on intuition supported by the use of graphical modeling software. Therefore, similar protein sequences do not always have similar three-dimensional structures (the designed protein by Regan and B1 domain of Streptococcal IgG-binding protein had different structures). That experiment supports the hypothesis that natural homologous sequences code similar structures because they have similar evolutionary origins and also the protein structure and function are evolutionarily conserved.

The aim of the present study was to check whether similar sequences tend to encode similar structures. We designed sets of protein sequences homologous to sequences of a known template three-dimensional structure, having the same number of identities, the same composition and equal level of homology, according to Blosum 60 substitution matrix [2] as known natural homologs using a novel computer program (this substitution matrix is widely used for homology searches and can be obtained from www page http://www.molgen.mpg.de/~service/scisoft/gcg/gcg10/moredata/blosum60.cmp). We also confirmed the results using blosum80 substitution matrix.

We called such sequences pseudohomologs. Later, we compared the structural features of the pseudohomologs and their native counterparts.

We used the existing knowledge of the structural features of natural globular proteins. Globular proteins are usually well packed with buried hydrophobic residues and exposed hydrophilic ones [11]; some residue pairs are often in contact [12] and some sequences have preferences for different types of secondary structures [13, 14]. This knowledge can be statistically described. Such mathematical descriptions are widely used for predicting the three-dimensional structure of a protein using the so-called threading, i.e., by searching known folds into which the given sequence structure fits best. One of the first such methods was called Profile 3D [15]. Each residue in the structure is described by its structural environment, i.e., by the residue class, secondary structure and solvent accessibility (buried/exposed). The log odds matrix called Profile 3D describes the probability of each residue being in a given structural environment. A very similar approach is based on Boltzman-like statistics first proposed by Tanaka and Scheraga [12].

According to the Boltzman equation, the free energy difference can be expressed in terms of the probability of an event (for example, a chemical reaction). Similar parameters can be used to describe our knowledge about protein structures, e.g., one can calculate the probability that a given residue is buried in the interior of the protein. Of course, such parameters do not represent real free energies and therefore are often called pseudoenergies or knowledge-based potentials. There are many types of knowledge-based potentials [16–20].

In the present study, we used such statistical knowledge to compare how well the sequences of native homologs and "pseudohomologs" fit the template structure of an evolutionarily related protein.

## 2 Materials and methods

### 2.1 PDB template structures

Structures of the following proteins were used in this study: hen lysozyme [21] PDB code 132L (alpha-helical protein), *Bacillus amyloliquefaciens* barnase [22, 23] (protein containing both beta sheets and alpha-helical secondary structures), PDB code 1RNB, point mutant of human carbonic anhydrase [24], PDB code 12CA (mainly beta sheet protein), and bovine ribonuclease A [25], PDB code 2AAS (protein containing both beta and alpha-helical structures). These proteins have well-known structures and functions.

### 2.2 Homology searches

We used the implementation of the Smith Waterman algorithm from the WWW server at the EBI http://www.ebi.ac.uk/MPsrch/ and found sequences from the SWISS-PROT database (release 41) having full-length non-gapped alignment to our template proteins. Some of these sequences were used later for comparison with pseudohomologs (see below).

### 2.3 Designing pseudohomologs

The design procedure started from a randomized sequence having a given composition. Two randomly chosen residues were exchanged and this exchange was repeated once again. Such two exchanges were the basic steps of the optimization, and the resulting change of sequence was accepted if the optimization function was lowered. The optimization function was calculated in the following way:

optimization function $= F_i + F_h$

$F_i$ is the function of the target number of identities and

$F_i = 1000 * \mathrm{ABS}(io - id)$

where io is the number of identities between the designed sequence and the sequence of the template protein, id is the target number of identities between the designed sequence and the template sequence. $F_h$ is the function of target homology equal to:

$$H_o - H_d \text{ when } H_o < H_d \quad (1)$$

$$0 \text{ when } H_d < H_o < H_d + 10 \quad (2)$$

$$H_d - H_o - 10 \text{ when } H_o > H_d + 10 \quad (3)$$

where $H_o$ is the homology between the designed sequence and the template measured using Blosum 60 matrix expressed in ½ bit score, $H_d$ is the target homology between the designed sequence and the template sequence measured using Blosum 60 matrix expressed in ½ bit score.

The optimization was complete when the optimization function reached 0. Therefore, the designed sequences have equal number of identities and nearly equal homology to their natural counterparts (it can be higher by up to 5 bits).

We used this algorithm for designing pseudohomologs, i.e., sequences homologous to natural sequences, having the same number of identities, the same composition and equal level of homology, according to Blosum 60 substitution matrix as known natural homologs.

We designed the following sets of 100 pseudohomolog sequences.

1. Pseudohomologs of bovine ribonuclease: we used the *Rattus rattus* (black rat) ribonuclease A SWISSPROT code RNP_RATRT as a natural sequence.
2. Pseudohomologs of hen lysozyme: natural homologous sequences were used:
   a. lysozyme of *Anas platyrhynchos* (domestic duck) SWISSPROT code LYC1_ANAPL;
   b. lysozyme of*Oncorhynchus mykiss* (rainbow trout) SWISSPROT code LYC2_ONCMY.
3. Pseudohomologs of barnase of *Bacillus amyloliquefaciens*. Natural homologous sequences were used:
   a. putative ribonuclease of the plague causing bacterium*Yersinia pestis* SP-TREMBL code Q8ZAUZ;
   b. ribonuclease of the thermophilic bacterium*Bacillus coagulans* SWISSPROT code RN_BACCO.
4. Pseudohomologs of artificial human carbonic anhydrase II mutant (PDB code 12ca).

Natural sequence of human carbonic anhydrase III SWISSPROT code CAH3_HUMAN was also used.

### 2.4 Measuring threading potentials

As mentioned above, we assumed that homologs of the template proteins have nearly identical 3D structures. Therefore, we measured threading potentials of the homologs in the known template PDB structures.

Two types of threading potentials were used, Godzik function [18] and Profile 3D [15]. None of these use information on sequence similarity of target and template. The sequence-structure fit is calculated using only query sequence and template structure. Homology models of the query are not used for calculations. Using this approach, we avoid problems caused by errors in homology modeling, as structural differences between template structure and homology models could be caused only by errors in protein modeling.

The threading potential is the pseudoenergy function as defined by Godzik and Skolnick [18]. The knowledge-derived potential of Godzik and Skolnick consists of three terms: one body (the probability that a given residue is buried or exposed), two body (the probability that a pair of residues are in spatial contact) and three body (the probability of the appearance of a cluster of three mutually interacting residues).

Because the three body potential usually is not statistically significant [18], we used only one and two body potentials. We also measured the fit of the homologous sequences to the template structures using an implementation of the second type of threading potentials: Profile 3D from the Biosym/MSI software (MSI, San Diego, USA, [15]). For each PDB structure, a so-called Profile 3D was calculated using the Create_Profile command and the fit of the homologous sequences was measured using the Find_Structure command versus a database containing only one calculated profile.

### 2.5 Homology models and validation of packing quality

Homology models were built using the HOMOLOGY module of the Biosym/MSI software (MSI, San Diego, USA). Crude models were built without energy minimization. A special program in BIOSYM command language was written, which made it a fully automated procedure. It is worth to note that a residue of a given type has approximately constant volume in different proteins. Therefore, the expected protein volume of the pseudohomologs and their native counterparts is equal. Sometimes, protein models are not perfect and neighboring residues overlap. Therefore, a larger volume indicates that there are less overlaps between residues of such non-perfect models. Of course, residue overlaps are impossible in a real protein structure. Always, such models can be improved by very

detailed modeling of protein structure, but occasionally such correction is not true and the protein cannot fold in the target structure. We assumed that the probability that such a correction was higher when the volume of overlaps in the crude model was smaller, i.e., the volume of the crude model was larger. Therefore, we used the volume of such crude models for validating the quality of the packing. The volumes of the models were measured by the Voiddo computer program of the Uppsala Software Factory [26].

## 3 Results and discussion

At the beginning, we checked whether the proper values of knowledge-based potentials were evolutionarily conserved in homologous proteins having full length non-gapped alignment. As shown in Table 1, the evolutionary relationships among such sequences can be quite different (from 97% identity to about 60% identity). All these proteins have equally good values of knowledge-based potentials in their native structure. The value of the knowledge-based potentials does not depend on the evolutionary relationship. This result is expected, because knowledge-based potentials measure how well a sequence fits to a given structure and not the evolutionary relationship. It is worth mentioning that such a result indicates that knowledge-based potentials used properly measure the quality of sequence-structure fit.

Sequence-structure fits were measured using the classical threading knowledge-based potentials of Godzik [18] and Profile 3D [15]. The quality of packing was expressed as the volume of homology models. Figures 1, 2, 3 show the detailed results for trout lysozyme C (swissprot code LYC_ONCMY). Native trout lysozyme has smaller (better) one body and two body potentials of Godzik and, as a result, also the sum of one and two body potentials, than the pseudohomologs (see Fig. 1). In the case of Profile 3D [15], the results are very similar and the pseudohomologs have worse quality of Profile 3D expressed as the so-called *Z* score. (see Fig. 2). We also tried to check the quality of

**Table 1** Comparison of knowledge-based potentials of template proteins from PDB database and their close structural homologs

| Swissprot accession id | Identity% | One body potential of Godzik and Skonick (notice that better is smaller) | Two body potential of Godzik and Skonick (notice that better is smaller) | Sum of one and two body potentials of Godzik and Skonick (notice that better is smaller) | Profile 3D score expressed as *Z* score (notice that better is bigger) |
|---|---|---|---|---|---|
| Comparison of hen lysozyme (PDB code 132 l, SWISSPROT code LYC_CHICK, length 129 aa) and their close structural homologs having swissprot ID LYC_COLVI, LYC_LOPCA, LYC_PAVCR, LYC_CHRAM, LYC_COTJA, LYC_MELGA, LYC_SYRRE, LYC_LOPLE, LYC_PHACO, LYC_SYRSO, LYC_NUMME, LYC_PHAVE, LYC1_ANAPL, LYC3_ANAPL, LYC_ORTVE, LYC2_ONCMY | | | | | |
| LYC_CHICK | 100 | −19.8 | −5.5 | −25.3 | 37.2 |
| LYC2_ONCMY | 60.5 | −19.2 | −8.3 | −27.5 | 31 |
| RANGE | 60.5, 100 | −20, −16 | −8.3, 0.7 | −27.5, −15.4 | 37.24, 31 |
| Comparison of barnase of *Bacillus amyloliquefaciens* (PDB code 1RNB, length 109 AA) and its close structural homologs having swissprot ID: RNBR_BACAM, RN_BACCI, RN_BACIN, RN_BACPU, RN_BACCO and SP-TREMBL ID Q8ZAUZ | | | | | |
| 1RNB | 100 | −8.3 | −4.7 | −13. | 27.9 |
| Q8ZAUZ | 56 | −16.7 | 0.5 | −16.2 | 22.1 |
| RANGE | 56, 100 | −4.3, −16.7 | 0.5, −5.1 | −6.3, −16.2 | 22.1, 27.9 |
| Comparison of artificial mutant of human carbonic anhydrase II (PDB code 12CA, length 255 AA) and its close structural homologs having swissprot ID: CAH2_HUMAN, CAH2_RABIT, CAH2_MOUSE, CAH2_RAT, CAH2_CHICK, CAHZ_BRARE, CAH3_RAT, CAH3_HORSE, CAH3_HUMAN | | | | | |
| 2CA | 100 | −31.3 | −29.9 | −61.2 | 58.9 |
| CAH3_HUMAN | 58 | −30.2 | −19.8 | −50 | 49.6 |
| RANGE | 58, 100 | −31.5, −24.2 | −30.4, −11.3 | −61.9, −38.5 | 49, 58.9 |
| Comparison of bovine ribonuclease A (PDB code 2aas, SWISSPROT code RNP_BOVIN, length 124 AA) and its close structural homologs having accession ID: RNP_SHEEP, RNP_AEPME, RNP_BUBBU, RNP_CONTA, RNP_ANTAM, RNP_GIRCA, RNP_RANTA, RNP_CAPCA,RNP_ALCAA,RNP_CEREL, RNP_DAMDA, RNP_AXIPR, RNP_HIPAM, RNBR_GIRCA, RNS_BOVIN, RNP_PIG, RNBR_BOVIN, RNBR_AXIPR, RNBR_CAPCA, RNBR_SHEEP, RNP_BALAC, RNP_HYSCR, RNPB_CAVPO, RNP_CHIBR, RNP_CHOHO, RNP_MYOCO, RNPA_CAVPO, RNP_HORSE, RNP_ACOCA, RNP_HYDHY, RNP_MESAU, RNP_PROGU, RNP_URARU, RNP_CRILO, RNP_ONDZI, RNP_HUMAN, RNP_MOUSE, RNP_MUSPA, RNP_MUSSA, RNP_PREEN, RNP_LEOED, RNP_RAT, RNP_NIVCR, RNP_RATRT | | | | | |
| RNP_BOVIN | 100 | −15.8 | 1 | −14.8 | 28 |
| RNP_RATRT | 63.7 | −18.4 | 2.4 | −16 | 25.2 |
| RANGE | 63.7, 100 | −12.3, −21.6 | −5.6, 6.4 | −5.9, −22.53 | 28.7, 21.5 |

We assumed that used homologs have almost identical structure to the template proteins from PDB

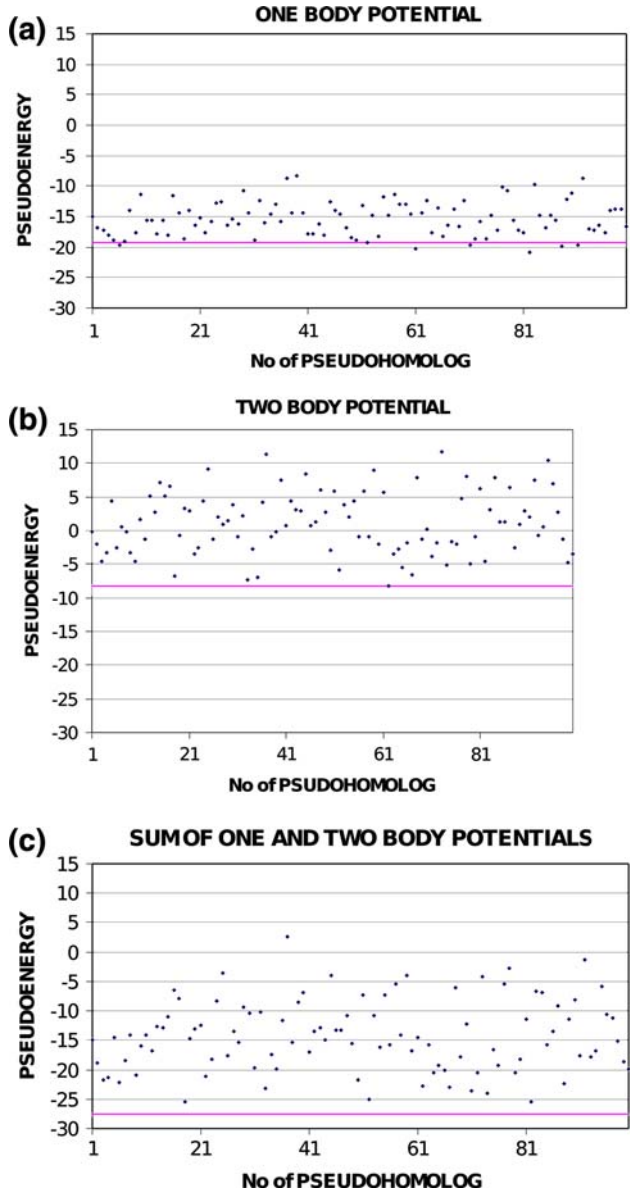


**Fig. 1** A comparison of the **a** one, **b** two and **c** sum of one and two body potentials of a natural hen lysozyme homolog, lysozyme C from *Oncorhynchus mykiss* (rainbow trout) and a set of 100 designed homologous (pseudohomologs). The *line* indicates the value for lysozyme C from *O. mykiss*, and *points* represent values for different pseudohomologs. We assumed that all these proteins have almost identical structures to hen lysozyme. Notice that in the case of one and two body potentials, smaller values indicate better sequence-structure fit

protein packing in the target protein structure [26]. The quality of protein packing was worse for the majority of pseudohomologs than in the case of native trout lysozyme (see Fig. 3).

We confirmed these results and checked that also pseudohomologs deigned using the Blosum 80 substitution matrix has worse Godzik potentials than trout lysozyme.

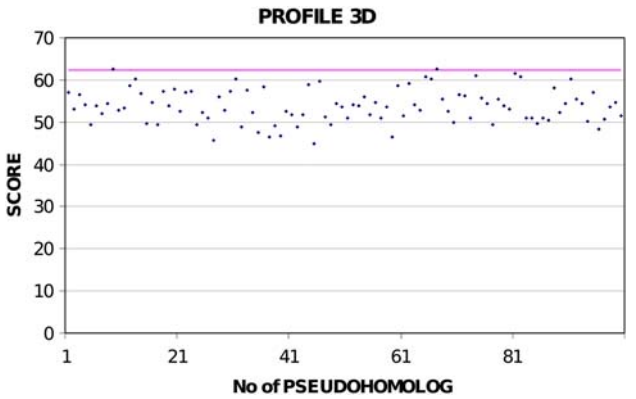


**Fig. 2** A comparison of Profile 3D of a natural hen lysozyme homolog, lysozyme C from *Oncorhynchus mykiss* (rainbow trout) and a set of 100 designed pseudohomologs. The *line* indicates the value for lysozyme C from *O. mykiss*, and *points* represent values for different pseudohomologs. We assumed that all these proteins have almost identical structures to hen lysozyme. Notice that in the case of Profile 3D scores, smaller values indicate worse sequence-structure fit

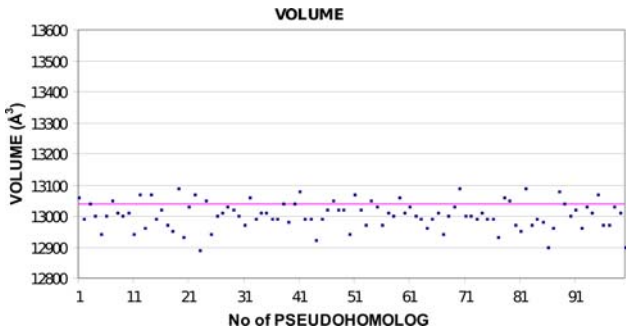


**Fig. 3** A comparison of homology model volume of a natural hen lysozyme homolog, lysozyme C from *Oncorhynchus mykiss* (Rainbow trout) and a set of 100 designed pseudohomologs. The *line* indicates the value for lysozyme C from *O. mykiss*, and *points* represent values for different pseudohomologs. We assumed that all these proteins have almost identical structures to hen lysozyme. In case of homology models volume, smaller values indicate worse sequence-structure fit

We checked also that similar results were obtained for the duck lysozyme, which is a closer homolog of hen lysozyme (82% identity) than trout lysozyme. The advantage of the native protein over the pseudohomologs in this case was smaller. Such a result is not surprising, because closer homologs of hen lysozyme (and pseudohomologs) are more similar to hen lysozyme and therefore to each other.

We also checked the results for other proteins. These results are presented in Table 2. In the majority of cases, real homologs fit better to the native structure than the designed pseudohomologs. An exception to the above rule was seen for the barnase of *Bacillus Coagulans*, a natural homolog of barnase from *Bacillus Amyloliquefaciens*: it had worse (higher) pseudoenergies of Godzik and Skolnick than the designed pseudohomologs.

**Table 2** Comparison of different sets of pseudohomologs and native homologous sequences

| Template protein structure | Natural homologous protein | Percentage of pseudohomologs having one body potential worse than natural homolog (%) | Percentage of pseudohomologs having two body potential worse than natural homolog (%) | Percentage of pseudohomologs having sum of one and two body potentials worse than natural homolog (%) | Percentage of pseudohomologs having profile 3D score worse than natural homolog (%) | Percentage of pseudohomologs packed worse than native homolog (%) |
|---|---|---|---|---|---|---|
| Hen lysozyme PDB code 132l | LYC2_ONCMY | 93 | 99 | 100 | 98 | 77 |
| | LYC1_ANAPL | 68 | 91 | 100 | 96 | 78 |
| Barnase of *Bacillus amyloliquefaciens* PDB code 1RNB | Q8ZAUZ | 100 | 62 | 100 | 100 | 98 |
| | RN_BACCO | 36 | 36 | 31 | 79 | 78 |
| Bovine ribonuclease A PDB code 2aas | RNP_RATR | 98 | 53 | 87 | 100 | 85 |
| Mutant of human carbonic anhydrase II PDB code 12CA | CAH3_HUMAN | 100 | 100 | 100 | 100 | 100 |

These unexpected results can be explained by the fact that *Bacillus Coagulans* is a moderate thermophilic organism [27]. Because proteins of extremophiles differ from that of non-extremophiles [28–30], we hypothesized that this exception was caused by the fact that the parameters of Godzik were probably not useful for proteins of thermophilic organisms. Therefore, we also ran a comparison for a native homolog of barnase from the non-thermophilic bacteria *Yersinia pestis* [31]. As expected, in this case the native protein had better Godzik knowledge-based potentials than the pseudohomologs. The results presented above show that proper sequence-structure fit is more strongly evolutionarily conserved than expected due to sequence conservation.

If this is true, it is expected that designed pseudohomologs with better homology level fit better to the target structure. We confirmed this expectation (see Fig. 4).

Therefore, the strict evolutionary conservation of protein structure cannot be explained using Markovian models of evolution based on the assumption that residues from different positions evolve independently. Apparently, residues from different positions co-evolve to conserve protein structure. An important implication of this fact is that the homology of two proteins alone is not a sufficient condition to assume that these proteins have similar folds.

The phenomenon of co-evolution has already been described. It has been shown that a correlation exists between mutations in different positions in protein structure [32–35]. Such co-evolving residues are often neighboring ones [33, 35]. It is still not well understood why the mutations are correlated [32–34].

This is in agreement with the fact that often correlated mutations appear in neighboring residues. It is worth mentioning that the observed strict evolutionary conservation of packing suggests that mutations in neighboring residues should be complementary in volume. Such a phenomenon was shown previously [32]. It is not clear if this rule is general, because such a correlation of residue volume complementation in neighboring positions in myoglobin does not exist [34].

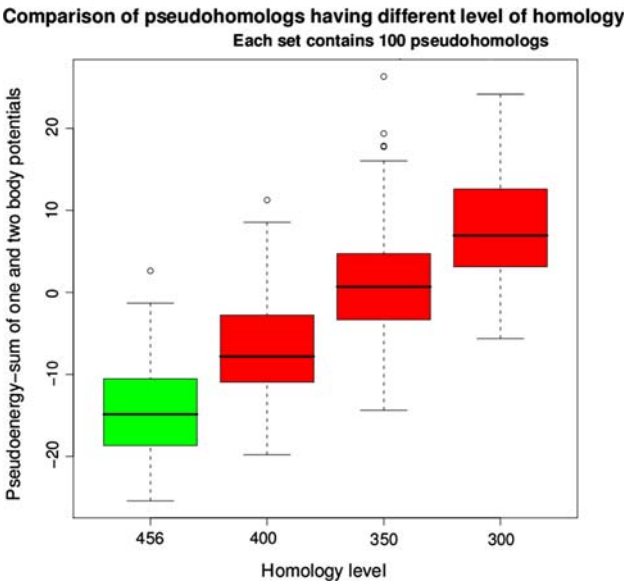


**Fig. 4** A comparison of one body and two body potentials for sets of 100 designed homologous sequences having the same number of identities to hen lysozyme and identical composition as lysozyme C from *Oncorhynchus mykiss* (Rainbow trout) and different level of homology (according to the bits scale of Blosum60). The *green box plot* indicates homologs having native-like level of homology. Each compared set contains 100 pseudohomologs

The importance of the evolutionary conservation of proper packing has also been shown in the case of lysozymes [36]. In [36], it was shown that extant lysozymes are more stable than the ancestral protein and that this was caused by adaptive mutations and improvement of protein packing.

The observed strict conservation of proper Godzik and Skolnick [18] one body potentials (describing preferences of given residue to be buried and exposed) and proper Profile 3D [15] (describing the preferences of a given residue to be in a given position in protein structure) suggests a second type of co-evolution. These two parameters are Markovian models of sequence-structure fit. Consequently, the quality of how a given residue fits to a given position in the protein structure is independent. Therefore, the mutations that lower the quality of protein structure, have to be compensated by mutations that improve the quality of sequence-structure fit. Such mutations need not concern neighboring residues only. This conclusion is in agreement with the fact that correlated mutations are often not neighboring [35].

## 4 Conclusions

We described the sequence-structure fit using knowledge-based potentials and measuring the quality of packing. The results indicate that proper structure fit is more strictly evolutionarily conserved than it would be caused only by conservation of sequence similarity as described by substitution matrices. In contrast, it is expected that artificial similar "homologous" sequences may encode very different three-dimensional structures. Our results indicate that it is unlikely that during protein evolution new folds will appear. One can hypothesise that there are folds, which are not realized by existing protein sequences. This conclusion is supported by the observation, that there is a limited number of protein folds existing in nature [37].

## References


1. Dayhoff MO, Schwartz RM, Orcutt BC (1978) A model of evolutionary change in proteins. In: Dayhoff MO (ed) Atlas of protein Sequence and Structure, vol 5. Nat. Biomed Res Found. Washington DC, pp 345–352
2. Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 89:10915–10919
3. Gonnet GH, Cohen MA, Benner SA (1992) Exhaustive matching of the entire protein sequence database. Science 256:1443–1445
4. Baussand J, Carbone A (2008) Inconsistent distances in substitution matrices can be avoided by properly handling hydrophobic residues. Evol Bioinform Online 4:255–261
5. Lesk AM, Chothia C (1980) How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. J Mol Biol 136:225–270
6. Moult J (2005) A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. Curr Opin Struct Biol 15:285–289
7. Padlan EA, Love WE (1985) Refined crystal structure of deoxyhemoglobin S. II. Molecular interactions in the crystal. J Biol Chem 260:8280–8289
8. Koshi JM, Goldstein RA (1995) Context-dependent optimal substitution matrices. Protein Eng 8:641–645
9. Koshi JM, Goldstein RA (1997) Mutation matrices and physical–chemical properties: correlations and implications. Proteins 27:336–344
10. Dalal S, Balasubramanian S, Regan L (1997) Protein alchemy: changing beta-sheet into alpha-helix. Nat Struct Biol 4:548–552
11. Chothia C (1975) Structural invariants in protein folding. Nature 254:304–308
12. Tanaka S, Scheraga HA (1976) Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. Macromolecules 9:945–950
13. Chou PY, Fasman GD (1974) Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins. Biochemistry 13:211–222
14. Chou PY, Fasman GD (1978) Prediction of the secondary structure of proteins from their amino acid sequence. Adv Enzymol Relat Areas Mol Biol 47:45–148
15. Bowie JU, Luthy R, Eisenberg D (1991) A method to identify protein sequences that fold into a known three-dimensional structure. Science 253:164–170
16. Miyazawa S, Jernigan RL (1985) Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. Macromolecules 18:534–552
17. Miyazawa S, Jernigan RL (1996) Residue–residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. J Mol Biol 256:623–644
18. Godzik A, Skolnick J (1992) Sequence-structure matching in globular proteins: application to supersecondary and tertiary structure determination. Proc Natl Acad Sci USA 89:12098–12102
19. Nishikawa K, Matsuo Y (1993) Development of pseudoenergy potentials for assessing protein 3-D-1-D compatibility and detecting weak homologies. Protein Eng 6:811–820
20. Jones DT, Thornton JM (1996) Potential energy functions for threading. Curr Opin Struct Biol 6:210–216
21. Rypniewski WR, Holden HM, Rayment I (1993) Structural consequences of reductive methylation of lysine residues in hen egg white lysozyme: an X-ray analysis at 1.8-A resolution. Biochemistry 32:9851–9858
22. Mauguen Y, Hartley RW, Dodson EJ, Dodson GG, Bricogne G, Chothia C, Jack A (1982) Molecular structure of a new family of ribonucleases. Nature 297:162–164
23. Baudet S, Janin J (1991) Crystal structure of a barnase-d(GpC) complex at 1.9 A resolution. J Mol Biol 219:123–132
24. Nair SK, Calderone TL, Christianson DW, Fierke CA (1991) Altering the mouth of a hydrophobic pocket. Structure and kinetics of human carbonic anhydrase II mutants at residue Val-121. J Biol Chem 266:17320–17325
25. Santoro J, González C, Bruix M, Neira JL, Nieto JL, Herranz J, Rico M (1993) High-resolution three-dimensional structure of ribonuclease A in solution by nuclear magnetic resonance spectroscopy. J Mol Biol 229:722–734
26. Kjeldgaard M, Jones TA (1994) Detection, delineation, measurement and display of cavities in macromolecular structures. Acta Crystallogr D Biol Crystallogr 1:178–185

27. Watanabe K, Kitamura K, Suzuki Y (1996) Analysis of the critical sites for protein thermostabilization by proline substitution in oligo-1, 6-glucosidase from *Bacillus coagulans* ATCC 7050 and the evolutionary consideration of proline residues. Appl Environ Microbiol 62:2066–2073
28. Jaenicke R, Bohm G (1998) The stability of proteins in extreme environments. Curr Opin Struct Biol 8:738–748
29. Madern D, Ebel C, Zaccai G (2000) Halophilic adaptation of enzymes. Extremophiles 4:91–98
30. Saunders NF, Thomas T, Curmi PM, Mattick JS, Kuczek E, Slade R, Davis J, Franzmann PD, Boone D, Rusterholtz K, Feldman R, Gates C, Bench S, Sowers K, Kadner K, Aerts A, Dehal P, Detter C, Glavina T, Lucas S, Richardson P, Larimer F, Hauser L, Land M, Cavicchioli R (2003) Mechanisms of thermal adaptation revealed from the genomes of the Antarctic Archaea *Methanogenium frigidum* and *Methanococcoides burtonii*. Genome Res 13:1580–1588
31. Perry RD, Fetherston JD (1997) *Yersinia pestis*–etiologic agent of plague. Clin Microbiol Rev 10:35–66
32. Altschuh D, Lesk AM, Bloomer AC, Klug A (1987) Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus. J Mol Biol 193:693–707
33. Gobel U, Sander C, Schneider R, Valencia A (1994) Correlated mutations and residue contacts in proteins. Proteins 18:309–317
34. Neher E (1994) How frequent are correlated changes in families of protein sequences? Proc Natl Acad Sci USA 91:98–102
35. Horner DS, Pirovano W, Pesole G (2008) Correlated substitution analysis and the prediction of amino acid structural contacts. Brief Bioinform 9:46–56
36. Malcolm BA, Wilson KP, Matthews BW, Kirsch JF, Wilson AC (1990) Ancestral lysozymes reconstructed, neutrality tested, and thermostability linked to hydrocarbon packing. Nature 345:86–89
37. Chothia C (1992) Proteins one thousand families for the molecular biologist. Nature 357:543–544