*Published online 28 April 2008* *Nucleic Acids Research, 2008, Vol. 36, Web Server issue* ***W185–W189***
*doi:10.1093/nar/gkn218*

# HOMCOS: a server to predict interacting protein pairs and interacting sites by homology modeling of complex structures

**Naoshi Fukuhara[1] and Takeshi Kawabata[1,2,*]**

[1]Graduate School of Information Science, Nara Institute of Science and Technology, 8916-5 Takayama, Ikoma, Nara 630-0192 and [2]CREST, Japan Science and Technology Agency, Japan

Received February 2, 2008; Revised April 4, 2008; Accepted April 9, 2008

## ABSTRACT

**As protein–protein interactions are crucial in most biological processes, it is valuable to understand how and where protein pairs interact. We developed a web server HOMCOS (Homology Modeling of Complex Structure, http://biunit.naist.jp/homcos) to predict interacting protein pairs and interacting sites by homology modeling of complex structures. Our server is capable of three services. The first is modeling heterodimers from two query amino acid sequences posted by users. The server performs BLAST searches to identify homologous templates in the latest representative dataset of heterodimer structures generated from the PQS database. Structure validity is evaluated by the combination of sequence similarity and knowledge-based contact potential energy as previously described. The server generates a sequence-replaced model PDB file and a MODELLER script to build full atomic models of complex structures. The second service is modeling homodimers from one query sequence. The third service is identification of potentially interacting proteins for one query sequence. The server searches the dataset of heterodimer structures for a homologous template, outputs the candidate interacting sequences in the Uniprot database homologous for the interacting partner template proteins. These features are useful for wide range of researchers to predict putative interaction sites and interacting proteins.**

## INTRODUCTION

Protein–protein interactions support a wide range of cellular functions in all forms of life, from bacterial cell division to mammalian immunity (1). Characterizing interacting protein pairs and interaction sites is necessary to fully understand the molecular mechanism of cellular activities. Recently, high-throughput screening methods, such as yeast-two-hybrid (Y2H) method and tandem affinity purification (TAP), have generated large datasets of protein–protein interactions. While these data provide a wealth of information about cellular processes, such experiments have been performed for only a few organisms, and may contain unreliable or inaccurate data (2–4). Large amounts of 3D data detailing protein complex structures have been accumulated in the wwPDB database (5); this source is thought to be more reliable than high-throughput methods. In addition, the wwPDB database provides atomic details of protein–protein interface, although number of 3D complex data sets is much smaller than that for high-throughput methods. Homology modeling approaches can be used to extend the accurate interaction data of 3D complex structures (6–13). Such studies have employed a common standard procedure. First, structures for the two target proteins in the complex are generated by comparative-modeling methods. The BLAST and PSI-BLAST programs (14) have been employed by most researchers to search for template complex structures. Next, the validity of the modeled structures is evaluated by calculation of interaction energies. Knowledge-based residue–residue contact energies were employed by most researchers. A number of researchers reported that combination of sequence and structural score was effective to improve prediction performances (9–11). A more detailed interaction energy function using a full atomic model of complex structures was also employed (12,13). Several web servers predicting protein–protein pairs based on homology modeling have been developed. The servers InterPreTS (15) and 3D-partner (9) are able to predict interacting partners for a query protein sequence posted by users. The MODBASE database (16) provides the putative complex models of yeast proteomes.

We propose a new server, HOMCOS (Homology Modeling of Complex Structure), for homology modeling of complex structures and predicting the interacting

*To whom correspondence should be addressed. Tel: +81 743 72 5396; Fax: +81 743 72 5396; Email: takawaba@is.naist.jp

partners of query protein sequences posted by users. The server has three services: modeling heterodimers, modeling homodimers and identifying putative interacting proteins. The basic approach of our server is similar to previously described related servers; however, our server has several advantages over these servers. First, we employed a new score function using the combination of sequence similarity and knowledge-based contact potential energy to validate the predicted interactions (11). Second, the server provides users, multiple ways to examine modeled complex structures. A simple sequence-replaced model can be viewed in the browser using the Jmol software, and downloaded from the server in PDB format. A MODELLER script (17) allows users to model complex structures when atomic details of protein–protein interface are desired. Third, the server facilitates the modeling of homodimers, which are common and important structures in a variety of molecular functions (18). Finally, we employed the latest representative dimer sets based on the PQS server (19) using a new similarity measure between dimmers to create more reasonable representatives.

## METHODS

### Modeling heterodimers and homodimers

To model heterodimer, the HOMCOS server accepts two query protein sequences. The heterodimeric complex structure is derived from a homologous template dimer structure, as summarized in Figure 1. After the two query sequences are input, the HOMCOS server performs two BLAST searches (14) for each query sequence against a sequence database of representative protein heterodimers. The database was generated using the PQS server (19), the details of which are described in the following section. The server then checks if a dimer template structure exists in the database that contains two proteins homologous to the query proteins. If a dimer template structure is found, model validity is evaluated by the score of sequence similarity $Z_{seq}$ and the score of statistical contact energy $Z_{con}$. The details of these scores are described in a previous report (11).

The server then generates a simple sequence-replaced model by replacing the residue names and numbers in the PDB file of the template structure with those of the query protein using the BLAST alignment. The atoms of the substituted side chains and inserted residues, however, are not correctly modeled; the sequence-replaced model has only a rough residue-level resolution. The structure, however, can be quickly obtained and is precise enough to identify the overall structural features of the complex. The model can then be downloaded from the server in the PDB format and visualized in the browser using Jmol software (http://www.jmol.org). Interacting residues and contact residue pairs are also shown, which can be estimated from the sequence-replaced model. The server also provides alignment and script files for the MODELLER program (17), which allows users to build a full atomic model of complex structures. The user can immediately start modeling using the files generated by the HOMCOS server, if the MODELLER program is available for the user. Screenshots of the service are shown in Figure 2.

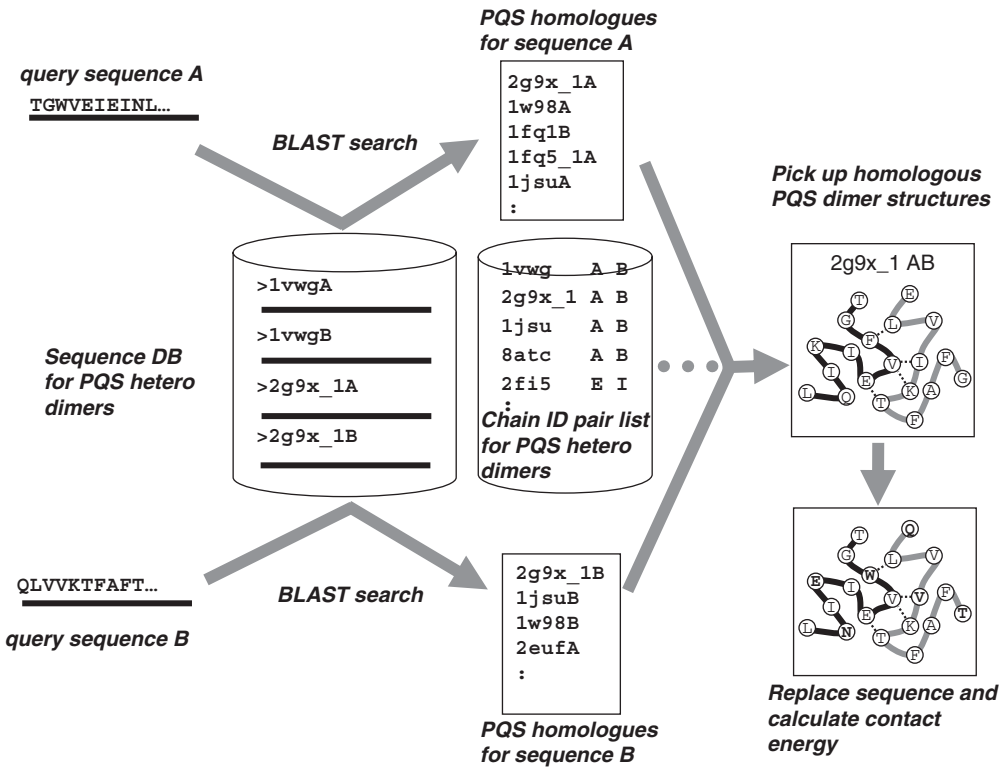


**Figure 1.** Overview of the procedures for modeling heterodimer structures.

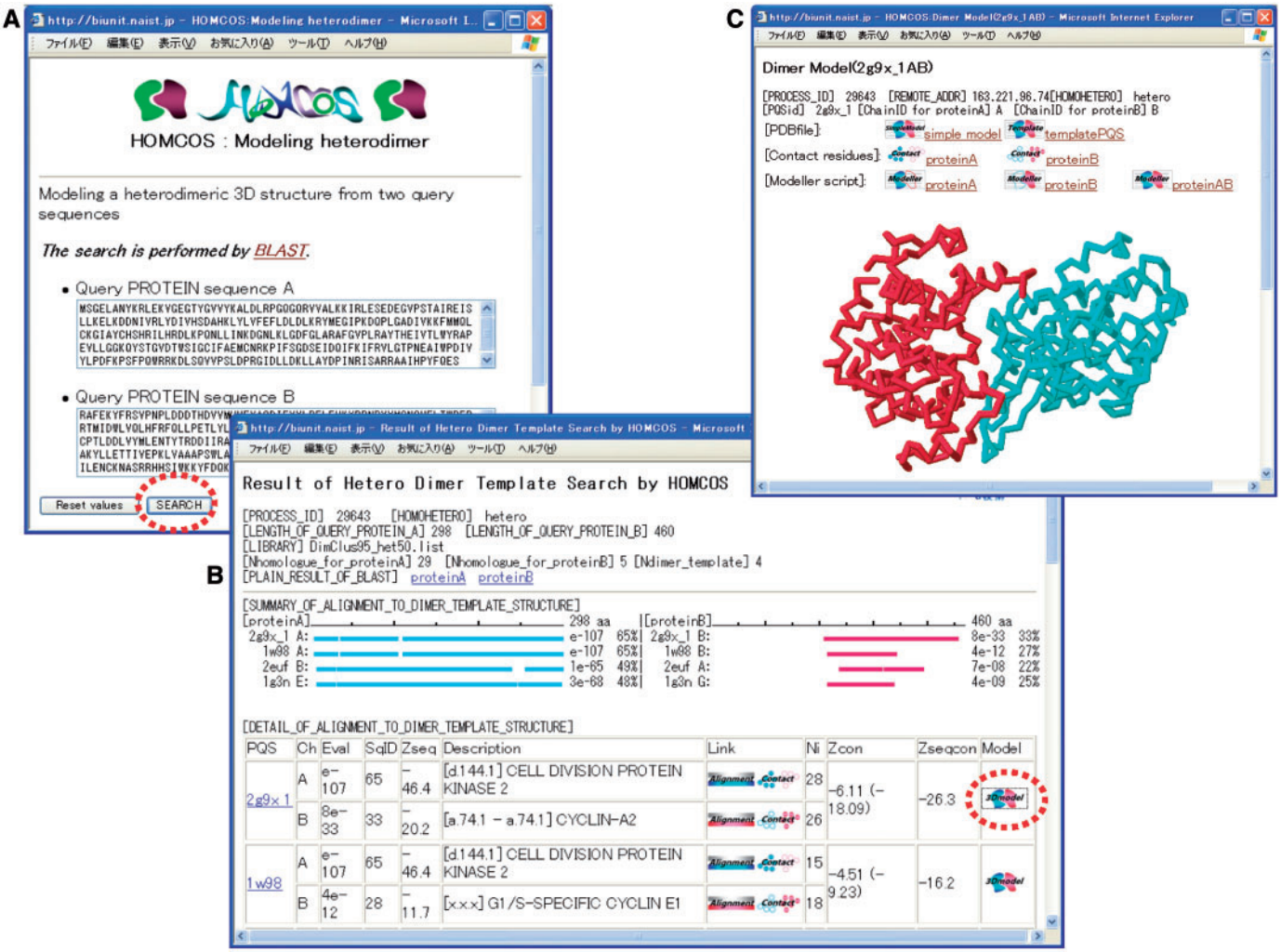


**Figure 2.** Screenshots of the service for modeling heterodimer structures. (**A**) The title page contains two forms in which a user can input two query protein sequences. (**B**) A result summarizing two BLAST searches against the heterodimer database. (**C**) A generated simple sequence-replaced model viewed with the Jmol software.

The procedures for the modeling of homodimers are similar to those for heterodimers. The HOMCOS server accepts only one query protein sequences and then performs a BLAST search against a sequence database of representative homodimers.

### Identifying putative interacting proteins

The HOMCOS server allows users to identify putative interacting protein that may interact with a query protein sequence, which is summarized in Figure 3. As for heterodimer modeling, the server initially performs a BLAST search for the query sequence against a sequence database of representative protein heterodimers. From the list of homologues and the pair list of PQS chains, candidate interacting proteins are identified from the PQS database. The server has a BLAST homologue table for each PQS protein of homologous Uniprot entry lists (20). From the candidate interacting PQS proteins and the table of Uniprot homologues for PQS proteins, the server displays candidate proteins that may interact with the query protein as a list of Uniprot entries. The candidate entries are grouped by organism. A user can then model complex structures of the query protein and one of putative interacting candidate proteins using our heterodimer modeling service (described above).

### Representative datasets of heterodimer and homodimer structures

Representative datasets of heterodimers and homodimers are generated from the quaternary structure database downloaded from the PQS server (19). These datasets were generated as follows. First, all multimers included in the PQS database were separated into dimers. Dimers with fewer than five interacting residues, which are defined as a residue with at least one heavy atom located within 4 Å of a heavy atom of another protein chain, were removed. Next, these dimers were classified as either into heterodimers or homodimers. Heterodimers were defined as proteins whose sequence identity was less than or equal to 50%, the other dimers whose sequence identity was

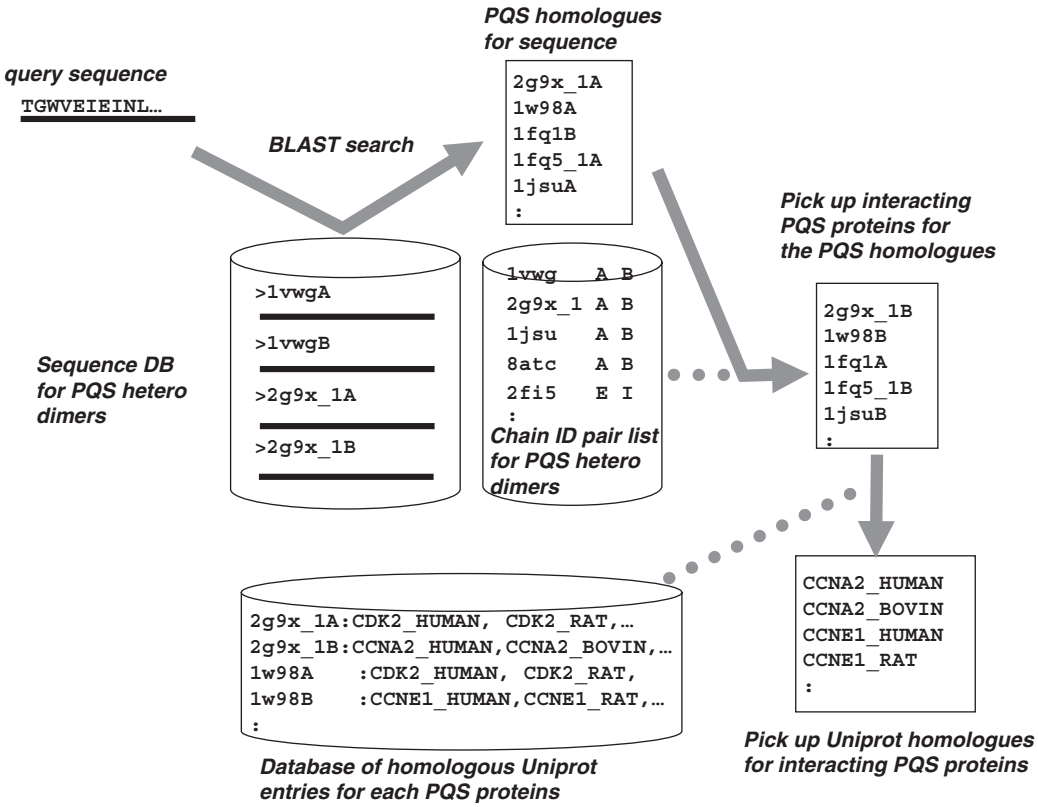


**Figure 3.** Overview of the procedures for identifying putative interacting proteins.

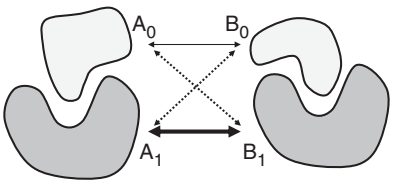


**Figure 4.** Definition of similarity between heterodimers for the representative heterodimer dataset. Similarity between a dimer of protein $A_0$ and $A_1$ and a dimer of protein $B_0$ and $B_1$ is defined as follows. The sequence similarities $S(A_0,B_0)$, $S(A_0,B_1)$, $S(A_1,B_0)$ and $S(A_1,B_1)$ are calculated. The value $S(A_i,B_j)$ is defined as the sequence similarity between protein $A_i$ and protein $B_j$. If $S(A_i,B_j)$ is the highest of the four similarity values, the corresponding pairs are $(A_i,B_j)$ and $(A_{i'},B_{j'})$ where $i' = (i + 1)\%2$ and $j' = (j + 1)\%2$. The sequence similarity between the two heterodimers is defined as the lower sequence similarity $S(A_{i'},B_{j'})$ of the two sequence similarities between corresponding proteins $S(A_i,B_j)$ and $S(A_{i'},B_{j'})$. For example, if $S(A_1,B_1)$ has the greatest value, the similarity between the dimer is $S(A_0,B_0)$.

greater than 50% were defined as homodimers. Using a single-linkage clustering algorithm (21), these dimers were clustered according to their sequence similarities. Sequence similarity was defined as the lower sequence of the two sequence similarities between corresponding proteins (described in Figure 4). Even if one protein of the dimer proteins is similar to a protein contained in another dimer, these dimers are considered to be different if the paring proteins are not similar. This is a reasonable definition, because several proteins, such as protease and immunoglobulin, exhibit a large number of dimer complex structures with different interacting proteins. For each cluster, the dimer protein with the largest number of interacting residues was chosen as the representative. We used the structural data from January 23, 2008 version of the PQS database with a threshold of 95% to define similar proteins. The heterodimer set contained 3305 dimers, while the homodimer set contained 8206 dimers.

## LIMITATIONS OF THE METHOD

Homology to a known 3D structure of a protein complex is a powerful tool to predict new interactions and their interacting sites. This methodology assumes that homologous protein pairs interact in a similar way. However, some exceptions have been reported. First, proteins belonging to multigene families often show different interaction specificities, even if their sequence similarity is high. A good example would be the interactions between Fibroblast Growth Factors (FGFs) and their receptors (6). The interaction specificities among many homologous protein pairs are biologically important, but difficult to be captured by our method even if the contact energy is employed. Users have to be aware of this limited accuracy of the predicted interaction specificity. Second, homologous interacting protein pairs sometimes show completely different interacting structural topologies. These different structural pairs of dimers mainly appear in a twilight zone of sequence similarity (< 30–40%) (22,23). Users have to be careful with our dimeric model based on

a remote-homologous template structures. This fact also indicates that our procedure of clustering dimer structures was not perfect, structural differences between homologous dimers should be considered in near future.

## CONCLUDING REMARKS

In comparison to homology modeling of a single protein, the modeling of complex structures has not been well studied. Only a few modeling servers for complex structures are currently working and available. The concept of the HOMCOS server is simple, but the updated dimer database and various output types for model complexes make the server useful for wide range of research needs. The complex structural models generated by our server can provide useful hypotheses to address the possible effects of natural or artificial mutation on protein–protein interactions, if users recognize the limited accuracies of the models. Putative interacting proteins identified by our server may be used as candidates to be confirmed experimentally. We plan to update the dimer database monthly and add a new service to model multimeric, not only binary complexes.

## ACKNOWLEDGEMENTS

We thank Mr Yuki Yoshii for designing the HOMCOS server logo. Mr Hiroyuki Miyakubo and Mr Junya Watanabe helped us test the server service. N. Fukuhara was supported by a Grand-in-Aid for the 21st Century COE Research from the Ministry of Education, Culture, Sports, Science and Technology of Japan. Funding to pay the Open Access publication charges for this article was provided by the Management Subsidy for Nara Institute of Science and Technology.

*Conflict of interest statement*. None declared.

## REFERENCES

1. Kleanthous,C. (ed.) (2000) *Protein-Protein Recognition*. Oxford University Press, Oxford.
2. von Mering,C., Krause,R., Snel,B., Cornell,M., Oliver,S.G., Fields,S. and Bork,P. (2002) Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, **417**, 399–403.
3. Deane,C.M., Salwinski,L., Xenarios,I. and Eisenberg,D. (2002) Protein interactions: two methods for assessment of the reliability of high throughput observations. *Mol. Cell. Prot.*, **1**, 349–356.
4. Sprinzak,E., Sattath,S. and Margalit,H. (2003) How reliable are experimental protein-protein interaction data? *J. Mol. Biol.*, **327**, 919–923.
5. Berman,H.M., Henrick,K. and Nakamura,H. (2003) Announcing the worldwide Protein Data Bank. *Nature Str. Biol.*, **10**, 980.
6. Aloy,P. and Russell,R.B. (2002) Interrogating protein interaction networks through structural biology. *Proc. Natl Acad. Sci. USA*, **99**, 5896–5901.
7. Lu,L., Lu,H. and Skolnick,J. (2002) MULTIPROSPECTOR: an algorithm for the prediction of protein-protein interactions by multimeric threading. *Proteins*, **49**, 350–364.
8. Davis,F.P., Braberg,H., Shen,M.Y., Pieper,U., Sali,A. and Madhusudhan,M.S. (2006) Protein complex compositions predicted by structural similarity. *Nucleic Acids Res.*, **34**, 2943–2952.
9. Cockell,S.J., Oliva,B. and Jackson,R.M. (2007) Structure-based evaluation of in silico predictions of protein-protein interactions using comparative docking. *Bioinformatics*, **23**, 573–581.
10. Cheng,Y.-C., Lo,Y.-S., Hsu,W.-C. and Yang,J.-M. (2007) 3D-partner: a web server to infer interacting partners and binding models. *Nucleic Acids Res.*, **35**, W561–W567.
11. Fukuhara,N., Go,N. and Kawabata,T. (2007) Prediction of interacting proteins from homology-modeled complex structures using sequence and structure scores. *Biophysics*, **3**, 13–26.
12. Grigoryan,G. and Keating,A.E. (2006) Structure-based prediction of bZIP partnering specificity. *J. Mol. Biol.*, **355**, 1125–1142.
13. Kiel,C., Wohlgemuth,S., Rousseau,F., Schymkowitz,J., F-Borg,J., Wittinghofer,F. and Serrano,L. (2005) Recognizing and defining true Ras binding domains II: in silico prediction based on homology modeling and energy calculations. *J. Mol. Biol.*, **348**, 759–775.
14. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
15. Aloy,P. and Russell,R.B. (2003) InterPreTS: protein interaction prediction through tertiary structure. *Bioinformatics*, **19**, 161–162.
16. Pieper,U., Eswar,N., Davis,F.P., Braberg,H., Madhusudhan,M.S., Rossi,A., Marti-Renom,M., Karchin,R., Webb,B.M., Eramian,D. *et al.* (2006) MODBASE: a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Res.*, **34**, D291–D295.
17. Sali,A. and Blundell,T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.*, **234**, 779–815.
18. Ispolatov,I., Yuryev,A., Mazo,I. and Maslov,S. (2005) Binding properties and evolution of homodimers in protein-protein interaction networks. *Nucleic Acids Res.*, **33**, 3629–3635.
19. Henrick,K. and Thornton,J.M. (1998) PQS: a protein quaternary structure file server. *Trends Biochem. Sci.*, **23**, 358–361.
20. Uniprot Consortium. (2008) The universal protein resource (UniProt). *Nucleic Acids Res.*, **25**, 3389–3402.
21. Johnson,R.A. and Wichern,D.W. (1998) *Applied Multivariate Statistical Analysis*. Prentice-Hall, London, p. 740.
22. Aloy,P., Ceulemans,H., Stark,A. and Russell,R.B. (2003) The relationship between sequence and interaction divergence in proteins. *J. Mol. Biol.*, **332**, 989–998.
23. Aloy,P., Pichaud,M. and Russell,R.B. (2005) Protein complexes: structure prediction challenges for 21st century. *Curr. Opin. Struct. Biol.*, **15**, 15–22.