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

Naoshi Fukuhara¹ and Takeshi Kawabata¹,²,*

¹Graduate School of Information Science, Nara Institute of Science and Technology, 8916-5 Takayama, Ikoma, Nara 630-0192 and ²CREST, 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://biounit.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 structures. 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...
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.
The procedures for the modeling of homodimers are similar to those for heterodimers. The HOMCOS server accepts only one query protein sequence 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
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 homol-
ogous 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, homol-
ogous interacting protein pairs sometimes show com-
pletely 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