*D360–D364* *Nucleic Acids Research, 2009, Vol. 37, Database issue* *Published online 4 October 2008*
*doi:10.1093/nar/gkn659*

# PiSite: a database of protein interaction sites using multiple binding states in the PDB

**Miho Higurashi[1], Takashi Ishida[1] and Kengo Kinoshita[1,2,]***

[1]Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639 and [2]Bioinformatics Research and Development, Japan Science and Technology Corporation, 4-1-8 Honcho, Kawaguchi, Saitama 332-0012, Japan

Received August 15, 2008; Revised September 19, 2008; Accepted September 21, 2008

## ABSTRACT

**The vast accumulation of protein structural data has now facilitated the observation of many different complexes in the PDB for the same protein. Therefore, a single protein complex is not sufficient to identify their interaction sites, especially for proteins with multiple binding states or different partners, such as hub proteins. PiSite is a database that provides protein–protein interaction sites at the residue level with consideration of multiple complexes at the same time, by mapping the binding sites of all complexes containing the same protein in the PDB. PiSite provides easy web interfaces with an interactive viewer working with typical web browsers, and the different binding modes can be checked visually. All of the information can also be downloaded for further analyses. In addition, PiSite provides a list of proteins with multiple binding partners and multiple binding states, as well as up-to-date statistics of protein–protein interfaces. PiSite is available at http://pisite.hgc.jp**

## INTRODUCTION

Protein–protein interactions are fundamental for proteins to exert their biological functions, and the molecular interactions of proteins can be understood by observing the 3D structure of their complexes. Therefore, many efforts have been made to solve complex structures experimentally, and a large number of structures are now available in the Protein Data Bank (PDB) (1). As a result of the vast accumulation of structural data from several genomic projects (2), we can now estimate the structural changes of proteins (3), the evolution of homo-oligomerization states (4,5) and the changes in protein–protein interactions (6) by analyzing all of the structures in the PDB simultaneously.

The structural characteristics of protein–protein interaction sites have been extensively studied (7,8), and the knowledge has been used for the prediction of binding sites (9–12). For the statistical analyses, one representative is usually selected for each group with similar amino acid sequences. However, some proteins, called hub proteins (13), interact with several kinds of different proteins, and thus one representative complex is not sufficient to describe all of the interfaces of a hub protein. For the identification of all of the binding sites of a hub protein, all of the complexes within the PDB should be considered at the same time.

Protein structures often consist of structural domains, and protein interactions can sometimes be interpreted as domain–domain interactions (14). Therefore, many databases to describe domain–domain interactions have been developed [3did (15), DIMA2 (15) and DOMINE (16)], and networks of domain interactions have been constructed. In these databases, large-scale interaction networks are the main focus, and the residue-level interactions are not described. On the other hand, for better understanding the molecular interactions of protein domains, SCOPPI (17) and iPFAM (18) provide residue-level information about the interacting domains, but the interacting residues are provided for each pair of interacting domains (iPfam) or a multiple alignment in the family with interface classification (SCOPPI), and thus it is difficult to observe several different binding modes of proteins with multiple partners. In addition, SCOPPI focuses on the diversity at the family level, while we would like to observe the multiple binding modes for individual proteins. In the similar way, PiBase (19) provides detailed physicochemical properties for all the protein complexes in the PDB as a correction of binary relationships.

In this article, we describe a new protein–protein interaction database, PiSite, based on the protein complexes in

*To whom correspondence should be addressed. Tel: +81 3 5449 5131; Fax: +81 3 5449 5133; Email: kino@ims.u-tokyo.ac.jp

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

the PDB. We tried to provide a simple user interface to observe the 'real' binding sites, by simultaneously considering multiple binding states of individual proteins at the residue level, and not just for protein domains but for entire protein chains. PiSite also provides a list of 'sociable proteins', proteins with multiple binding states and multiple binding partners, which are considered as the key molecules in protein interaction networks. It should be noted that the sociable proteins are somewhat different from the so-called hub proteins, because the so-called hub proteins, which are usually defined as proteins with multiple binding partners in a protein interaction network obtained by large-scale experiments (20–23), are sometimes the subunits of a supermolecule (6). Furthermore, PiSite provides up-to-date statistics of protein interfaces. PiSite will be periodically updated every 3 months, using the most recent version of the PDB.

## CONSTRUCTION METHOD

### Data set

The current version of PiSite was constructed from the PDB entries as of July 2008. We did not use the asymmetric unit of the coordinates, but employed the biological units distributed by RCSB, to eliminate the crystallographic interfaces (ftp://ftp.rcsb.org/pub/pdb/data/biounit/coordinates). All protein chains with more than 30 residues were considered as proteins and we excluded the entries with >5.0 Å resolution and the models with only Cα coordinates. As a result, we selected 110 325 protein chains from the 51 482 PDB entries in our data set.

### Mapping

The binding sites that appeared in the PDB were gathered by mapping from all complexes to each protein chain (Fig. 1). For this purpose, at first, a similarity search by BLAST (24) against all protein chains in the data set was carried out for each protein chain. Ideally, the exact match of the amino acid sequence may be sufficient to find other complex structures in the PDB, but to enlarge the complex information and to avoid minor errors in sequence records due to missing residues and/or modifications, the proteins with sequence identity >90% and with coverage of the smaller protein >80% were used to select the entries for mapping. Then, mapping of the binding sites from one complex to a query chain was performed according to the BLAST alignment. The identification of the binding site residues was achieved by the distance criteria: when the minimum distance between the atoms in a residue pair was <4.0 Å, then the pair of residues was defined as contact residues. If the number of contacting residues between a pair of proteins was less than two, then the pair of protein chains was not used for mapping. It should be noted that we sometimes refer to the similar proteins used for the mapping as similar proteins for simplicity, but our focus is not to analyze the interfaces among the family members, in contrast to other protein–protein databases.

### Definition of the binding state

By selecting similar entries, as mentioned, we can enumerate all complexes containing the protein chain being considered, and can identify all of the binding partners that interact with the considering protein chain. The binding

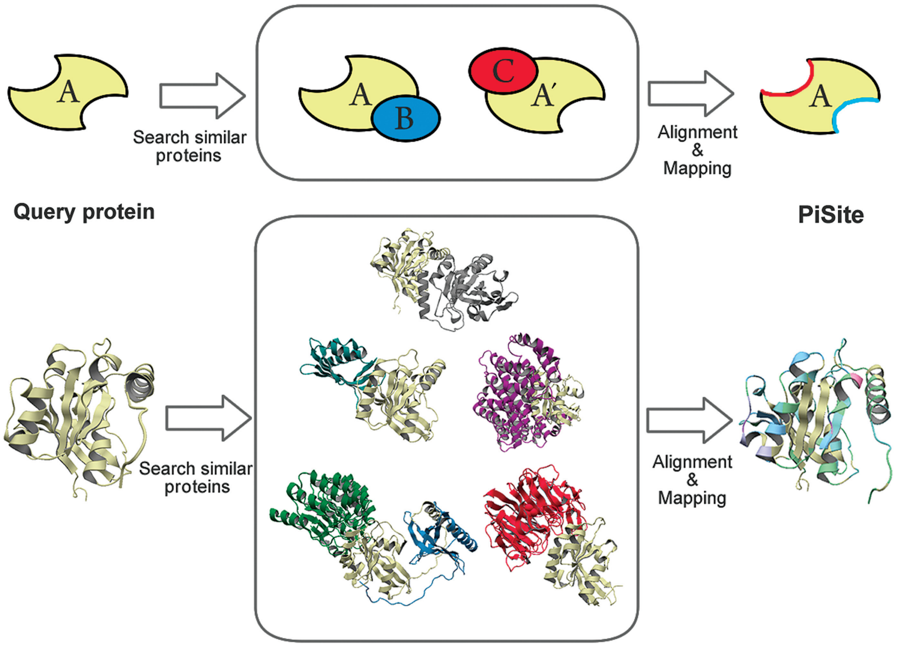


**Figure 1.** An explanation of residue mapping. The upper panel shows a schematic representation of residue mapping, and the lower panel shows an example of the mapping by using 3D models. In the example, the GTP binding protein RAN (PDB: 1byu, chain A) was used. The gray, light blue, purple, green, blue and red chains were taken from 1byuB, 1a2kD, 1ibrB, 1k5gKL and 1l1mB, respectively (the four letter code indicates the PDB ID and the fifth letter means the chain ID).

partners were grouped by sequence identity, and the number of groups was used as the number of binding partners shown in PiSite. Here, we regarded two proteins as being in the same group, if they have >30% sequence identity and >50% coverage of the smaller chain. Each binding state of a protein was defined as the combination and the number of binding partners appeared in the PDB. For example, when we consider the binding state of protein A with the complexes A–B and A–B–C in the PDB, then the number of binding states of protein A is two. If we have another complex A–B–B, then the number of binding states is three, because the number of binding partner B is different from that in the complex A–B. It should be noted that the similarities between binding sites are not considered explicitly in our approach, thus the number of binding state can be smaller than that of the binding modes, while we used all complexes for the residue mapping to get the comprehensive mapping.

The definitions of binding states and binding partners are virtually same as those used in Higurashi *et al.* (6), but we modified the original protocol to handle all PDB entries in PiSite, as described.

### Definition of sociable proteins

We basically followed the sociable protein definition by Higurashi *et al.* (6). However, since automated processes are required in this study and we could not apply manual curation, we used a stricter definition. We defined proteins with three or more binding states and three or more binding partners as sociable proteins. We excluded proteins with more than 10 chains in a single PDB entry as supermolecules. In addition, we also excluded proteins with four or fewer similar proteins in the data set. This last condition is to ensure the reliability of sociable protein identifications. If the number of similar proteins is too small, then this definition may contain some errors.

## CONTENTS OF PiSite

The main content of the PiSite is an interactive view of the multiple binding states, and we refer to the corresponding web page as the interaction viewer. The interaction viewer (Fig. 2) shows all binding states appearing in the PDB, and is prepared for each protein chain. It consists of four parts: title table including a link to UniprotKB (25), a molecule viewer, a binding state viewer and a download menu. The title table contains brief descriptions of the protein chain and the numbers of similar proteins, binding states and binding partners.

The molecule viewer shows the protein, colored according to the number of binding partners for each residue in default. The amino acid sequence is also colored according to the same procedure, and is shown just below the viewer. The position of each amino acid can be checked by clicking a residue in the sequence, which changes the specified residue into a CPK model (Fig. 2). Visualization of the molecule was done by jV [formally known as pdbjviewer (26)], and thus the view of the molecule can be interactively rotated and translated by mouse operations, and the residue position can be inspected by clicking the residue on the screen. More detailed options for jV are also available, by opening the option screen at the top of the molecule viewer (Fig. 2).

The binding state viewer shows a different binding state of the proteins, by communicating with the molecule viewer. By switching to a different binding state with the radio button, the complex structure is shown in the molecule viewer with the same color as the background of the partner name. It may be noteworthy that two or more protein names appear as binding partners as in the cases of the 9th and 10th binding states in Fig. 2, which means that the binding state contains three or more chains, including the chain under consideration. The names of the binding partners are taken from the DBREF record in the PDB entry. All of the data obtained after mapping the similar proteins can be downloaded in a flat file or XML format. The format is described in the download page.

## OTHER CONTENTS

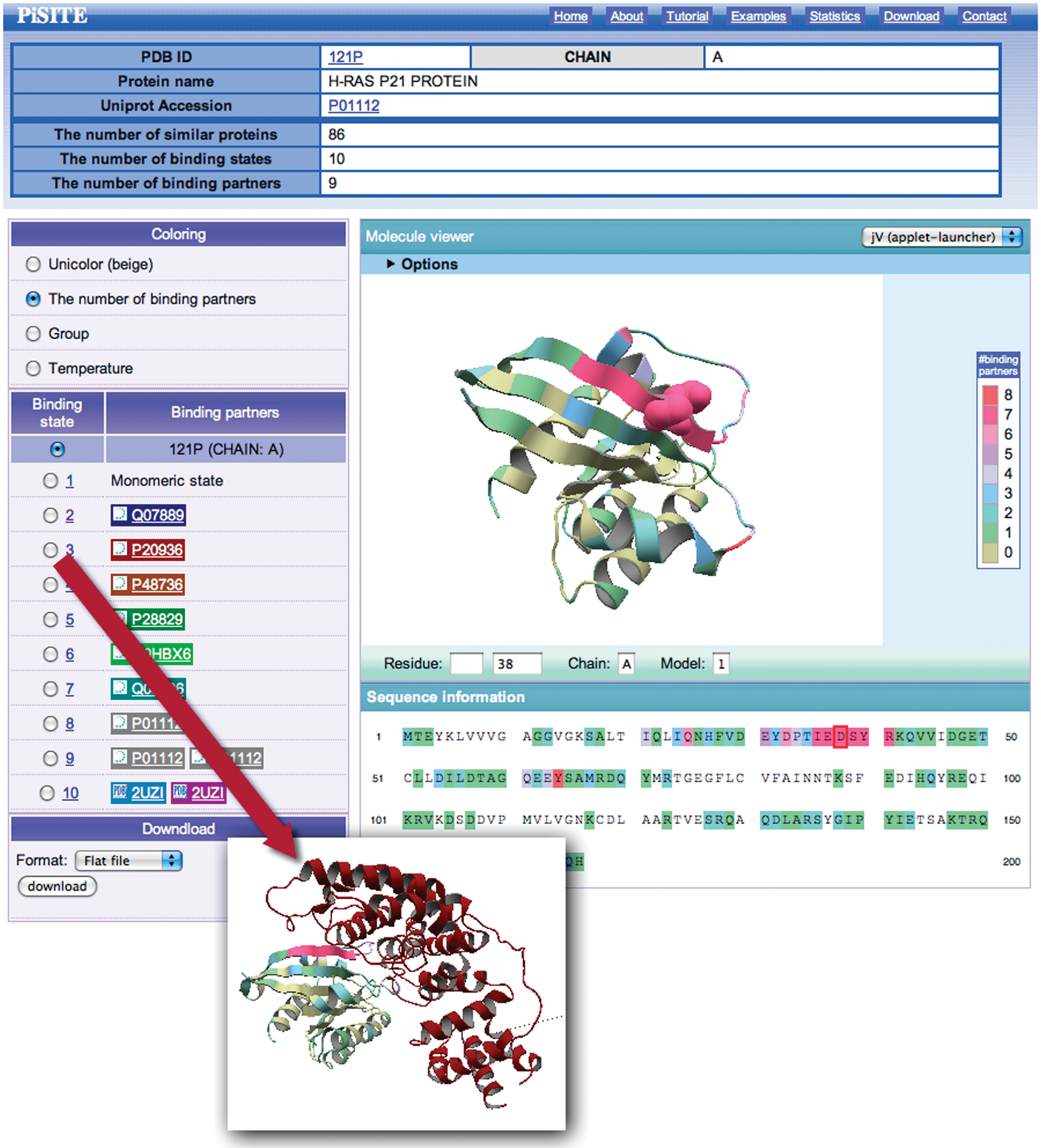
PiSite also provides a list of sociable proteins, which are proteins with multiple binding states and multiple binding partners. In our original paper, the number of sociable proteins was 86 (6), where only the structures with a resolution of 3.0 Å or better were selected from the PDB as of July 2006. In contrast, the number of nonredundant sociable proteins in the current release of PiSite is 102, where the redundancies were eliminated by selecting one representative from 30% identity sequence clusters. The difference arises mainly from the date of the PDB and the inclusion of NMR data in PiSite.

On the statistics page, we also show some statistical analyses of amino acid occurrence in the protein–protein interfaces based on the latest release of PiSite, which will be useful to develop statistical parameters to evaluate the likelihood of protein–protein interfaces.

All of the raw data are also available from the statistics page for further analyses.

## ACCESS TO EACH ENTRY, WITH AN EXAMPLE

PiSite provides two different ways to access each entry. The first way is access from a compiled list. PiSite provides a list of representative sociable proteins and supermolecules, as mentioned above. These lists contain links for the entries and the user can access them by clicking the links. The other way is to search by sequence or keywords. Both sequence and keyword searches are available, through the search form on the top page. A sequence search requires amino acid sequence of a protein chain as an input and is performed by using BLAST. A more general search is by keywords. If the user wants to check the PiSite entry for Ras p21 proteins, for example, then the user can search ‘RAS p21’ as a keyword. As a result, the search results, including a moleculer description of the target and information about the number of binding states and partners are obtained. Each PiSite entry can be accessed from links in a search result page. In this case, the user can access the PiSite entry of the hRas P21 protein (PDBID: 121p, chain A) by clicking the link of the top hit of the search



| | | | |
|---|---|---|---|
| PDB ID | 121P | CHAIN | A |
| Protein name | H-RAS P21 PROTEIN | | |
| Uniprot Accession | P01112 | | |
| The number of similar proteins | 86 | | |
| The number of binding states | 10 | | |
| The number of binding partners | 9 | | |

Coloring

○ Unicolor (beige)

● The number of binding partners

○ Group

○ Temperature

| Binding state | Binding partners |
|---|---|
| ● | 121P (CHAIN: A) |
| ○ 1 | Monomeric state |
| ○ 2 | Q07889 |
| ○ 3 | P20936 |
| ○ | P48736 |
| ○ 5 | P28829 |
| ○ 6 | HBX6 |
| ○ 7 | Q0 6 |
| ○ 8 | P0111 |
| ○ 9 | P01112 1112 |
| ○ 10 | 2UZI 2UZI |

Downdload

Format: Flat file

download

Residue: 38 Chain: A Model: 1

Sequence information

```
1    MTEYKLVVVG AGGVGKSALT IQLIQNHFVD EYDPTIEDSY RKQVVIDGET   50
51   CLLDILDTAG QEEYSAMRDQ YMRTGEGFLC VFAINNTKSF EDIHQYREQI  100
101  KRVKDSDDVP MVLVGNKCDL AARTVESRQA QDLARSYGIP YIETSAKTRQ  150
     QH                                                     200
```

**Figure 2.** An example of a PiSite entry. See the main text for details.

results (Fig. 2). If the PDB ID of a target protein is already known, then rapid search can be available selecting the ‘search by PDB ID’ radio button. The interaction viewer shows that the protein has 10 different binding states and 9 different binding partners. The molecule viewer shows the number of binding partners of each residue, and the residues contacting many partners are colored magenta. Interestingly, this magenta-colored region, including residues 32–42, corresponds to the effector region of a Ras protein. This region of a Ras protein is known as a hot spot and has been identified as a neutralizing epitope of hRas (27). More mapping result details can be obtained from the download menu.

## FUNDING

Institute for Bioinformatics Research and Development; Japan Science and Technology Corporation (BIRD-JST);

Scientific Research on Priority Areas from the Ministry of Education, Culture, Sports, Science and Technology of Japan (to K.K.). Funding for open access charge: Annual budget for Human Genome Center.

*Conflict of interest statement*. None declared.

## REFERENCES

1. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.
2. Chandonia,J.M. and Brenner,S.E. (2006) The impact of structural genomics: expectations and outcomes. *Science*, **311**, 347–351.
3. Koike,R., Amemiya,T., Ota,M. and Kidera,A. (2008) Protein structural change upon ligand binding correlates with enzymatic reaction mechanism. *J. Mol. Biol.*, **379**, 397–401.
4. Levy,E.D., Boeri Erba,E., Robinson,C.V. and Teichmann,S.A. (2008) Assembly reflects evolution of protein complexes. *Nature*, **453**, 1262–1265.
5. Levy,E.D., Pereira-Leal,J.B., Chothia,C. and Teichmann,S.A. (2006) 3D complex: a structural classification of protein complexes. *PLoS Comp. Biol.*, **2**, e155.
6. Higurashi,M., Ishida,T. and Kinoshita,K. (2008) Identification of transient hub proteins and the possible structural basis for their multiple interactions. *Protein Sci.*, **17**, 72–78.
7. Jones,S. and Thornton,J.M. (1995) Protein-protein interactions: a review of protein dimer structures. *Prog. Biophys. Mol. Biol.*, **63**, 31–65.
8. Keskin,O., Gursoy,A., Ma,B. and Nussinov,R. (2008) ) Principles of Protein-protein interactions: what are the preferred ways for proteins to interact? *Chem. Rev.*, **108**, 1225–1244.
9. Huang,B. and Schroeder,M. (2008) Using protein binding site prediction to improve protein docking. *Gene*, **422**, 14–21.
10. Zhou,H.-X. and Qin,S. (2007) Interaction-site prediction for protein complexes: a critical assessment. *Bioinformatics*, **23**, 2203–2209.
11. Valdar,W.S. and Thornton,J.M. (2001) Protein-protein interfaces: analysis of amino acid conservation in homodimers. *Proteins*, **42**, 108–124.
12. Bahadur,R.P., Chakrabarti,P., Rodier,F. and Janin,J. (2004) A dissection of specific and non-specific protein-protein interfaces. *J. Mol. Biol.*, **336**, 943–955.
13. Jeong,H., Mason,S.P., Barabasi,A.L. and Oltvai,Z.N. (2001) Lethality and centrality in protein networks. *Nature*, **411**, 41–42.
14. Kim,W., Bolser,D. and Park,J. (2004) Large-scale co-evolution analysis of protein structural interlogues using the global protein structural interactome map (PSIMAP). *Bioinformatics*, **20**, 1138–1150.
15. Stein,A., Russell,R.B. and Aloy,P. (2005) 3did: interacting protein domains of known three-dimensional structure. *Nucleic Acid Res.*, **33**, D413–D417.
16. Raghavachari,B., Tasneem,A., Przytycka,T.M. and Jothi,R. (2007) DOMINE: a database of protein domain interactions. *Nucleic Acids Res.*, **36**, 656–661.
17. Winter,C., Henschel,A., Kim,W.K. and Schroeder,M. (2006) SCOPPI: a structural classification of protein-protein interfaces. *Nucleic Acids Res.*, **34**, D310–D314.
18. Finn,R., Marshall,M. and Bateman,A. (2005) iPfam: visualization of protein-protein interactions in PDB at domain and amino acid resolutions. *Bioinformatics*, **21**, 410–412.
19. Davis,F.P. and Sali,A. (2005) PIBASE: a comprehensive database of structurally defined protein interfaces. *Bioinformatics*, **21**, 1901–1907.
20. Ito,T., Chiba,T., Ozawa,R., Yoshida,M., Hattori,M. and Sakaki,Y. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl Acad. Sci. USA*, **98**, 4569–4574.
21. Uetz,P., Giot,L., Cagney,G., Mansfield,T.A., Judson,R.S., Knight,J.R., Lockshon,D., Narayan,V., Srinivasan,M., Pochart,P. *et al.* (2000) A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature*, **403**, 623–627.
22. Gavin,A.C., Bosche,M., Krause,R., Grandi,P., Marzioch,M., Bauer,A., Schultz,J., Rick,J.M., Michon,A.M., Cruciat,C.M. *et al.* (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**, 141–147.
23. Ho,Y., Gruhler,A., Heilbut,A., Bader,G.D., Moore,L., Adams,S.L., Millar,A., Taylor,P., Bennett,K., Boutilier,K. *et al.* (2002) Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature*, **415**, 180–183.
24. 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.
25. UnitProt Consortium. (2008) The universal protein resource (UniProt). *Nucleic Acids Res.*, **36**, D190–D195.
26. Kinoshita,K. and Nakamura,H. (2004) eF-site and PDBjViewer: database and viewer for protein functional sites. *Bioinformatics*, **20**, 1329–1330.
27. Sigal,I.S., Gibbs,J.B., D'Alonzo,J.S. and Scolnick,E.M. (1986) Identification of effector residues and a neutralizing epitope of Ha-ras-encoded p21. *Proc. Natl Acad. Sci. USA*, **83**, 4725–4729.