Can bioinformatics help in the identification of moonlighting proteins?

Sergio Hernández*, Alejandra Calvo†, Gabriela Ferragut†, Luís Franco*, Antoni Hermoso*[1], Isaac Amela*, Antonio Gómez‡, Enrique Querol* and Juan Cedano†[2]

*Institut de Biotecnologia i Biomedicina and Departament de Bioquímica i Biologia Molecular, Universitat Autònoma de Barcelona, 08193 Cerdanyola del Vallès, Barcelona, Spain

†Laboratorio de Inmunología, Universidad de la República Regional Norte-Salto, Rivera 1350, CP 50000 Salto, Uruguay

‡Cancer Epigenetics and Biology Program (PEBC), Institut d'Investigació Biomèdica de Bellvitge (IDIBeLL), L'Hospitalet de Llobregat, 08908 Barcelona, Spain

Abstract

Protein multitasking or moonlighting is the capability of certain proteins to execute two or more unique biological functions. This ability to perform moonlighting functions helps us to understand one of the ways used by cells to perform many complex functions with a limited number of genes. Usually, moonlighting proteins are revealed experimentally by serendipity, and the proteins described probably represent just the tip of the iceberg. It would be helpful if bioinformatics could predict protein multifunctionality, especially because of the large amounts of sequences coming from genome projects. In the present article, we describe several approaches that use sequences, structures, interactomics and current bioinformatics algorithms and programs to try to overcome this problem. The sequence analysis has been performed: (i) by remote homology searches using PSI-BLAST, (ii) by the detection of functional motifs, and (iii) by the co-evolutionary relationship between amino acids. Programs designed to identify functional motifs/domains are basically oriented to detect the main function, but usually fail in the detection of secondary ones. Remote homology searches such as PSI-BLAST seem to be more versatile in this task, and it is a good complement for the information obtained from protein–protein interaction (PPI) databases. Structural information and mutation correlation analysis can help us to map the functional sites. Mutation correlation analysis can be used only in very restricted situations, but can suggest how the evolutionary process of the acquisition of the second function took place.

Introduction

Multitasking or moonlighting proteins refer to those proteins in which two or more distinct biological functions are performed by the same single polypeptide chain. Moonlighting proteins present alternative functions which are mostly related to cellular localization, cell type, oligomeric state, concentration of cellular ligands, substrates, cofactors, products or post-translational modifications [1–10]. In many cases, a protein uses a combination of these mechanisms to switch between functions. Although some findings suggest involvement of a protein in extra functions, i.e. they can be found in different cellular localizations or in amounts exceeding those required for their canonical function, moonlighting proteins are usually revealed experimentally by serendipity. Therefore any alternative method to identify moonlighting proteins would be valuable. During the development of our previous work, aimed at trying to find bioinformatics approaches to predict multitasking proteins, we encountered the difficulty of collecting enough examples of such proteins because of the lack of a broad database, thus we have collected and compiled a database of multifunctional proteins: MultitaskProtDB [11].

In previous studies, we have explored the possibility of identifying moonlighting proteins using bioinformatics approaches [12,13]. Protein–protein interaction (PPI) databases should contain information on moonlighting proteins and could provide suggestions for further analysis in order to prove the multifunctionality [13,14]. It is generally considered that experimental data from proteomics contain many false positives, estimated to be up to ~20%, which may induce proteomics researchers to consider most of the unexpected partners as false positives. We have screened different algorithms using the MultitaskProtDB database as a benchmark and suggest the best combination of programs to predict multitasking proteins. However, although bioinformatics analyses can help to disclose multifunctional proteins for the moment, true moonlighting proteins need to be verified experimentally.

Databases

The database of multifunctional proteins, MultitaskProtDB [11], is accessible at http://wallace.uab.es/multitask/.

Key words: bioinformatics, moonlighting protein, multitasking protein.

Abbreviations: PPI, protein–protein interaction.

[1] Present address: Bioinformatics Core Facility, Centre for Genomic Regulation (CRG) and Universitat Pompeu Fabra (UPF), Av. Dr Aigüader 88, 08003 Barcelona, Spain.

[2] To whom correspondence should be addressed (email jcedano@unorte.edu.uy).

Biochem. Soc. Trans. (2014) 42, 1692–1697; doi:10.1042/BST20140241

Table 1 | **Examples of moonlighting protein prediction combining PPI databases and BYPASS**

Canonical function	Moonlighting function	PPI partners (only some are shown)	BYPASS output (only some are shown)
Phosphoglucose isomerase	Neurotrophic factor; neuroleukin; autocrine motility factor; nerve growth factor	GO:4842, autocrine motility factor receptor 2; GO:31994, insulin-like growth factor-binding protein 3	gi|17380385; glucose-6-phosphate isomerase; autocrine motility factor; neuroleukin
Pyruvate kinase	Thyroid hormone-binding protein	GO:3707, nuclear hormone receptor member nhr-111; GO:9914 sex hormone-binding globulin; GO:5179, atrial natriuretic factor	gi|20178296; pyruvate kinase isoenzymes; cytosolic thyroid hormone-binding protein
Ribosomal protein S3 (human)	Apurinic/apyrimidinic (AP) endonuclease	GO:31571, DNA damage-binding protein 1; GO:3735, S27 ribosomal protein	gi|290275; ribosomal protein S3; AP endonuclease DNA repair
Ure2	Glutathione peroxidase	GO:6808, nitrogen regulatory protein	gi|173152; gi449015276; GST-like protein; nitrogen catabolite repression transcriptional regulator

PPI partners for the moonlighting proteins have been checked in the APID (Agile Protein Interaction DataAnalyzer) server [15] at http://bioinfow.dep.usal.es/apid/index.htm. We have considered that proteomics data disclose the second function of a moonlighting protein if the PPI database identifies a molecular function or, in some cases, a biological process according to the Gene Ontology annotation (http://www.geneontology.org) upon an ontology-enrichment analysis using the GOStat R package [16] as reported previously [13].

Protein sequence analyses

Remote homology analysis on a non-redundant database was carried out using PSI-BLAST [17] (http://www.ncbi.nlm.nih.gov/BLAST). The search was performed with default settings with a maximum of five iterations. Up to 100 hits with *E*-value scores better than 0.01 and with functional annotation were inspected in order to check whether they show canonical and moonlighting annotations, the last in any position of the final output. Because true biological hits can be found, not in the top positions, but in lower positions, we have rearranged the BLAST/PSI-BLAST output by means of the BYPASS fuzzy logic program at http://bypass.uab.cat/wiki/ [18].

Motif and domain screening was performed using InterPro [19], accessible at http://www.ebi.ac.uk/Tools/pfa/iprscan/. Blocks [20] (http://blocks.fhcrc.org), a database of protein alignment blocks, was also used. In theory, it is more advisable to use InterPro because Blocks has not been updated since 2006. However, we considered it an interesting task to carry out searches with Blocks because this tool is a non-curated database that can disclose several signatures related to more than one function (which is a characteristic of moonlighting proteins), whereas InterPro databases have been programmed for an accurate identification of the major motif/domain.

Mapping the structural/functional sites on the protein sequence from homology to 3D structures

To check whether the main structural/functional sites for both functions can be disclosed from the protein sequence, we have used PiSITE [21] (http://pisite.hgc.jp/), a program for mapping them. PiSITE works by aligning the query protein sequence with those present in the PDB.

Interactomics database searches

We have proposed previously that PPI databases should contain information on moonlighting proteins and provide suggestions for further analyses in order to prove their multitasking properties [13]. Column 3 of Table 1 shows some examples of disclosing the moonlighting function. Nevertheless, the number of interaction partners found in the PPI databases can be high. Therefore to pick the true partners is not an easy task if the researcher has no additional hints. We have found that, by crossing PPI database information and remote homology searches, the accuracy of the results is improved (see Table 1).

Remote homology searches

The remote homology algorithm PSI-BLAST identifies stretches of amino acid residues from different domains, therefore it is suitable to disclose moonlighting proteins [12,22]. As in the PPI database searches, the output depicts a large list of hits and the researcher does not know *a priori* which of them will be true positives, and it is the careful analysis of the different predictions and the experimental data that can suggest a true hit. Column 4 of Table 1 shows some examples of moonlighting proteins identified by remote homology. From the 288 moonlighting proteins of the MultitaskProtDB database, PSI-BLAST identifies 42 % of them. However, only 8 % of the moonlighting proteins of the database are identified by PSI-BLAST and InterPro at the same time.

Combining interactomics database and PSI-BLAST/BYPASS searches

Some 250 proteins from the MultitaskProtDB database reported partners in the APID PPI server. As stated above, each moonlighting protein can present a large set of putative PPI partners and also a large output of putative remote homologues from the PSI-BLAST algorithm. We have manually inspected both outputs to check whether those proteins present in both lists match the canonical and moonlighting functions of the query protein in order to narrow the number of candidate hits. This careful inspection has been necessary because there is a problem related to the different annotation descriptors depicted by both outputs. Most BLAST/PSI-BLAST outputs from sequence alignments do not report semantic curated annotations, whereas many PPI databases usually have GO annotations. This fact complicates the automatic matching of the outputs. We suggest to take as putative positive matches those describing a function in any position of the PSI-BLAST/BYPASS output that matches with a PPI database partner, as shown in the examples in Table 1. To design a program able to automatically match both outputs would be very useful. Moreover, current GO annotations contain only ∼9500 molecular functions, so most functional annotations from sequence databases cannot be found at GO. Therefore, at present, it is advisable to perform a manual inspection of the outputs.

To combine the outputs of interaction partners and remote homologues from PSI-BLAST is the best approach for reducing the sometimes large outputs from both servers and to improve the bioinformatics prediction of putative moonlighting proteins. Upon overlapping PSI-BLAST and interaction databases, ∼50 % of the moonlighting proteins from our database can be predicted. Table 1 shows some examples of proteins whose moonlighting function is disclosed by the partners found by interactomics and also using the PSI-BLAST algorithm.

Sequence searches using motif/domain programs

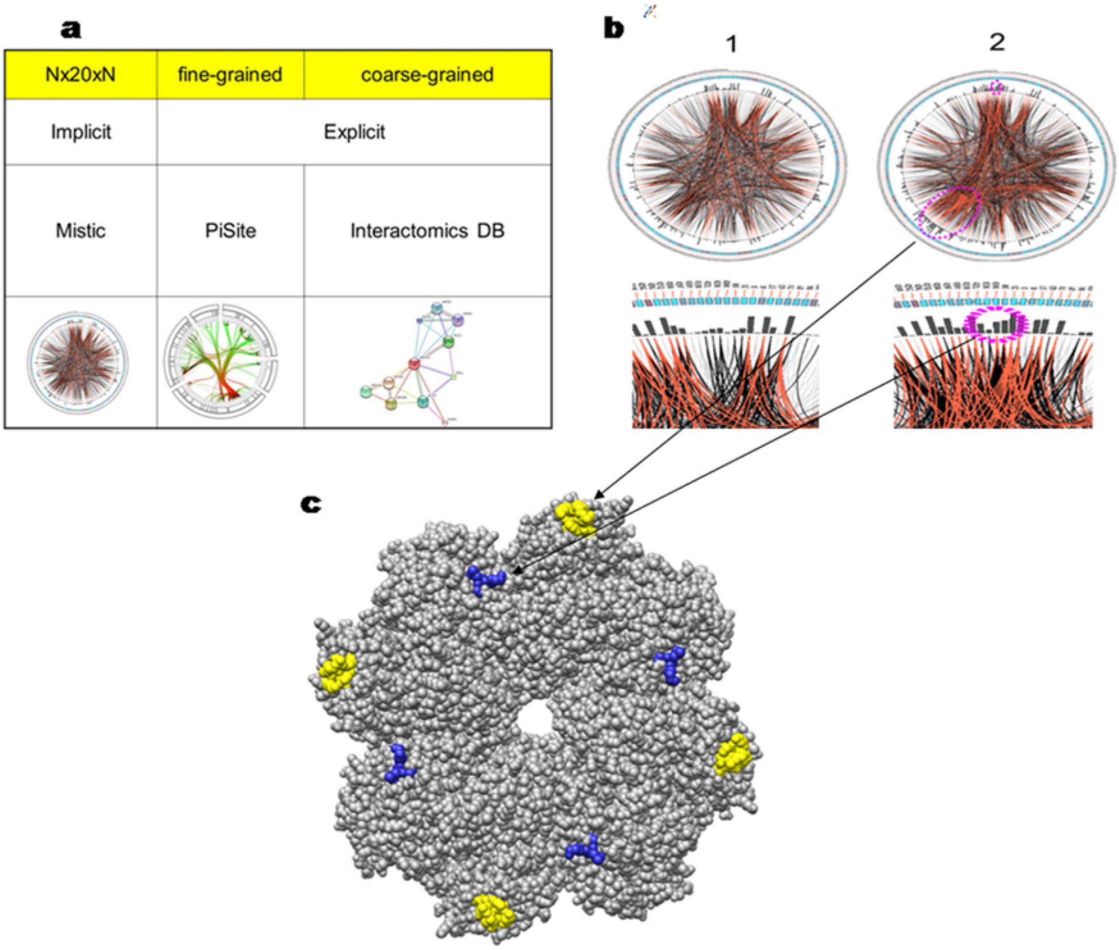
Searching for different motifs/domains linked to each different function in a target protein sequence using InterPro can help to identify moonlighting proteins. However, there are two main problems: (i) the relatively low number of domains and signatures currently known, and (ii) the current version of InterPro programs such as PROSITE have been programmed for a more accurate prediction of the major motif/domain, but loses less scored signatures. This would explain the fact that, on using InterPro on the MultitaskProtDB proteins, it discloses the canonical function in 64 % of them, but the moonlighting function in only 8 % of cases. These are also the few cases in which PSI-BLAST and InterPro identify both functions. For instance, a classical moonlighting example is glyceraldehyde-3-phosphate dehydrogenase (GAPDH)/uracil DNA glycosylase (UDG), in which the PSI-BLAST output discloses both functions with high scores, but InterPro fails and only identifies a motif for this protein. However, they are identified by Blocks. The fact that pattern detection of secondary function using obsolete systems such as Blocks is better than using more modern and refined ones made us think that this phenomenon may be due to a problem between sensitivity and specificity. Pattern detection tools traditionally have been developed to have a good ratio between specificity and sensitivity. When a gold-standard dataset is built to train these applications, it is usually assumed that all of the proteins included in the database have only a unique function. Therefore, if this assumption is not true, as is the case for multifunctional proteins, we are actually making a bias in terms of loss of sensitivity, so that the tools tend to detect a low number of secondary functions. In this sense, the trend of using very curated seeds of sequences to build these patterns could explain why obsolete tools such as Blocks are more effective to detect secondary functions. If so, this would indicate that to detect such secondary functions of proteins, tools such as BLAST or PSI-BLAST can be appropriate, because they are not dependent on the pre-existence of previously constructed patterns with a limited seed.

Mutation correlation analyses

Co-evolution studies of catalytic amino acids, also termed mutation correlation networks, have been used to predict key catalytic residues of enzymes. We have checked whether the MISTIC server [23] can help to predict moonlighting proteins. The main limitation of algorithms such as MISTIC is that they require a large multi-alignment; however, current families of moonlighting proteins are scarce, with enolases being the best example. We have analysed the correlation matrix of the amino acids of enolase with the extra function of binding to plasminogen, including them in a set of enolases with less than 35 % identity. At the same time, we have compared the same set of enolases, removing all those that bind plasminogen, except for the one used as a sequence reference. As shown in Figures 1(b) and 1(c), the correlation is lost in the C-terminal region, a major binding site for plasminogen, owing to the interaction introduced by the architecture that distorts the previous net of dependencies among the amino acids. Still, some distortions are also spread around position 250, another region involved in the interaction. This region is clearly flanking the loop that interacts with plasminogen. That is, the acquisition of new functions seems not to be confined to amino acids normally

Figure 1 | Some examples of the tools used for moonlighting protein prediction

(**a**) Interactomics analyses. (**b**) Correlations of the enolase muti-alignment with (1), considering all proteins and (2), without moonlighting proteins. (**c**) Comparison of putative plasminogen-binding sites. The functional sites of plasminogen are shown in blue and yellow according to [24].



associated with the binding pattern, but it may involve more global changes in the protein.

Mapping the structural/functional sites on the protein sequence from homology to 3D structures

We have used PiSITE [21] with the MultitaskProtDB database. The main limitation of this program is that it requires that the query protein, or a domain from it, must have a significant similar amino acid stretch in the PDB. Nevertheless the program has identified 266 PDB hits and, in addition to the canonical function, it identifies the moonlighting function for 28 proteins. Figure 2 shows an example of a successful match for fatty acid multifunctional protein (MFP) where PiSITE identifies both canonical and moonlighting functions. Moreover, using the SwissPDBViewer tool, both functions can be structurally mapped with a good RMSD. PiSITE alone cannot disclose as many multifunctional proteins as overlapping PSI-BLAST and PPI databases, but it can support putative hits as true positives upon running those programs and, interestingly, suggest a location of the moonlighting function.

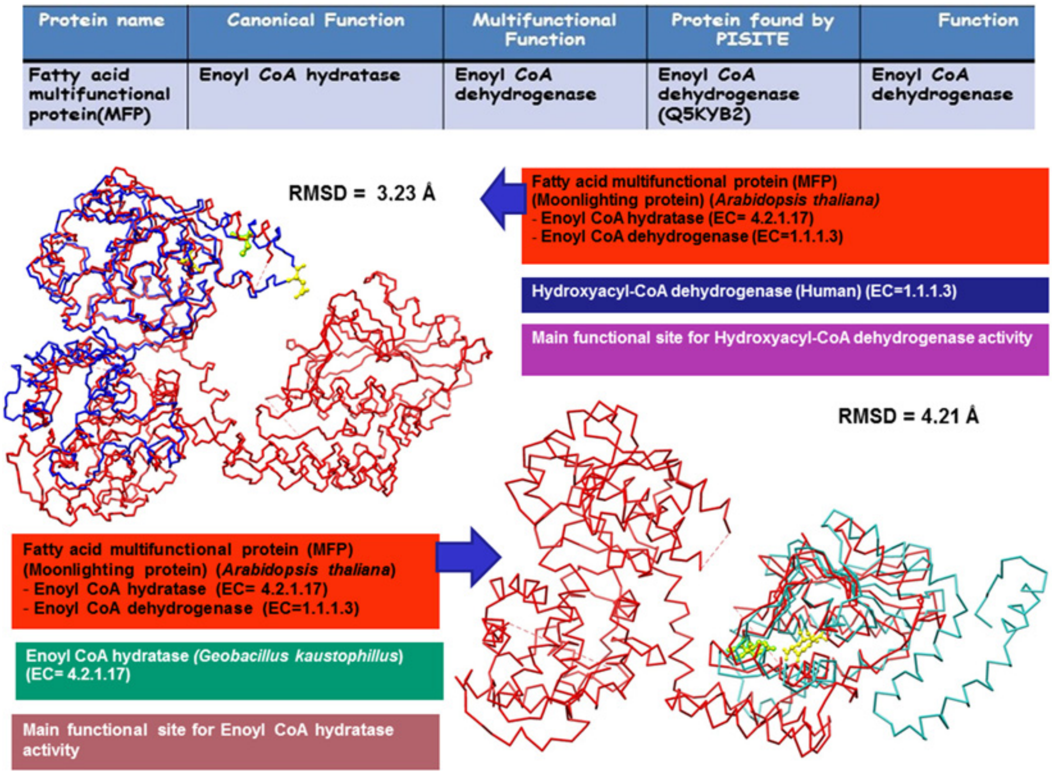
Discussion

Prediction of moonlighting proteins should be very useful for researchers when designing knockout experiments because they might have off-target or side effects owing to moonlighting proteins with hidden phenotypic traits.

Globally, remote homology algorithm PSI-BLAST performs well, but in real situations it is difficult for the researcher to pick out the best hits from a large output. As described above, combining different bioinformatics algorithms for protein sequence analysis can help to reduce the targets and disclose putative moonlighting proteins. The best approach is to combine interactomics databases with PSI-BLAST outputs, although, at present, it has to be

Figure 2 | **Example of the PiSite result for fatty acid multifunctional protein (MFP)**

Here, a significant level of structural similarity to two already solved PDB structures representing the canonical and moonlighting function areas was found. This allows for structural superposition and it is very useful to map both functions. The key active residues involved in the canonical and moonlighting functions are depicted in the 'ball-and-stick' format and coloured yellow.



Protein name	Canonical Function	Multifunctional Function	Protein found by PISITE	Function
Fatty acid multifunctional protein(MFP)	Enoyl CoA hydratase	Enoyl CoA dehydrogenase	Enoyl CoA dehydrogenase (Q5KYB2)	Enoyl CoA dehydrogenase

RMSD = 3.23 Å
Fatty acid multifunctional protein (MFP)
(Moonlighting protein) (*Arabidopsis thaliana*)
- Enoyl CoA hydratase (EC= 4.2.1.17)
- Enoyl CoA dehydrogenase (EC=1.1.1.3)
Hydroxyacyl-CoA dehydrogenase (Human) (EC=1.1.1.3)
Main functional site for Hydroxyacyl-CoA dehydrogenase activity
RMSD = 4.21 Å
Fatty acid multifunctional protein (MFP)
(Moonlighting protein) (*Arabidopsis thaliana*)
- Enoyl CoA hydratase (EC= 4.2.1.17)
- Enoyl CoA dehydrogenase (EC=1.1.1.3)
Enoyl CoA hydratase (*Geobacillus kaustophillus*) (EC= 4.2.1.17)
Main functional site for Enoyl CoA hydratase activity

carried out by manual inspection. This combination leads to the prediction of ~50% of the moonlighting proteins, albeit from a database of proteins previously demonstrated to be multitasking. Two additional approaches described previously can help us to consider a protein hit that comes from the methods described as a true positive moonlighting protein. One of them is the alignment with known 3D structures, which in addition helps to map both functions, and amino acid mutation correlation which can suggest clues to the evolution of multifunctionality when compared with monofunctional examples in the family. The second one is the mutation correlation analysis applied to enolases. In this case, the structure of the enolase is shaped to complement plasminogen, which does not happen with other proteins of the pathogen, although probably the original enolase lacks some of the correct amino acids for allowing a sufficiently strong union. In this case, we can see that we need between five and ten concurrent mutations for the adaptation to occur, but this would eventually involve restructuring other regions of the protein not directly related to the newly acquired function, but maintaining the structure and folding. Enolase has a secondary function that appears quite often in different micro-organisms and this protein allows us to test whether the acquisition of a new multifunctionality is a frequently occurring phenomenon or not. If this only occurs very occasionally, the repetition of the same function will be linked to the similarity among the different proteins, indicating that this function has emerged from a common ancestor of the micro-organisms containing these enolases. In the opposite case, if none of the proteins share this extra function with any closely related sequence, this could mean that multifunctionality is a frequent event in evolution. The result of these analyses is more consistent with the second hypothesis. This is not a conclusive result, but an interesting clue in the sense that the current list of multifunctional proteins is only a minimal representation of what we expect.

At the present state of the art, the bioinformatics analysis is better for checking specific protein cases in which the researcher suspects the possibility of it being a moonlighting protein from experimental results or paradoxes. Protein function prediction is a daunting task, therefore it is even more daunting when the protein is multifunctional.

Funding

This research was supported by Ministerio de Ciencia e Innovación of Spain [grant numbers BFU2010-22209-C02-01 and BIO2013-48704-R], the Centre de Referència de R+D de Biotecnologia de la Generalitat de Catalunya, La Marató de TV3 [grant number 101930/31/32/33] and the Comisión Coordinadora del Interior de Uruguay.

References

1 Wool, I.G. (1996) Extraribosomal functions of ribosomal proteins. Trends Biochem. **21**, 164–165 CrossRef

2 Jeffery, C.J. (1999) Moonlighting proteins. Trends Biochem. Sci. **24**, 8–11 CrossRef PubMed

3 Jeffery, C.J. (2003) Moonlighting proteins: old proteins learning new tricks. Trends Genet. **19**, 415–417 CrossRef PubMed

4 Jeffery, C.J. (2004) Molecular mechanisms for multitasking: recent crystal structures of moonlighting proteins. Curr. Opin. Struct. Biol. **14**, 663–668 CrossRef PubMed

5 Jeffery, C.J. (2009) Moonlighting proteins: an update. Mol. Biosyst. **5**, 345–350 CrossRef PubMed

6 Piatigorsky, J. (2007) Gene Sharing and Evolution: the Diversity of Protein Function, Harvard University Press, Cambridge, MA, U.S.A.

7 Gancedo, C. and Flores, C.-L.M. (2008) Moonlighting proteins in yeast. Microbiol. Mol. Biol. Rev. **72**, 197–210 CrossRef PubMed

8 Nobeli, I., Favia, A.D. and Thornton, J.M. (2009) Protein promiscuity and its implications for biotechnology. Nat. Biotechnol. **27**, 157–167 CrossRef PubMed

9 Huberts, D. and van der Klie, I.J. (2010) Moonlighting proteins: an intriguing mode of multitasking. Biochim. Biophys. Acta **1803**, 520–525 CrossRef PubMed

10 Copley, S.D. (2012) Moonlighting is mainstream: paradigm adjustment required. BioEssays **34**, 578–588 CrossRef PubMed

11 Hernández, S., Ferragut, G., Amela, I., Cedano, J., Perez-Pons, J.A., Piñol, J., Mozo-Villarias, A.J., Cedano, J. and Querol, E. (2014) MultitaskProtDB: a database of multitasking proteins. Nucleic Acids Res. **42**, D517–D520 CrossRef PubMed

12 Gómez, A., Domedel, N., Cedano, J., Piñol, J. and Querol, E. (2003) Do current sequence analysis algorithms disclose multifunctional (moonlighting) proteins? Bioinformatics **19**, 895–896 CrossRef PubMed

13 Gómez, A., Hernández, S., Amela, I., Piñol, J., Cedano, J. and Querol, E. (2011) Do protein–protein interaction databases identify moonlighting proteins? Mol. Biosyst. **7**, 2379–2382 CrossRef

14 Becker, E., Robisson, B., Chapple, C.E., Guénoche, A. and Brun, C. (2012) Multifunctional proteins revealed by overlapping clustering in protein interaction network. Bioinformatics **28**, 84–90 CrossRef PubMed

15 Prieto, C. and de la Rivas, J. (2006) APID: Agile Protein Interaction DataAnalyzer. Nucleic Acids Res. **34**, W298–W302 CrossRef PubMed

16 Beissbarth, T. and Speed, T.P. (2004) GOstat: find statistically overrepresented Gene Ontologies within a group of genes. Bioinformatics **20**, 1464–1465 CrossRef PubMed

17 Altschul, S.F., Madden, T.L., Shaffer, 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 CrossRef PubMed

18 Gómez, A., Cedano, J., Hermoso, A., Piñol, J. and Querol, E. (2008) Prediction of protein function improving sequence remote alignment search by a fuzzy logic algorithm. Protein J. **27**, 130–139 CrossRef PubMed

19 Hunter, S., Jones, P., Mitchell, A., Apweiler, R., Attwood, T.K., Bateman, A., Bernard, T., Binns, D., Bork, P., Burge, S. et al. (2012) InterPro in 2011: new developments in the family and domain prediction database. Nucleic Acids Res. **40**, D306–D312 CrossRef PubMed

20 Henikoff, S., Henikoff, J.G. and Pietrokoski, S. (1999) Blocks+: a non-redundant database of protein alignment blocks derived from multiple compilations. Bioinformatics **15**, 471–479 CrossRef PubMed

21 Higurashi, M., Ishida, T. and Kinoshita, K. (2009) PiSite: a database of protein interaction sites using multiple binding states in the PDB. Nucleic Acids Res. **37**, D360–D364 CrossRef PubMed

22 Khan, I.K., Chitale, M., Rayon, C. and Kihara, D. (2012) Evaluation of function predictions by PFP, ESG, and PSI-BLAST for moonlighting proteins. BMC Proc. **6** (Suppl. 7), S5 CrossRef PubMed

23 Simonetti, F.L., Teppa, E., Chernomoretz, A., Nielsen, M. and Marino Buslje, C. (2013) MISTIC: mutual information server to infer coevolution. Nucleic Acids Res. **41**, W8–W14 CrossRef PubMed

24 Ehinger, S., Schubert, W.D., Bergmann, S., Hammerschmidt, S. and Heinz, D.W. (2004) Plasmin(ogen)-binding α-enolase from *Streptococcus pneumoniae*: crystal structure and evaluation of plasmin(ogen)-binding sites. J. Mol. Biol. **343**, 997–1005 CrossRef PubMed

Received 4 September 2014
doi:10.1042/BST20140241