Molecular BioSystems

RSC Publishing

PAPER

View Article Online
View Journal | View Issue

Cite this: *Mol. BioSyst.*, 2013, **9**, 70–76

Incorporating information on predicted solvent accessibility to the co-evolution-based study of protein interactions†

David Ochoa,[a] Ponciano García-Gutiérrez,‡[a] David Juan,[b] Alfonso Valencia[b] and Florencio Pazos*[a]

Received 9th August 2012,
Accepted 9th October 2012

DOI: 10.1039/c2mb25325a

www.rsc.org/molecularbiosystems

A widespread family of methods for studying and predicting protein interactions using sequence information is based on co-evolution, quantified as similarity of phylogenetic trees. Part of the co-evolution observed between interacting proteins could be due to co-adaptation caused by inter-protein contacts. In this case, the co-evolution is expected to be more evident when evaluated on the surface of the proteins or the internal layers close to it. In this work we study the effect of incorporating information on predicted solvent accessibility to three methods for predicting protein interactions based on similarity of phylogenetic trees. We evaluate the performance of these methods in predicting different types of protein associations when trees based on positions with different characteristics of predicted accessibility are used as input. We found that predicted accessibility improves the results of two recent versions of the *mirrortree* methodology in predicting direct binary physical interactions, while it neither improves these methods, nor the original *mirrortree* method, in predicting other types of interactions. That improvement comes at no cost in terms of applicability since accessibility can be predicted for any sequence. We also found that predictions of protein–protein interactions are improved when multiple sequence alignments with a richer representation of sequences (including paralogs) are incorporated in the accessibility prediction.

Introduction

Computational methods for predicting protein interactions and functional relationships complement experimental techniques in deciphering the networks of protein interactions underlying cellular processes. These techniques are not only faster and cheaper but, in certain situations and for certain types of interactions, their levels of accuracy/coverage are comparable to their experimental counterparts.[1] The tendency now is to combine both approaches in order to obtain reliable interactomes.[2,3]

These computational techniques are based on genomic and sequence features intuitively related to interaction (see ref. 4–7 for recent reviews). A widely used computational approach for detecting interacting proteins is based on similarity of phylogenetic trees (co-evolution). It was repeatedly observed that the phylogenetic trees of interacting proteins are more similar than those of non-interacting ones (see ref. 8, 9 and references therein).

This relationship between protein co-evolution (measured as similarity of trees) and interactions is being exploited in many different ways, ranging from the detailed study of particular interacting families which now can be performed with on-line interactive tools,[10] to the prediction of interactomes in a high-throughput way (*e.g.* ref. 11 and 12), to the prediction of the associations between the members of two protein families known to be related (*e.g.* a family of ligands and the corresponding receptors[13,14]).

The underlying cause for this observed relationship between protein co-evolution and interactions is still a matter of certain debate. The possible explanations range from specific co-adaptation between the interacting partners to general global similarities between their evolutionary rates.[8,15,16] The co-adaptative hypothesis proposes that a long process of specific co-adaptation at the residue

[a] *Computational Systems Biology Group, National Centre for Biotechnology (CNB-CSIC), C/Darwin, 3, Cantoblanco, 28049 Madrid, Spain. E-mail: pazos@cnb.csic.es; Fax: +34 91 5854506; Tel: +34 91 5854669*

[b] *Structural Bioinformatics Group, Spanish National Cancer Research Centre (CNIO), C/Melchor Fernández Almagro 3, 28029 Madrid, Spain*

† Electronic supplementary information (ESI) available. See DOI: 10.1039/c2mb25325a

‡ Present address: Chemistry Department, Universidad Autónoma Metropolitana-Iztapalapa, 09340 México D.F., Mexico.

level (in which interaction de-stabilizing changes in one protein are compensated by changes of similar magnitude in the other) would be the responsible for the observed similarity of evolutionary histories. In the other extreme, it is proposed that this observed similarity could be simply due to the similarity between the evolutionary rates of interacting and functionally related proteins. These two possible explanations for the observed relationship between co-evolution and interactions had been already proposed in the first works dealing with this subject.[17] While these two factors could be jointly contributing to the observed co-evolution, it is possibly the similarity of evolutionary rates that having a major effect, since compensatory changes would need to occur in large numbers in order to really affect the phylogenetic trees.[8]

A number of works have tried, more or less directly, to get some insight into the contribution of co-adaptation to the observed co-evolution.[15,18] The simplest way to approach this problem is to evaluate co-evolution using only the regions of the proteins amenable to co-adaptation (compensatory changes), that is, interaction surfaces (interfaces) or the whole surface, depending on the available information. If co-evolution is (mainly) due to the similarity in evolutionary rates, it would be "spread" through the whole sequence of the proteins, while if it were mainly due to compensatory changes it would be more evident in the surface/interface residues. However, not only surface residues can suffer inter-protein compensatory changes, but also those partially buried or even internal ones *via* indirect and allosteric effects. Moreover, the "intersection" between data on protein three-dimensional (3D) structures and interactions is not high, leading to small or eventually biased datasets to perform these studies. The scarcity in 3D data has another effect: if a methodology is eventually developed which combines co-evolution with structural information (solvent accessibility) for improving the accuracy in predicting interactions, its range of applicability would drop drastically compared to its counterparts which require only sequence information. Based on the above, it would be desirable to study the effect of incorporating predicted solvent accessibility information on co-evolution methods, instead of the "real" solvent accessibility extracted from experimental 3D structures. Predicted solvent accessibility can be obtained for any sequence, and with good levels of accuracy: above 75% for two-state predictions ("buried/exposed").[19,20]

In this work we assess for the first time the effect of including predicted solvent accessibility information on the results and range of applicability of three co-evolution based methods for predicting protein interactions. We used a number of datasets representing different types of interactions (physical, functional, ...) as gold standards in order to interpret the results in terms of the type of interaction of interest.

Methods

We aim to evaluate the effect of the incorporation of information on predicted solvent accessibility in the performance of three *mirrortree*-related methods in predicting interactions of different nature. This has been done by generating, for all proteins in the model organism *E. coli*, different sets of phylogenetic trees constructed using (i) the whole protein and (ii) only the protein residues above certain thresholds of predicted accessibility, and evaluating the performance of the methods based on these different trees. In order to evaluate the performance we used different datasets of protein interactions representing interactions of different nature (*e.g.* physical and functional). The process is illustrated in Fig. 1, and details are given below.

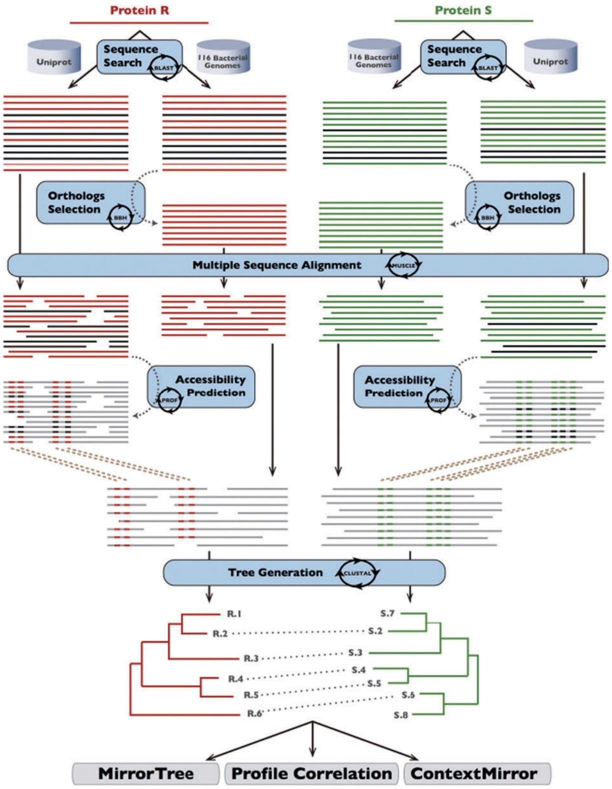


Fig. 1 Scheme of the methodology. In order to evaluate the co-evolution between proteins R and S based on their residues fulfilling a given predicted accessibility criterion, the first step is to look for their orthologs in a set of 116 fully sequenced genomes. For each protein, a multiple sequence alignment is generated with these orthologs, which will serve as a basis for the generation of the trees. In parallel, another multiple alignment is generated for the same protein based on the homologs found in the whole Uniprot database (hence including orthologs and paralogs). This second alignment will be used for the prediction of solvent accessibility. A tree is generated based on the first alignment but using only the positions with a given predicted accessibility criterion. The trees generated in this way are the input for the three methods for evaluating co-evolution.

Solvent accessibility prediction

First, for each *E. coli* protein, a list of candidate homolog protein sequences was retrieved searching with BLAST[21] in the non-redundant Uniprot database.[22] Sequences with an *E*-value greater than 1×10^{-4} or an identity (based on the BLAST alignment) less than 20% were excluded. Alignment coverages lower than 60% (either respect to the hit or the query protein) were also excluded.

A multiple sequence alignment (MSA) for the remaining sequences was generated using MUSCLE.[23] The identity of each aligned sequence with the *E. coli* reference sequence was then calculated using only positions with less than 90% gaps. If this identity was less than 20%, the sequence was discarded. Additionally, sequence redundancy was removed at 95% to avoid the overrepresentation of some sequences, which could influence the accessibility predictions.

Finally, this multiple sequence alignment was used as input for the PROF program for predicting solvent accessibility,[24,25] and the predictions for the columns of the MSA were mapped to the positions of the original *E. coli* protein. For comparative purposes equivalent accessibility predictions were also generated based on the MSAs of orthologs used for constructing the phylogenetic trees (described in the next section).

Generation of phylogenetic trees

We used a set of 116 fully sequenced organisms previously used in other works[11,26] to look for orthologs of *E. coli* proteins and construct the trees based on them. This set does not contain very similar organisms, thus avoiding phylogenetic redundancy.

We used the ''BLAST best bi-directional hit'' criterion for detecting the ortholog of a given *E. coli* protein in each genome, with an *E*-value cut-off of 1×10^{-5}, and requiring an alignment coverage of 70%. All orthologs found for this *E. coli* protein were aligned with MUSCLE[23] using the default parameters of this program. Then, a phylogenetic tree was generated from this alignment using the neighbor-joining algorithm implemented in ClustalW,[27] excluding the gaps for the distance calculation.

Equivalent trees were generated but using only the positions of the alignment fulfilling the following criteria of predicted accessibility:

- eRIA0: positions predicted as accessible by PROF with any value of ''reliability''.
- eRIA3: positions predicted as accessible with reliability ≥ 3 (PROF reliability values range from 0 to 9).
- pACC2, pACC12 and pACC50: positions with a predicted solvent accessible surface ≥ 2, 12 and 50 Å^2, respectively.

Finally, distance matrices containing the pair-wise distances between all orthologs were generated for the original tree (based on the whole length of the protein) as well as for these trees based on (predicted) accessible positions. These distances are calculated by summing the lengths of the branches separating the corresponding leaves. These distance matrices are the input for the *mirrortree*-based methods described in the next point.

Prediction of protein interactions based on phylogenetic trees

The original *mirrortree* (MT) approach[17] evaluates the co-evolution between two protein families by calculating the linear correlation coefficient between the values of their corresponding distance matrices. A minimum of 15 species in common is required in order to evaluate a given pair of proteins. Moreover, only correlation values supported by a tabulated *P*-value of 1×10^{-5} or better are used.

The *profile-correlation* (PC) method[11] takes as input the *mirrortree* raw scores for all pairs of proteins in a given organism. Hence, in this case the input is a squared matrix the size of the *E. coli* proteome with the correlation values for all pairs of proteins (actually, those with 15 or more organisms in common and supported by a *P*-value $\leq 1 \times 10^{-5}$). A row in this matrix, known as ''co-evolutionary profile'', represents the co-evolutionary behaviour of a protein respect to the rest of the proteome. Within the context of the PC method, the co-evolution between two proteins is re-evaluated as the correlation between their corresponding co-evolutionary profiles, with the same significance thresholds used for the original *mirrortree*. The idea is that two proteins whose trees are similar and, additionally, that tend to be similar to the same set of proteins (and dissimilar to the complementary set) are more likely to represent a case of true co-evolution.

The *context-mirror* (CM) method[11] takes into account the influence of ''third proteins'' in a given co-evolutionary signal observed for a given pair of proteins using a partial correlation criterion. In this way it is possible to separate specific co-evolution (particular to a given pair of proteins) from general co-evolutionary trends involving many proteins. For a given pair of proteins, this method produces results at different ''levels'' of specificity, being ''level 1'' the one representing the most specific co-evolution.

Datasets of protein interaction and functional relationship

The performance of the three methods when fed with phylogenetic trees generated with residues of different predicted accessibility was evaluated using three datasets representing protein interactions of different nature in *E. coli* as gold standard.

- Binary physical direct interactions obtained from MPIDB.[28] This database contains binary interactions manually curated from the literature or imported from other databases. We retrieved the 2103 binary interactions between 1538 different *E. coli* proteins stored on it.
- Physical (sometimes indirect) interactions inferred as co-presence in experimentally determined macromolecular complexes obtained from EcoCyc.[29] This dataset contains 1354 experimentally determined interactions between 591 proteins.
- Functional interactions inferred as membership in the same metabolic pathways, also taken from the EcoCyc. This dataset contains 4419 relations between 719 proteins.

In the three cases, the sets of negatives (pairs of proteins regarded as non-interacting) were constructed by generating all possible pairs between the proteins in the corresponding positive (interacting) sets, excluding those pairs already annotated as interacting.

Performance evaluation

For each combination of a method, an input set of trees (generated from residues of different predicted accessibility) and an interaction dataset we obtain a list of protein pairs, sorted by the score of the corresponding method. Each pair can be labelled as positive or negative depending on whether it is a reported interaction in that particular dataset or not. A combination method-set of trees will be better for predicting interactions (for that particular set of interaction evidences) as the positives tend to cluster at the top of these sorted lists (associated to high scores) and the other way around for the negatives.

The Area Under the ROC Curve (AUC) was calculated for these lists, as a global estimator of the accuracy and coverage of the corresponding predictions. The ROC ("receiver operating characteristic") analysis[30] generates a plot of "true positives rate" (TPR) against "false positives rate" (FPR) when varying the classification threshold (score of the method). Curves above the diagonal in this plot represent methods with some discriminative power, being this discriminative capacity better as the curve gets closer to the top-left corner of the plot. Consequently, areas under these curves range from 0.5 (random classifier, diagonal in the plot, positives and negatives uniformly distributed through the list) to 1.0 (perfect classifier, all positives at the top of the list). ROC analysis was performed with the ROCR library of the R statistical package (http://www.r-project.org).

Results and discussion

Fig. 2 and Fig. S1 (ESI†) show the performance (AUC value) of the three *mirrortree*-based methods, when using the phylogenetic trees constructed from residues of different predicted accessibility, and evaluated based on the three different datasets of protein interactions.

As previously seen,[11,26] *mirrotree*-based methods predict better physical interactions (binary and complexes) than functional associations (*e.g.* pathways). Within physical interactions, those representing co-membership to macromolecular complexes are better detected than those representing binary (eventually transient) interactions. About the methods, the PC and CM methods work

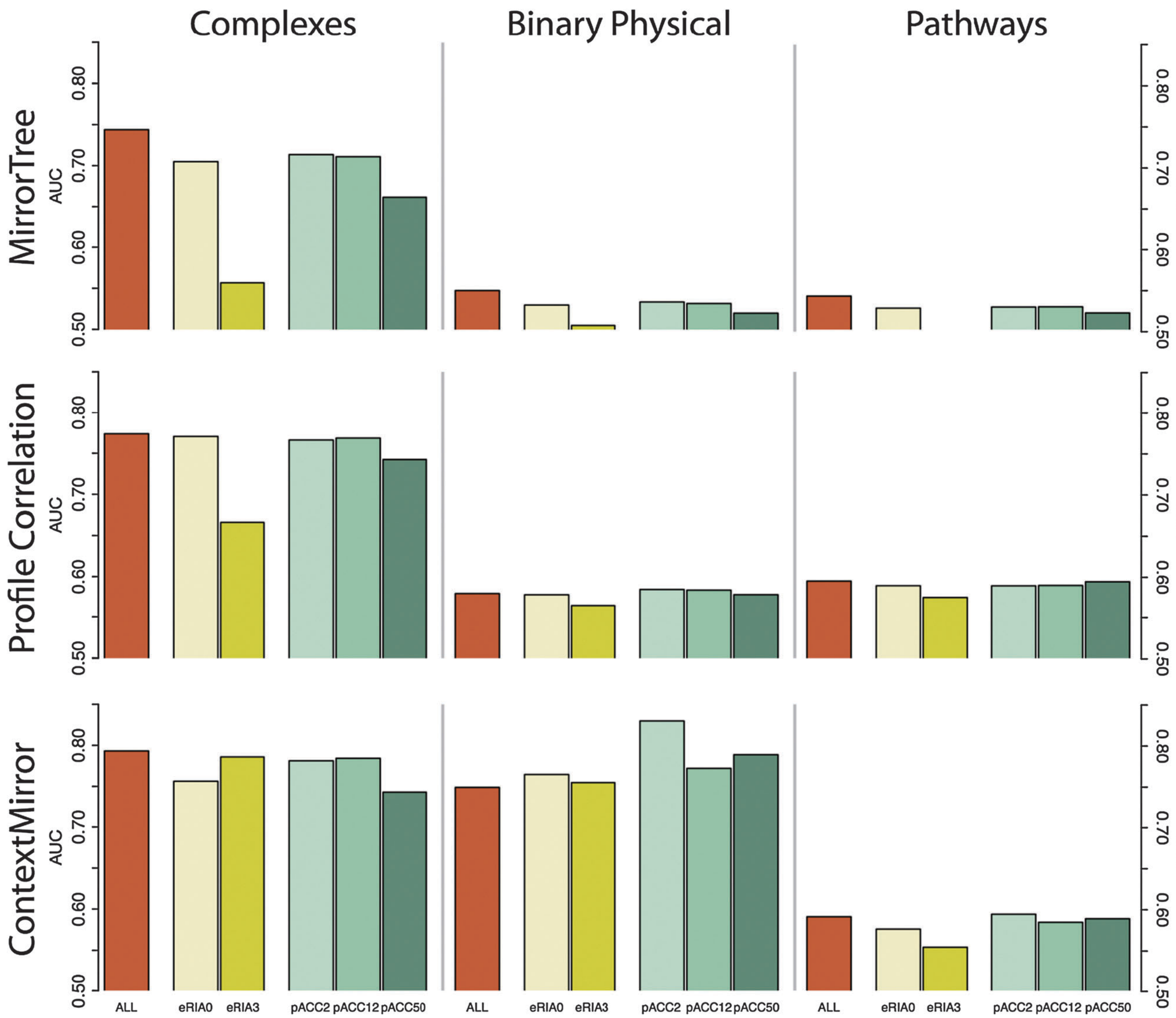


Fig. 2 Performances for different combinations of: phylogenetic tree comparative methods, interaction evidence and predicted accessibility filter. Performance is evaluated as the "Area Under the [ROC] Curve" (AUC). The same figure with different scales for each plot is available as Fig. S1 (ESI†). Equivalent figures with the results obtained using predicted accessibility derived from MSAs of orthologs are available as Fig. S2 and S3 (ESI†).

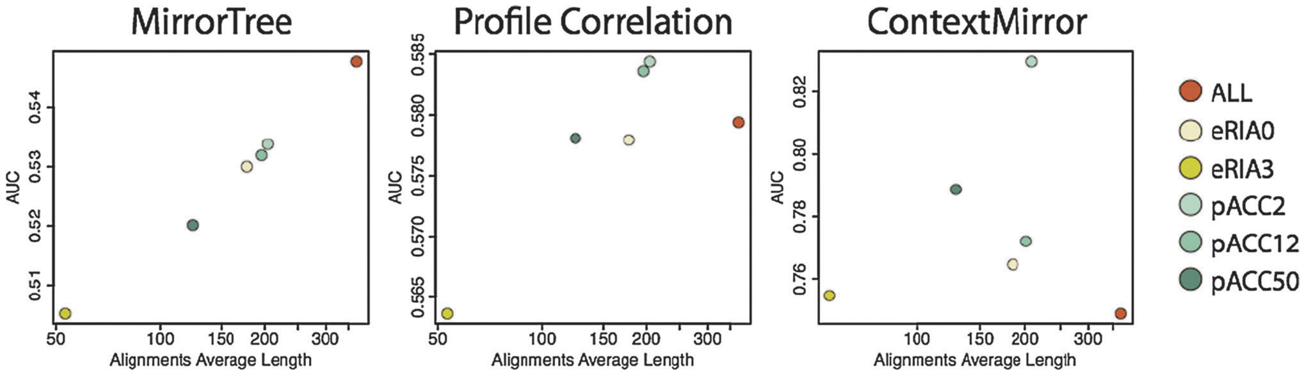


Fig. 3 Relationship between the performances and the lengths of the virtual alignments. The length of the virtual alignment is the number of positions (fulfilling a given predicted accessibility criteria – colors) used for deriving the trees. The data shown here are for binary physical interactions. The corresponding plots for the other interaction datasets are available as Fig. S4 (ESI†).

better than the original MT approach, which presents ''usable'' levels of performance only for detecting interactions of macromolecular complexes.

Predicted accessibility only helps the PC and CM methods in detecting binary physical interactions

For most cases, the use of predicted solvent accessibility within *mirrortree*-based methodologies worsens the results (Fig. 2 and Fig. S1, ESI†). The AUC values for these methods working with trees derived from different sets of (predicted) accessible residues are worse than those based on full sequences.

Interestingly, for the case of binary physical interactions, the results of the PC and CM methods are improved when using predicted solvent accessibility. The best results are obtained when using all residues with a minimum of solvent accessible area (''pACC2'', area $\geq$ 2 Å^{2}). Restricting to residues predicted to be highly accessible ($\geq$ 12 and $\geq$ 50 Å^{2}), or those predicted as ''accessible'' by PROF's two-state predictor (eRIA0 and eRIA9) works worse than with $\geq$ 2 Å^{2}.

For most cases, there is a correlation between the performances obtained with the trees based on different predicted accessibilities and the average lengths of the virtual alignments used for deriving them (number of positions fulfilling that particular accessibility cut-off) (Fig. 3 and Fig. S4, ESI†). This trend is broken for the results of the PC and CM methods predicting direct physical interactions: in these two cases pACC2 renders the better results in spite of not having the largest virtual alignments (Fig. 3). This general decrease in performance when incorporating predicted accessibility could be partly due to the intrinsic errors associated with the prediction. Nevertheless, it is probably more related to the largest contribution of the similarity of evolutionary rates to the observed co-evolution (see above): the co-evolutionary signal would be spread through the whole sequence and not restricted to certain parts (surfaces, *etc.*) This is reinforced by the observation that, in general, performances correlate positively with the number of positions used for building the trees.

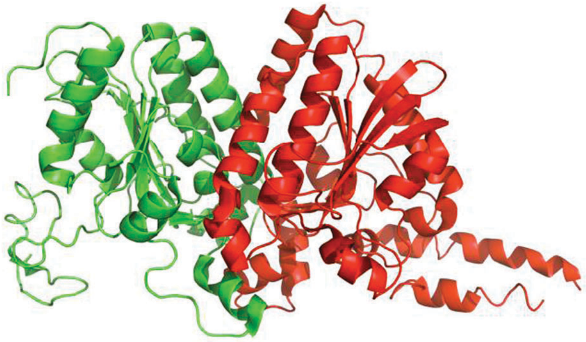
Our interpretation for the fact that accessibility predictions do not help the original MT (but the other way around) is that this methodology is mainly detecting non-specific co-evolution associated with global similarities in evolutionary rates (reflected in the whole sequence, as commented above). The more recent PC and CM methods benefit by the use of predicted accessibility when applied for the prediction of binary direct physical interactions. These two methods have been previously associated with the detection of more specific co-evolution,[11] cases where the co-adaptive part of the co-evolutionary signal is probably higher. In the same line, these specific co-evolutions (with larger proportions of co-adaptations) are intuitively more related to direct physical interactions than, for example, to those relating the members of macromolecular complexes.

It is also interesting that the set of predicted accessible residues which renders the best results include residues with a minimal predicted solvent accessible area (> = 2 Å^{2}), which are better than those with higher levels of predicted accessibility (> = 12 Å^{2}, > = 50 Å^{2}). That could be explained by the fact that co-adaptation is not necessarily restricted to totally exposed residues but can also happen between their neighbours or even buried residues (through allosteric effects).

It is better to use accessibility predicted from MSAs constructed for this purpose, than that based on MSAs with orthologs only

Fig. S2 and S3 (ESI†) show the same AUC results as Fig. 2 and Fig. S1 (ESI†) but using accessibility predicted from the same alignments used for constructing the phylogenetic trees input of the *mirrortree*-based methods (composed by orthologs only). The general drop in performance detected when incorporating accessibility information can be observed to be sharper here. Moreover, for those cases in which predicted accessibility improved the results (PC and CM predicting physical interactions, previous point) the improvement obtained with these alignments of orthologs is still present but smaller. Therefore, accessibility predicted from the same alignments used for constructing the phylogenetic trees renders worse results than that predicted from MSAs constructed *ad hoc* for this purpose.

The fact that accessibility predicted from ''richer'' alignments (including eukaryotic sequences and eventually paralogs) is



	ALL	eRIA0	eRIA3	pACC2	pACC12	pACC50
MirrorTree	0.9083	0.8982	0.8331	0.8998	0.9069	0.8924
Profile Correlation	0.9516	0.9531	0.9435	0.9605	0.9592	0.9328
ContextMirror	0.6068	0.6650	0.5410	0.6818	0.6828	0.5407

Fig. 4 Example illustrating the effect of incorporating predicted solvent accessibility on the evaluation of tree similarity. The structure of the complex between the α and β chains of *E. coli* acetyl-CoA carboxylase carboxyl transferase is shown in ribbon representation. The table contains the scores of the three methods for this interacting pair of proteins based on the trees derived with the six different criteria of predicted accessibility.

better in helping these co-evolution based methods than that based on alignments containing only bacterial orthologs was expected. It was previously shown that the quality of the MSA is critical for obtaining good sequence-based predictions of protein features such as accessibility or secondary structure.[19] Nevertheless, we wanted to make a test with MSAs of orthologs due to a methodological reason: these MSAs have to be generated in order to apply *mirrortree* and related methods. Consequently, if the accessibility predicted from them turned out to perform similarly to that predicted from richer alignments, it would be trivial to add this accessibility prediction step to current *mirrortree* workflows. Unfortunately, although some improvement is obtained with that accessibility, the best results are obtained when using that predicted from richer alignments. Consequently, in order to obtain these optimal results the workflow has to be "bifurcated", generating one alignment for tree construction and another one for accessibility prediction, as shown in Fig. 1.

Example

For illustrative purposes only, we include an example of an interacting pair of proteins for whose co-evolution is more evident when evaluated using solvent accessible predicted residues. Fig. 4 shows the results of *mirrotree* and related methods evaluating the co-evolution between the α and β subunits of the *E. coli* acetyl-CoA carboxylase carboxyl transferase. It can be seen that the similarity between the evolutionary histories of these two interacting proteins is more evident when evaluated from trees constructed using the residues predicted as accessible, except for the original MT method. For example, the score of the ContextMirror method increases from 0.60, when it is based on the trees derived from the whole sequence of these proteins, to 0.68 (trees based on predicted solvent accessible residues).

Conclusions

The underlying cause for the observed relationship between protein co-evolution and interactions is still not totally clear. The possible explanations range from unspecific co-evolution due to the similarity of evolutionary rates of interacting proteins, to specific co-adaptation involving inter-protein compensatory changes.[8,16] It is possibly the first factor the one playing a major role since evolutionary rate and interactions have been previously related through a number of direct and indirect paths.[15,31] The co-evolution observed in pairs of functionally related proteins which do not necessarily interact physically (*e.g.* ref. 32 and 33) is also easier to understand under this hypothesis. Nevertheless, compensatory changes have been repeatedly observed in protein interfaces (*e.g.* see ref. 8) and are surely playing a role in the co-evolution of interacting proteins at a local level. However, it is difficult to conceive these changes as mostly responsible for the observed tree similarity, since a very large number of such compensatory changes would be necessary to have an effect on the shapes of the trees. Previous studies tried to disentangle these two factors by comparing the co-evolution of protein regions amenable to compensatory changes (surfaces and interfaces) to that of the whole protein length.[15,18] In this work we tackle this problem but using predicted solvent accessibility, instead of real surfaces.

We have demonstrated that using predicted solvent accessibility helps in the co-evolution based prediction of protein interaction under some circumstances. Besides the implications of these results for the debate on the contribution of co-adaptation to the observed relationship between tree similarity and interactions, this work has also practical implications for the application of these methodologies, and these are not only related to the improvement in the prediction of protein interaction. Since this method goes on a step further in the detection of the protein regions actually co-evolving, it opens interesting possibilities for studying how the residues at the interfaces change and co-adapt during evolution. This could give some insight into the physico-chemical basis of protein interactions since the coordinated changes at the interfaces would provide a picture of possible interactions modes for a particular protein family. Moreover co-evolution has been proposed as a mechanism for maintaining interactions between proteins while allowing them to change at the same time. In many interacting protein families co-evolution is reflected in a set of specific surface residues which concomitantly change in both interacting partners. These residues are good candidates for mutagenesis experiments aimed at switching the interaction specificity of the proteins and/or adapting them to new interaction partners.

It is also important to highlight that the improvement obtained when incorporating predicted solvent accessibility does not have any associated cost in terms of coverage/applicability, since accessibility predictions can be generated for any sequence.

Acknowledgements

We want to thank Prof. Burkhard Rost and his group (currently at Technische Universität München) for help with local installations of PROF. We also want to thank the members of the

Computational Systems Biology Group (CNB-CSIC) for insightful discussions. This work was funded by a project from the Spanish Ministry for Economy and Competitiveness (BIO2010-22109). DO is recipient a fellowship from the Basque Country. PGG also thanks Mexican CONACyT for a graduate studies grant.

References

1 C. von Mering, R. Krause, B. Snel, M. Cornell, S. G. Oliver, S. Fields and P. Bork, *Nature*, 2002, **417**, 399–403.
2 I. Lee, S. V. Date, A. T. Adai and E. M. Marcotte, *Science*, 2004, **306**, 1555–1558.
3 C. von Mering, M. Huynen, D. Jaeggi, S. Schmidt, P. Bork and B. Snel, *Nucleic Acids Res.*, 2003, **31**, 258–261.
4 B. A. Shoemaker and A. R. Panchenko, *PLoS Comput. Biol.*, 2007, **3**, e43.
5 A. Valencia and F. Pazos, in *Protein–protein interactions and networks*, ed. A. R. Panchenko and T. M. Przytycka, Springer-Verlag, London, 2008, pp. 67–81.
6 E. D. Harrington, L. J. Jensen and P. Bork, *FEBS Lett.*, 2008, **582**, 1251–1258.
7 M. N. Wass, A. David and M. J. Sternberg, *Curr. Opin. Struct. Biol.*, 2011, **21**, 382–390.
8 F. Pazos and A. Valencia, *EMBO J.*, 2008, **27**, 2648–2655.
9 D. Juan, F. Pazos and A. Valencia, *FEBS Lett.*, 2008, **582**, 1225–1230.
10 D. Ochoa and F. Pazos, *Bioinformatics*, 2010, **26**, 1370–1371.
11 D. Juan, F. Pazos and A. Valencia, *Proc. Natl. Acad. Sci. U. S. A.*, 2008, **105**, 934–939.
12 E. R. Tillier and R. L. Charlebois, *Genome Res.*, 2009, **19**, 1861–1871.
13 A. K. Ramani and E. M. Marcotte, *J. Mol. Biol.*, 2003, **327**, 273–284.
14 J. M. Izarzugaza, D. Juan, C. Pons, J. A. Ranea, A. Valencia and F. Pazos, *Nucleic Acids Res.*, 2006, **34**, W315–W319.
15 L. Hakes, S. Lovell, S. G. Oliver and D. L. Robertson, *Proc. Natl. Acad. Sci. U. S. A.*, 2007, **104**, 7999–8004.
16 S. C. Lovell and D. L. Robertson, *Mol. Biol. Evol.*, 2010, **27**, 2567–2575.
17 F. Pazos and A. Valencia, *Protein Eng.*, 2001, **14**, 609–614.
18 M. G. Kann, B. A. Shoemaker, A. R. Panchenko and T. M. Przytycka, *J. Mol. Biol.*, 2009, **385**, 91–98.
19 B. Rost, in *Structural Bioinformatics*, ed. P. E. Bourne and J. Gu, Wiley-Blackwell, 2nd edn, 2009, pp. 679–714.
20 I. Y. Y. Koh, V. A. Eyrich, M. A. Marti-Renom, D. Przybylski, M. S. Madhusudhan, N. Eswar, O. Grana, F. Pazos, A. Valencia, A. Sali and B. Rost, *Nucleic Acids Res.*, 2003, **31**, 3311–3315.
21 S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller and D. J. Lipman, *Nucleic Acids Res.*, 1997, **25**, 3389–3402.
22 U. Consortium, *Nucleic Acids Res.*, 2009, **37**, D169–D174.
23 R. C. Edgar, *Nucleic Acids Res.*, 2004, **32**, 1792–1797.
24 B. Rost, *Methods Enzymol.*, 1996, **266**, 525–539.
25 B. Rost, in *The Proteomics Protocols Handbook*, ed. J. E. Walker, Humana, Totowa, NJ, 2005, pp. 875–901.
26 D. Herman, D. Ochoa, D. Juan, D. Lopez, A. Valencia and F. Pazos, *BMC Bioinf.*, 2011, **12**, 363.
27 R. Chenna, H. Sugawara, T. Koike, R. Lopez, T. J. Gibson, D. G. Higgins and J. D. Thompson, *Nucleic Acids Res.*, 2003, **31**, 3497–3500.
28 J. Goll, S. V. Rajagopala, S. C. Shiau, H. Wu, B. T. Lamb and P. Uetz, *Bioinformatics*, 2008, **24**, 1743–1744.
29 I. M. Keseler, J. Collado-Vides, S. Gama-Castro, J. Ingraham, S. Paley, I. T. Paulsen, M. Peralta-Gil and P. D. Karp, *Nucleic Acids Res.*, 2005, **33**, D334–D337.
30 T. Fawcett, *Pattern Recogn. Lett.*, 2006, **27**, 861–874.
31 H. B. Fraser, A. E. Hirsh, L. M. Steinmetz, C. Scharfe and M. W. Feldman, *Science*, 2002, **296**, 750–752.
32 Z. Liang, M. Xu, M. Teng, L. Niu and J. Wu, *FEBS Lett.*, 2010, **584**, 4237–4240.
33 N. L. Clark, E. Alani and C. F. Aquadro, *Genome Res.*, 2012, **22**, 714–720.