In Silico Two-Hybrid System for the Selection of Physically Interacting Protein Pairs

Florencio Pazos and Alfonso Valencia*
Protein Design Group, CNB-CSIC, Madrid, Spain

ABSTRACT **Deciphering the interaction links between proteins has become one of the main tasks of experimental and bioinformatic methodologies. Reconstruction of complex networks of interactions in simple cellular systems by integrating predicted interaction networks with available experimental data is becoming one of the most demanding needs in the postgenomic era. On the basis of the study of correlated mutations in multiple sequence alignments, we propose a new method (in silico two-hybrid, i2h) that directly addresses the detection of physically interacting protein pairs and identifies the most likely sequence regions involved in the interactions. We have applied the system to several test sets, showing that it can discriminate between true and false interactions in a significant number of cases. We have also analyzed a large collection of *E. coli* protein pairs as a first step toward the virtual reconstruction of its complete interaction network. Proteins 2002;47:219–227. © 2002 Wiley-Liss, Inc.**

Key words: protein–protein interactions; correlated mutations; co-evolution; interaction networks

INTRODUCTION

Recent advances in molecular biology have provided a vast amount of genetic information for many different organisms. A great challenge is to identify the possible interactions between different protein components in what has been called neighborhood relations.[1] Accumulated experimental evidence on metabolic and signaling pathways, supplemented by emerging techniques such as expression arrays, mass spectrometry applied to two-dimensional gels, and automated yeast two-hybrid systems,[2] is leading to the experimental identification of large numbers of relationships between sets of proteins.

A number of computational techniques for physical docking have been developed to tackle the problem of predicting protein complexes and protein interactions from knowledge of the corresponding structures.[3] Unfortunately, the number of cases for which the three-dimensional structures of both proteins are known is still very small, and even in these cases the predictions are of limited success.[4]

Another approach focuses on the prediction of functional relationships irrespective of physical interaction. Dandekar et al.[5] identified a relationship between genes that are contiguous in several bacterial chromosomes and proteins that form complexes. This observation was unfortunately limited to the small set of genes whose proximity in the genome is conserved in several bacterial species. Gaasterland and Ragan[6] and Pellegrini et al.[7] predicted functional interactions based on comparisons of the species distributions of gene pairs ("phylogenetic profiles"). These methods assumed that genomes coding for one member of an interaction pair would necessarily code for its interacting partner. Among other technical limitations, this approach would not be applicable to the many essential proteins that are widely distributed among different genomes.[8] Marcotte et al.[9] and Enright et al.[10] predicted protein interactions for those multidomain proteins that show a variety of domain arrangements in different genomes (e.g., the interaction between proteins A and B would be predicted if it is possible to find a single protein composed of two domains similar to A and B, respectively). Nonetheless, this method can be applied only to the few cases in which these types of molecular arrangement are found.

The in silico two-hybrid (i2h) approach proposed here is based on previous studies of sequence correlation between distant positions of multiple sequence alignments. We have shown that the study of such correlations can be used as a weak but significant predictor of interresidue contact.[11–13] Indeed, we showed that correlation information can be used systematically to improve protein structure prediction methods based on threading alignments.[14]

We also showed that correlation information is sufficient for selecting the correct structural arrangement of heterodimers and protein domains in a representative number of cases, because the correlated pairs between the monomers tend to accumulate at the contact interface.[15] This early approach addressed the detection of the structural interaction region between proteins known to interact. The conceptual advance proposed here is the direct search for interacting protein pairs. In practice, we search for pairs of multiple sequence alignments with a distinctive co-variation signal, based on the hypothesis that co-adaptation of interacting proteins can be detected by the presence of a distinctive number of compensatory

Grant sponsor: CICYT (Spain).

*Correspondence to: Alfonso Valencia, Protein Design Group, National Centre for Biotechnology, CNB-CSIC, Cantoblanco, Madrid E-28049, Spain. E-mail: valencia@cnb.uam.es

Received 2 February 2001; Accepted 16 November 2001

mutations in the corresponding proteins of different species.

MATERIALS AND METHODS

Calculation of Correlation Values

Correlated mutations evaluate the similarity in variation patterns between positions in a multiple sequence alignment. The similarity of those variation patterns is thought to be related to compensatory mutations. The basic method for calculating correlated mutations was originally proposed by Göbel et al.[11] A position-specific matrix is calculated for each position in the sequence; this matrix contains the distances, defined as in McLachlan,[16] between the residues corresponding to all sequence combinations at that position. The position-specific matrices are compared with a correlation formula. In the i2h system, we obtained the best results with a variation of the method in which the correlation formula was replaced by a rank correlation calculation,[17] in which the numerical values are replaced by their ordinal ranking number. The rank correlation coefficient (r_{ij}) between positions i and j is given by

$$r_{ij} = \frac{\sum_{k,j} (S_{ikl} - \bar{S}_i)(S_{jkl} - \bar{S}_j)}{\sqrt{\sum_{k,l} (S_{ikl} - \bar{S}_i)^2}\ \sqrt{\sum_{k,l} (S_{jkl} - \bar{S}_j)^2}}$$

where the summations run for every possible pair of proteins k and l in the multiple sequence alignment. S_{ikl} is the ranked similarity between residue i in protein k and residue i in protein l. S_{jkl} is the same for residue j. $\bar{S}_i$ and $\bar{S}_j$ are the means of S_{ikl} and S_{jkl}, respectively.

Protein Alignment and Distribution of Correlation Values

Original multiple sequence alignments were obtained by searching for homologous proteins with BLAST[18] and aligning them with Clustalw[19] in the *E. coli* test sets, or taken from the HSSP database[20] in the cases of the structural domains and the interacting proteins of known structure (see "test sets" below).

Starting with the multiple sequence alignments of the two proteins, we reduce them, leaving only sequences of coincident species. The entries in the position-specific matrices of the positions of the two proteins are thus comparable, and calculation of interprotein correlated mutations can be performed by using compatible protein sequences. This step is easy to visualize by imagining that the corresponding sequences of the same species in each one of the two proteins are linked in a "virtual concatenated alignment" (Fig. 1). If we have more than one homologous protein in the same species (paralogs), we choose that closest to the *E. coli* sequence in the *E. coli* test sets, and that closest to the HSSP master in the structural domains and the interacting proteins of known structure test sets.

The interaction index is obtained by comparing the distribution of correlation values that correspond to pairs of positions, one from each protein in the concatenated alignment, with the distributions corresponding to pairs of positions in the individual proteins. This normalization step is introduced to decrease the potential noise produced by the presence of divergent sequences that could introduce atypical high correlation values in both concatenated and individual alignments (Fig. 1). Thus, we obtain the interaction index score for proteins 1 and 2 with the following formula.

$$C_{12} = \sum_{i=\mathrm{incorr}}^{1.0} \frac{P_{12i}}{P_{11i} + P_{22i}} \cdot i$$

where the summation runs for all correlation bins from an initial value (*incorr*; 0.4 in this study) to 1.0. P_{12i} is the percentage of interprotein pairs with correlation value in bin i. P_{11i} and P_{22i} are the same for intraprotein pairs (Fig. 1).

We also compare the interaction indices of one protein with all its possible partners to evaluate the significance of each of the interactions. The comparison is presented in a Z-score.

Sets of Protein Families for Application of the i2h System

The i2h system was applied to various test sets. In each of them, all possible protein pairs were explored (for N proteins, a total of Nx(N-1)/2 possible pairs) by building the corresponding concatenated alignments. The main limitation was that it was not possible to find in all cases enough sequences after the alignment reduction step (see "obtaining the alignments"). In the current application, we fixed a minimal threshold of 11 sequences for each of the two proteins from the same species. Under these conditions, the number of protein pairs for which it was possible to perform the i2h calculation was considerably lower than the total number of possible pairs.

The first set was composed of structural domains rather than of proteins. We took 14 two-domain proteins with a tight intradomain interaction from Pazos et al.[15] The PDB codes of the proteins are 4mt2, 3dfr, 4tnc, 1rnd, 4mts, 3pgk, 1alc, 3blm, 2pf2, 3adk, 9pap, 2c2c, 3trx, and 1sgt. Calculations were made for 133 domain pairs.

The second set was started with 53 proteins analyzed by Dandekar et al.,[5] which form 31 known interactions. The total number of possible pairs was 1378, of which we could explore a reasonable fraction of 244 protein pairs by concatenating sequences extracted from 14 completely sequenced genomes, namely, *M. tuberculosis, N. Gonorrhoeae, E. coli, H. pylori, Synechocystis sp., M. thermoautotrophicum, A. aeolicus, B. burgdorferi, P. horikoshii, T. pallidum, B. subtilis, M. jannaschii, H. influenzae,* and *A. fulgidus.* Among those 244 pairs, 8 pairs corresponded to well-documented interactions, and some other pairs could be considered candidates for possible interactions, including different ribosomal proteins.

A third set contains 195 pairs with 15 possible interactions, derived from the 749 predicted interactions reported by Marcotte et al.[9] after selecting sequences from corresponding species as described above. The Marcotte set was derived by selecting pairs of *E. coli* proteins that are fused into a single protein in different organisms, which in many

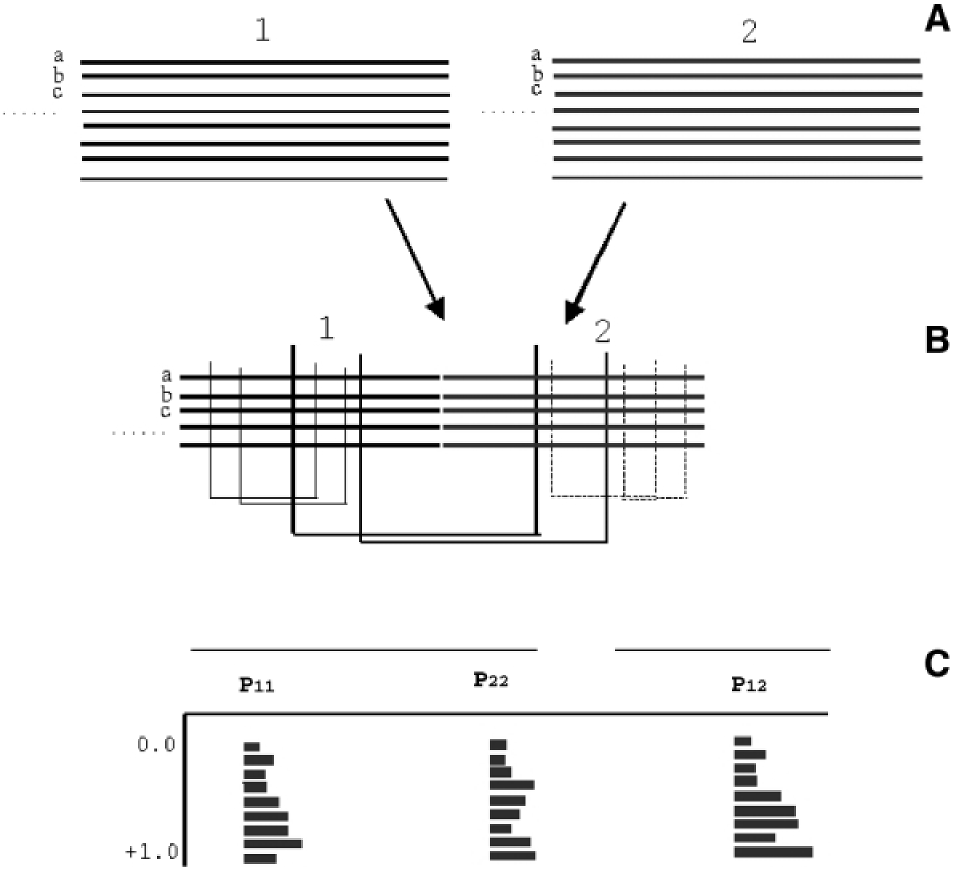

Fig. 1. Schematic representation of the i2h method. **A:** Family alignments are collected for two different proteins, 1 and 2, including corresponding sequences from different species (a, b, c, ...). **B:** A virtual alignment is constructed, concatenating the sequences of the probable orthologous sequences of the two proteins. Correlated mutations are calculated as described in Materials and Methods. **C:** The distributions of the correlation values are recorded. We used 10 correlation levels. The corresponding distributions are represented for the pairs of residues internal to the two proteins (P11 and P22) and for the pairs composed of one residue from each of the two proteins (P12). The interaction index is calculated from these distributions (see Materials and Methods).

cases implies a functional relationship between the proteins, but not necessarily a direct physical interaction.[9] In most cases, neither the functional nor the physical interactions have been verified experimentally.

The fourth test set was composed of interacting proteins of known structure. The interacting chains were extracted from the SPIN database (http://trantor.bioc.columbia.edu/cgi-bin/SPIN/), which contains all the protein complexes contained in the PDB Protein Data Bank.[21] By using the SPIN search engine, it is possible to search the set of protein complexes for specific characteristics. We searched in its sequence-unique set, excluding homodimers and complexes involving proteases. This was done to eliminate strong peculiar signals. We also excluded chains involved in more than one interaction in an attempt to limit the set to heterodimers. The set was then filtered by using the SCOP[22] structural classification to eliminate the chains labeled as "membrane peptides," "small proteins," and "coiled coils." This filtering resulted in 226 interacting protein chains (113 interactions). In this case, we need 12 sequences from common species for a pair to be evaluated. From the final list of 25,425 possible pairs, we excluded those with sequence identity >40%. With those restrictions we could make calculations for 321 pairs, 17 of which are known to interact.

The fifth protein set was formed of a collection of 67,238 pairs corresponding to *E. coli* proteins that fulfill the requirements for number of sequences in the concatenated alignments as for the previous *E. coli* sets, representing a small fraction of the possible pairs formed by all *E. coli* proteins. Even if this set represents only a fraction of all possible proteins, it contains on average >300 pairs for each protein, adding relevance to the Z-score calculation associated to the interaction indices. The corresponding database entries and alignments can be found in http://pdg.cnb.uam.es/i2h.

These data sets include all pairs for which it was possible to obtain a sufficient number of sequences for the alignments. They do not correspond to any selection bias.

RESULTS

To test the quality of the i2h predictions, we first analyzed a previously derived set of two domain proteins of known three-dimensional structure. A large collection of

TABLE I. List of Pairs in the Structural Domains Data Set

Pair	Interaction index	Pair	Interaction index
2c2c_2-1alc_1	3.503	3adk_2-4nc_2	0.961
1sgt_2-4mt2_1	3.448	1alc_1-1md_2	0.957
9pap_1-9pap_2*	**3.042**	1sgt_1-2c2c_2	0.889
1alc_1-1alc_2*	**2.852**	2c2c_2-3pgk_1	0.878
2c2c_1-4mt2_1	2.825	3trx_1-9pap_2	0.857
4tms_1-4tms_2*	**2.735**	4tnc_1-4mt2_2	0.853
3trx_1-3trx_2*	**2.571**	4tnc_2-4mt2_2	0.836
4mt2_1-4mt2_2*	**2.469**	3trx_1-3pgk_2	0.829
2c2c_2-4mt2_1	2.355	3trx_1-9pap_1	0.814
2c2c_2-4mt2_2	2.331	2c2c_2-1rnd_2	0.813
4tnc_1-4tnc_2*	**2.238**	4tms_2-3dfr_2	0.809
3blm_1-3blm_2*	**2.206**	9pap_2-3adk_2	0.805
3pgk_1-3pgk_2*	**2.197**	4tms_1-3dfr_2	0.804
2c2c_1-4mt2_2	2.139	1sgt_2-1alc_1	0.799
1sgt_2-2c2c_1	2.068	9pap_1-3adk_2	0.790
2c2c_1-1alc_1	2.011	3trx_2-9pap_2	0.761
2c2c_1-1alc_2	1.886	4tnc_2-4mt2_1	0.747
3adk_1-3adk_2*	**1.862**	3adk_2-3pgk_2	0.726
1sgt_2-2c2c_2	1.835	4tnc_1-4mt2_1	0.718
2c2c_1-2c2c_2*	**1.787**	9pap_2-4tnc_2	0.702
3adk_1-3pgk_1	1.624	3trx_1-3adk_1	0.673
1rnd_1-4mt2_1	1.530	**3dfr_1-3dfr_2***	**0.657**
2c2c_1-9pap_2	1.520	2pf2_2-1alc_2	0.628
3adk_2-3dfr_2	1.507	3adk_1-4tnc_1	0.617
1sgt_2-2pf2_2	1.489	3adk_1-4tnc_2	0.614
9pap_1-3adk_1	1.488	2pf2_2-1alc_1	0.595
3adk_1-3pgk_2	1.444	3adk_2-4tnc_1	0.539
2c2c_2-1alc_2	1.415	4tms_1-3dfr_1	0.507
2c2c_1-3pgk_2	1.389	3trx_2-3pgk_1	0.489
1sgt_1-4mt2_1	1.387	3trx_2-3pgk_2	0.471
3adk_1-3dfr_1	1.367	3trx_1-3adk_2	0.471
1rnd_2-4mt2_1	1.359	1sgt_1-1alc_1	0.455
2c2c_2-3adk_1	1.319	3trx_1-2c2c_2	0.453
1rnd_1-1rnd_2*	**1.314**	3trx_1-2c2c_1	0.446
3pgk_1-4tms_1	1.299	4tms_2-4tnc_2	0.444
2c2c_1-3adk_1	1.297	2c2c_1-1rnd_2	0.442
3pgk_1-4tms_2	1.292	1sgt_2-1alc_2	0.435
3trx_1-3pgk_1	1.279	3trx_2-3adk_1	0.427
2c2c_1-3pgk_1	1.278	4tms_1-4tnc_2	0.413
1alc_1-4mt2_1	1.278	1sgt_1-1rnd_1	0.403
2c2c_2-9pap_2	1.274	4tms_1-4tnc_1	0.401
1rnd_1-4mt2_2	1.258	4tms_2-3dfr_1	0.398
3adk_2-3pgk_1	1.252	1alc_2-4mt2_2	0.362
1rnd_2-4mt2_2	1.240	1sgt_1-1rnd_2	0.358
3adk_1-3dfr_2	1.209	1sgt_1-4mt2_2	0.356
3trx_2-2c2c_1	1.196	1sgt_2-1rnd_1	0.352
3pgk_2-4tms_2	1.178	3trx_1-4tnc_2	0.316
3pgk_2-4tms_1	1.170	2c2c_1-4tnc_1	0.303
3adk_2-3dfr_1	1.136	4tms_2-4tnc_1	0.299
3trx_2-9pap_1	1.133	9pap_2-4tnc_1	0.289
1sgt_1-1sgt_2*	**1.113**	1sgt_1-1alc_2	0.278
1sgt_1-2pf2_2	1.098	9pap_1-4tnc_2	0.254
2c2c_2-9pap_1	1.094	2c2c_1-4tnc_2	0.236
2c2c_1-3adk_2	1.067	3trx_1-4tnc_1	0.233
1sgt_2-4mt2_2	1.063	1sgt_2-1rnd_2	0.229
3trx_2-2c2c_2	1.058	1alc_1-4mt2_2	0.227
2c2c_2-3pgk_2	1.047	2c2c_2-4tnc_2	0.218
1alc_2-4mt2_1	1.037	3pgk_1-4tnc_1	0.202
1alc_2-1rnd_1	1.033	3pgk_1-4tnc_2	0.199
2c2c_1-1rnd_1	1.029	9pap_1-4tnc_1	0.186
9pap_2-3adk_1	1.014	2c2c_2-4tnc_1	0.174
2c2c_2-3adk_2	1.008	3trx_2-3adk_2	0.161
2c2c_2-1rnd_1	1.004	3pgk_2-4tnc_1	0.150
1alc_2-1rnd_2	1.003	3pgk_2-4tnc_2	0.127
1sgt_1-2c2c_1	0.991	3trx_2-4tnc_1	0.103
2c2c_1-9pap_1	0.977	3trx_2-4tnc_2	0.070
1alc_1-1rnd_1	0.968		

*List of possible pairs in the structural domains data set (Fig. 2), with corresponding interaction index values. Pairs are labeled "pdbid1_domain1_pdbid2_domain2." The pairs corresponding to known interactions are in bold. The table is sorted by interaction index value.

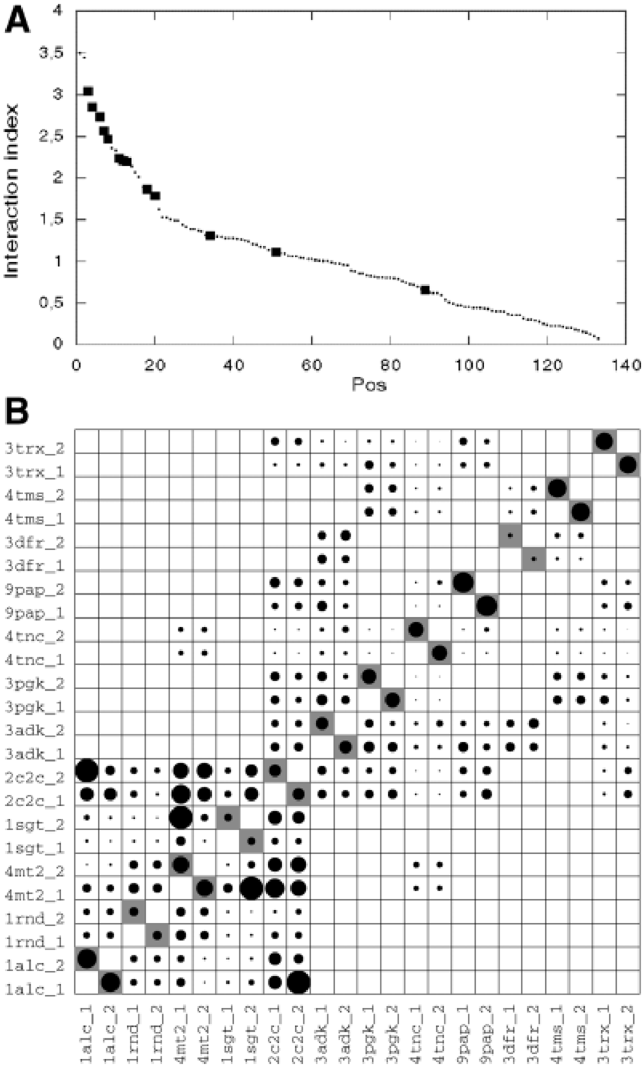

Fig. 2. Results obtained in a set of two domain proteins of known structure. The i2h method was applied to the set of 14 proteins with two "interacting" domains, previously used for the prediction interaction regions.[15] **A:** The interaction index is represented for the 133 pairs with 11 or more sequences in common. The true positive hits are highlighted with filled squares. **B:** Representation of i2h results, reminiscent of those obtained in the experimental yeast two-hybrid system. The diameter of the black circles is proportional to the interaction index; true pairs are highlighted with gray squares. Empty spaces correspond to those cases in which the i2h system could not be applied, because they contained <11 sequences from different species in common for the two domains.

possible domain pairs was prepared by treating different domains as independent proteins. In most cases, the i2h system scored the correct pair of protein domains above all other possible interactions (Table I, Fig. 2). Most of the false positives involve the domains of two proteins. One of them is a metallothionein (pdb code: 4mt2), a Cys-rich protein predicted to interact with many other domains as strongly as with its own second domain. We have no explanation for this behavior, which may be related to artifacts in the multiple sequence alignment produced by the peculiar sequence composition of this protein. The other false positive is cytochrome-C2 (2c2c). The result for this well-characterized set of proteins can be used as an indication of the relationship between score and quality of prediction. Scores > 1.5 can be considered as an initial indication of interaction; scores > 2.0 correspond mostly to true interactions.

Remember that i2h analysis is based on the prediction of physical proximity between residue pairs and, together with the predictions of interactions between proteins, it is possible to recover directly the predicted interacting residues.[15] In the case of the hemoglobin α/β dimmer, for example, the strongest interaction predictions were found for residues 37, 52, 88, and 84 in the α-subunit and 102, 57, 46, and 64 in the β-subunit. Together with a set of conserved residues, these form part of the interaction surface of the two subunits (data not shown). This observation adds value to the i2h predictions because they include not only the possible protein partners but also the prediction of their possible region of interaction.

A second test set was derived from that used by Dandekar et al.[5] to select those proteins for which we could obtain enough sequences in at least 11 genomes from a total of 14 used to build the alignments (see Materials and Methods). This set included 8 known pairs of interactions buried in 244 possible protein pairs. Figure 3 shows that seven of the known interactions were found among the high scoring pairs in i2h analysis. In addition, 11 protein pairs with a high interaction index were possible interacting proteins, including different ribosomal proteins and translation factors. Only one pair of known interacting proteins, SecD-SecF, gave a relatively low score. It is of interest that, although the interaction index of this pair is low in absolute terms, it is among the best for the pairs formed by SecD with other potential partners, indicating the possibility of detecting interactions based on comparison of their indices when the absolute score is insufficiently high.

We examined the influence of the species distributions on these results, that is, if the presence or absence of sequences of given species could always be related with high (or low) scores. The results are shown in Figure 3(c). There is no obvious relation between the species distribution ("phylogenetic profile") and the interaction index for pairs whose alignment of common species contains that distribution. Most of the phylogenetic profiles produce a complete range of scores, from low to high. There are only a few exceptions [e.g., the phylogenetic profile "11011111101111" (the alignment of common species contains all the 14 species but the ones from *N. Gonorrhoeae* and *T. Pallidum*)] that seem to be related always with low scores.

We analyzed a third set of protein interactions derived from the study by Marcotte et al.[9] The results using the i2h system revealed an interesting complementation with the approach of Marcotte et al. This is not surprising, because the i2h system is based on the prediction of physical interactions, and Marcotte's method addresses the prediction of general functional relations. From their protein set we extracted 195 pairs of proteins that contained enough sequences for application of i2h analysis, including 15 of the interactions predicted by Marcotte et al. Three were strongly predicted as physically interacting pairs by the i2h system, with absolute scores > 2.0; two other pairs were found in the range of significant predictions (scores > 1.5). Among those pairs predicted by both systems, we found, for example, thioredoxin with Thr-rRNA synthase and Mo cofactor biosynthesis protein C with Mo Cofactor biosynthesis protein A. For 10 other cases, i2h predictions did not agree with Marcotte's predictions. For example, the interaction predicted by Marcotte et al. between a hypothetical protein similar to ferredoxin reductase and thioredoxin may be due to their functional relation as part of the redox-related reactions rather than the physical interaction required by the i2h analysis. The i2h system predicts additional interactions between proteins in that set, including seven pairs with scores > 2.0. Even if some are false positives, it is difficult to disregard them completely (e.g., certain interactions between ribosomal proteins and aminoacyl-tRNA synthetases).

The results for the set of interacting proteins of known structure extracted from the SPIN database are shown in Figure 4. Again, most of the pairs of interacting chains plot at high scores. Among the false positives we find, for example, the pair formed by 1aisA with 1volB (two TATA binding proteins), many pairs among hemoglobins, 1lgbC-1hdsA (N2 fragment of lactotransferrin with hemoglobin), and 1cpcB-1phnA (two phycocyanins). Among the false negatives we find 1ttpB-1ttpA (subunits of the tryptophan synthase), 1aqdA-1aqdB (two chains of the Hla-Dr1 class II histocompatibility complex), and 1outA-1outB (α- and β-chains of trout hemoglobin).

Finally, we applied the i2h system to a large collection of *E. coli* proteins to show the feasibility of the systematic reconstruction of interaction networks in full genomes. In its current state, the database contains 67,238 pairs, for which we found at least 11 homologous sequences of common species for the two proteins in each pair. The extension of this experiment will require enlargement of the corresponding alignments, increasing the number of protein pairs and thus the number of potential interaction candidates. The number of predicted interactions at different interaction index cutoffs is shown in Figure 5.

Despite the limited size of the experiment, interesting relationships were found among the high scoring pairs,

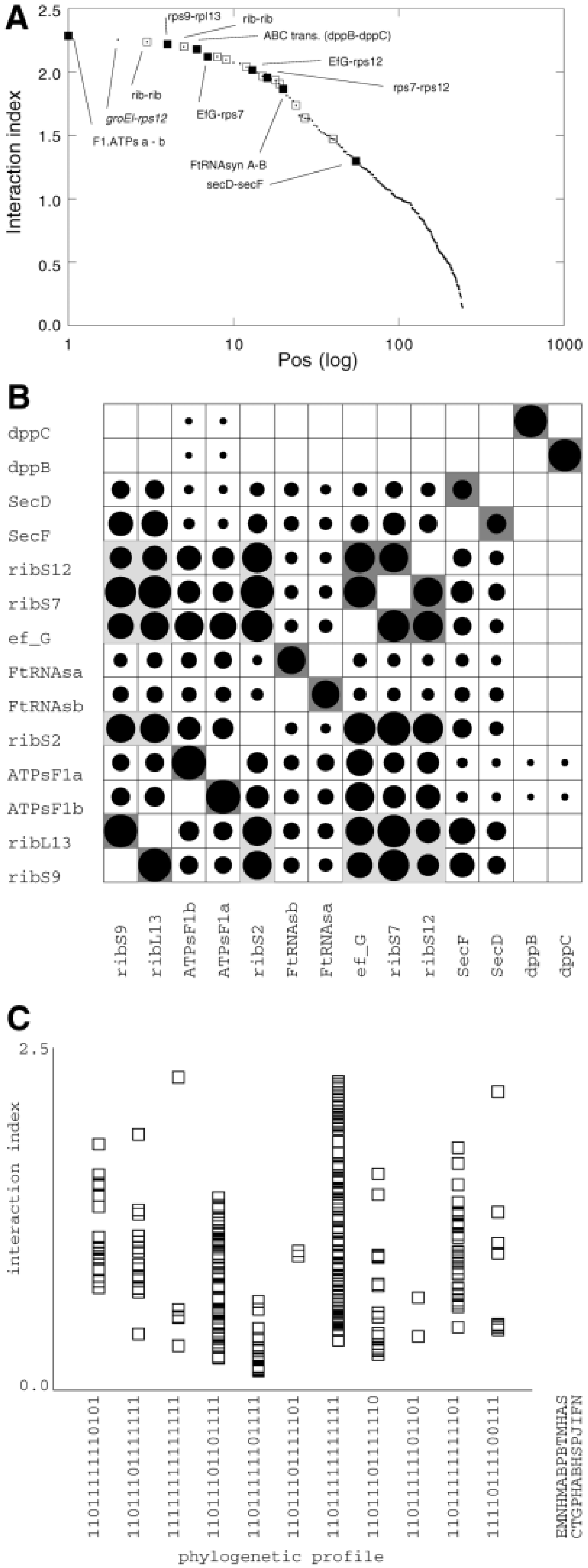

Fig. 3. Results obtained in the Dandekar data set. The i2h method was applied to the set of bacterial interacting proteins analyzed by Dandekar et al.,[5] using multiple sequence alignments compiled from 14 fully sequenced genomes. **A:** The interaction index is represented for the 244 possible pairs as in Figure 2(a). In this case, possible interactions are indicated with empty squares, including different ribosomal proteins and elongation factors. **B:** Representation of i2h results reminiscent of the typical representation of yeast two-hybrid experimental data. In this case, a subset of the results of (A) is represented, corresponding to proteins that form part of protein pairs with experimentally verified interactions and protein families with enough alignments. As in Figure 2(b), the diameter of the black circles is proportional to the interaction index, positive cases are highlighted with dark gray squares, and plausible interactions with light gray squares. Empty spaces correspond to those cases with <11 sequences from different species in common, as in Figure 2. **C:** Values of interaction indexes for the different phylogenetic profiles in this data set. A phylogenetic profile represents the pattern of presence (1)/absence (0) of that species in the alignment of common species for a pair of proteins. Abbreviations for the names of the species are shown at the right. The values of interaction indexes for all pairs of proteins containing a given phylogenetic profile are drawn.

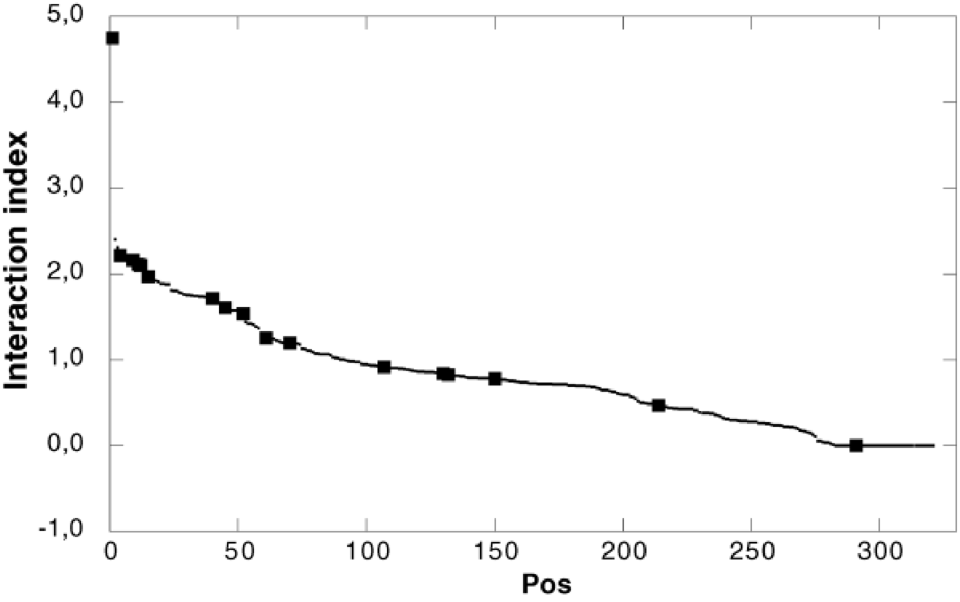

Fig. 4. Results obtained with the set of interacting proteins of known structure. Results of the application of the i2h method to the set of interacting proteins of known structures extracted from the SPIN database (see Materials and Methods). The interaction index is represented for the 321 pairs calculated. The black squares correspond to the 17 real complexes.

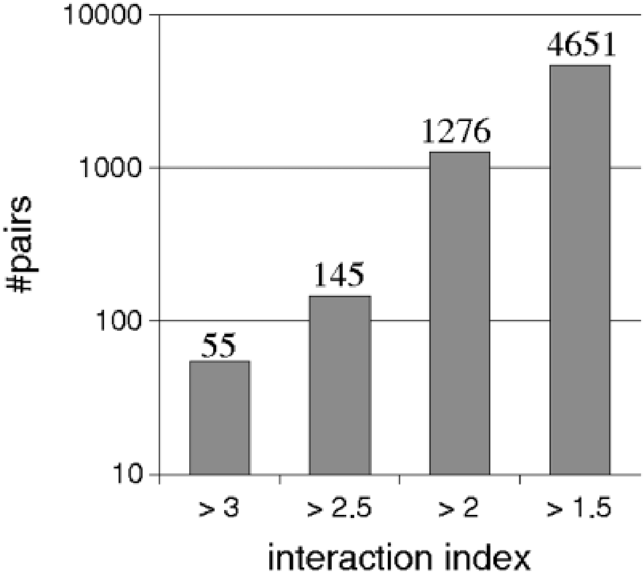

Fig. 5. Number of predicted interactions for *E. coli.* The bars represent the number of predicted interactions obtained from the 67,238 calculated pairs, depending on the interaction index cutoff established as a limit to consider interaction. The *y* axis scale is logarithmic.

corresponding in many cases to known interacting proteins such as membrane transporters implicated in spermidine/putrescine transport (PotB-PotH), transcription factors of the bacterial two-component system (ArcA-TorR), and the two isoforms (N and O) of formate dehydrogenase. Other interesting interactions were predicted, including different hypothetical proteins for which the possible relationships could be a first clue to their function. For example, YABK_ECOLI (Swissprot ID code) was predicted to interact with different iron and zinc transporters. Further analysis revealed that this hypothetical protein contains a possible transmembrane region, as well as distant similarity to an iron transporter from *S. marcescens*. Both observations added credibility to its possible function as part of a metal transport system (Fig. 6).

DISCUSSION

The application of i2h analysis to different sets of interactions revealed a considerable capacity for detection of true interactions with distinct scores. The information provided by the i2h systems is an indication not only of a possible interaction but also of the possible protein region involved. The main practical limitation of the current application of the i2h system is the difficulty in obtaining large multiple sequence alignments of corresponding sequences for each possible pair of proteins. This problem may disappear with the incorporation of new sequences from the continuous stream of newly sequenced genomes. Regarding the prediction of correlated mutations, it was previously shown that more informative alignments led to improved predictions of intraprotein contacts.[12]

Another important limitation, common to the other approaches for the prediction of protein interaction networks,[5,9,10,23] is the difficulty in obtaining good test sets to quantify the performances of the different methods.

Current experimental descriptions of full interaction networks (i.e., systematic applications of yeast-2-hybrid screenings) are still far from complete in both number of known interactions and reliability.[24] Moreover, known interactions are only a small part of the many still unknown reactions of biological relevance. The development of the field of protein interaction prediction would require the creation of gold standards as has occurred in

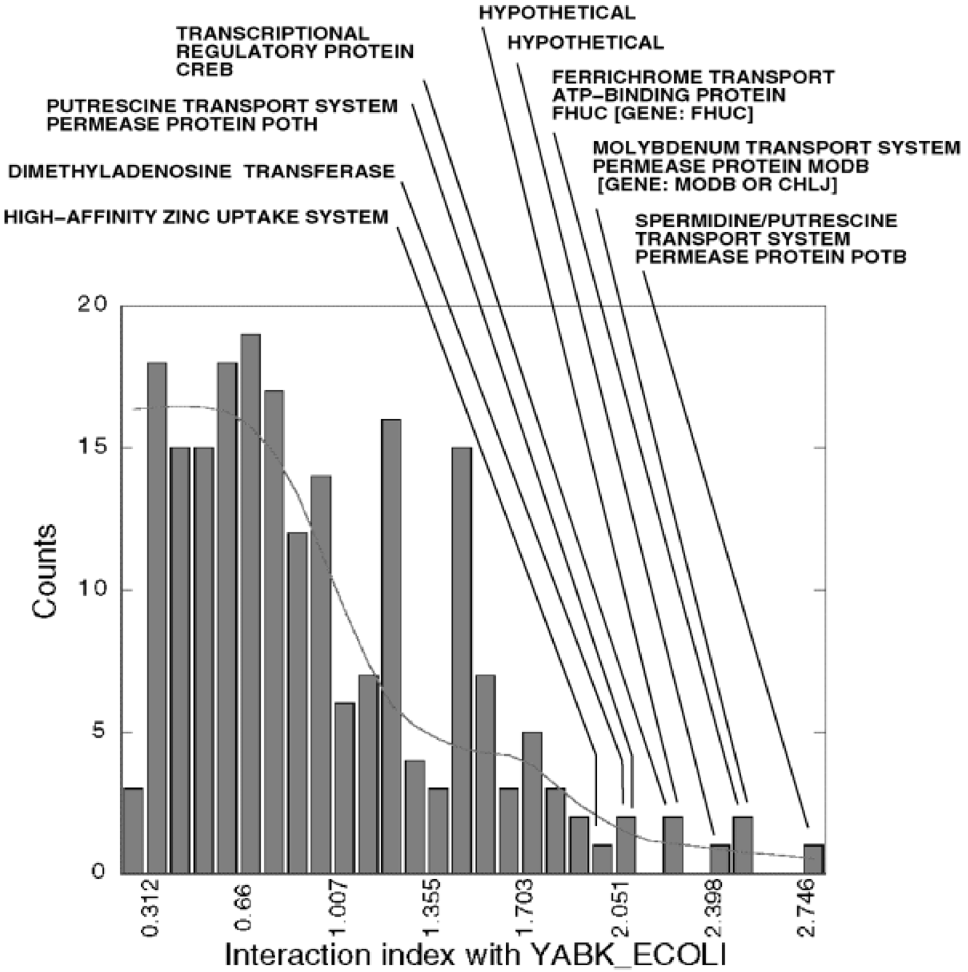

Fig. 6. Example of data analysis using the *E. coli* i2h database. Analysis of predicted interaction partners for the hypothetical protein YABK_ECOLI, one of the *E. coli* proteins included in the prototype database. The interaction index distribution for the different possible pairs is compared in an interactive Web-based interface that facilitates inspection of their functions by following links to the information deposited in Swissprot[35] and other databases, localization in the *E. coli* genome, and the possible relationship to *E. coli* operons. In this case, the different functions highlight the relationships of the hypothetical protein with iron and zinc transport mechanisms, as well as with other hypothetical proteins. The experimental i2h *E. coli* database can be accessed at http://pdg.cnb.uam.es/i2h.

fields such as gene finding, threading, and secondary structure prediction. Efforts are underway to construct databases of annotated interactions. For example, DIP[25]: http://dip.doe-mbi.ucla.edu; MIPS: http://www.mips.biochem.mpg.de.

Because of these difficulties in obtaining good test sets, other methods have been tested by analyzing the presence of similar keywords in their database descriptions[9] or by manually screening some of the proposed interactions.[5] We used a first test set composed of protein domains, because it provides unambiguous information about stable, permanent protein interaction. In addition, we tested our method with some of the data sets previously used by other authors finding a reasonable level of accuracy.

The reduction of the initial multiple sequence alignments, taken only sequences from common species (see Materials and Methods), restricts the number of available sequence pairs and the quality of the alignments. We previously showed that including a broad range of sequences in the alignments improves the predictions of intraprotein contacts.[12] Therefore, it is conceivable that the implementation of a more sophisticated process for sequence selection (i.e., selecting putative orthologous sequences) would improve the results even if it would imply an increase in computational complexity.

The results derived from the i2h system may also be improved by increasing the accuracy of the basic correlation method used to detect interaction sites, by reformulating the current algorithm (see Refs. 26–29 for alternative approaches) or by combining correlated mutations with other methods for the prediction of functionally important residues.[30–32]

We aim to apply the i2h system to the prediction of networks of physical interactions in complete genomes to provide suggestions about protein interactions and their biological function. The large set of interactions for the *E.*

coli genome here presented could be considered a first step in this direction. The experience of Marcotte et al.,[23] combining theoretical and experimental data, provides an excellent departure for the inclusion of the i2h predictions, because it represents a distinct type of prediction concerned with the detection of physical interactions.

ACKNOWLEDGMENTS

We thank A. de Daruvar (Lion Biosciences), J.L. Carrascosa (CNB-CSIC), and K. Harshman (DIO-CNB) for critical reading of the manuscript and suggestions and members of the Protein Design Group for continuous support.

REFERENCES

1. Nitschke P, Guerdoux-Jamet P, Chiapello H, Faroux G, Henaut C, Henaut A, Danchin A. Indigo: a World-Wide-Web review of genomes and gene functions. FEMS Microbiol Rev 1998;22:207–227.
2. Mendelsohn AR, Brent R. Protein interaction methods—toward an endgame. Science 1999;284:1948–1950.
3. Lengauer T, Rarey M. Methods for predicting molecular complexes involving proteins. Curr Opin Struct Biol 1996;5:402–406.
4. Dixon JS. Evaluation of the CASP2 Docking Section. Proteins 1997;S1:198–204.
5. Dandekar T, Snel B, Huynen M, Bork P. Conservation of gene order: a fingerprint of proteins that physically interact. Trends Biochem Sci 1998;23:324–328.
6. Gaasterland T, Ragan MA. Microbial genescapes: phyletic and functional patterns of ORF distribution among prokaryotes. Microb Comp Genomics 1998;3:199–217.
7. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci USA 1999;96: 4285–4288.
8. Tatusov RL, Koonin EV, Lipman DJ. A genomic perspective of protein families. Science 1997;278:631–637.
9. Marcotte EM, Pellegrini M, Ho-Leung N, Rice DW, Yeates TO, Eisenberg D. Detecting protein function and Protein–protein interactions from genome sequences. Science 1999;285:751–753.
10. Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA. Protein interaction maps for complete genomes based on gene fusion events. Nature 1999;402:86–90.
11. Göbel U, Sander C, Schneider R, Valencia A. Correlated mutations and residue contacts in proteins. Proteins 1994;18:309–317.
12. Olmea O, Valencia A. Improving contact predictions by the combination of correlated mutations and other sources of sequence information. Fold Design 1997;2:S25–S32.
13. Pazos F, Olmea O, Valencia A. A graphical interface for correlated mutations and other structure prediction methods. Comput Appl Biosci 1997;13:319–321.
14. Olmea O, Rost B, Valencia A. Effective use of sequence correlation and conservation in fold recognition. J Mol Biol 1999;293:1221–1239.
15. Pazos F, Helmer-Citterich M, Ausiello G, Valencia A. Correlated mutations contain information about Protein–protein interaction. J Mol Biol 1997;271:511–523.
16. Mclachlan AD. Test for comparing related aminoacid sequences. J Mol Biol 1971;61:409–424.
17. Press WH, Teukolsky SA, Vetterling WT, Flannery BP. Numerical recipes in C: the art of scientific computing. Cambridge: Cambridge University Press; 1992.
18. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215:403–410.
19. Higgins DG, Bleasby AJ, Fuchs R. CLUSTAL V: improved software for multiple sequence alignment. Comput Appl Biosci 1992;8: 189–191.
20. Sander C, Schneider R. The HSSP data base of protein structure-sequence alignments. Nucleic Acids Res 1993;21:3105–3109.
21. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
22. Murzin AG, Brenner SE, Hubbard T, Chotia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.
23. Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D. A combined algorithm for genome-wide prediction of protein function. Nature 1999;402:83–86.
24. Legrain P, Wojcik J, Gauthier JM. Protein–protein interaction maps: a lead toward cellular functions. Trends Genet 2001;17:346–352.
25. Xenarios I, Fernandez E, Salwinski L, Duan XJ, Thompson MJ, Marcotte EM, Eisenberg D. DIP: the database of interacting proteins: 2001 update. Nucleic Acids Res 2001;29:239–241.
26. Altschuh D, Lesk AM, Bloomer AC, Klug A. Correlation of coordinated amino acid substitutions with function in virus related to tobacco mosaic virus. J Mol Biol 1987;193:693–707.
27. Chelvanayagam G, Eggenschwiler A, Knecht L, Gonnet GH, Benner SA. An analysis of simultaneous variation in protein structures. Protein Eng 1997;10:307–316.
28. Pollock DD, Taylor WR, Goldman N. Coevolving protein residues: maximum likelihood identification and relationship to structure. J Mol Biol 1999;287:187–198.
29. Shindyalov IN, Kolchanov NA, Sander C. Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations. Protein Eng 1994;7:349–358.
30. Andrade MA, Casari G, Sander C, Valencia A. Classification of protein families and detection of the determinant residues with an improved self-organizing map. Biol Cybern 1997;76:441–450.
31. Casari G, Sander C, Valencia A. A method to predict functional residues in proteins. Nat Struct Biol 1995;2:171–178.
32. Lichtarge O, Bourne HR, Cohen FE. An evolutionary trace method defines binding surfaces common to protein families. J Mol Biol 1996;257:342–358.
33. Bairoch A, Boeckmann B. The SWISS-PROT protein sequence data bank. Nucleic Acids Res 1992;20:2019–2022.