Improved Homology-driven Computational Validation of Protein-Protein Interactions Based on Evolutionary Gene Duplication and Divergence Hypothesis

Christian Frech*[1], Michael Kommenda[1], Viktoria Dorfer[1], Thomas Kern[1], Helmut Hintner[2], Johann W. Bauer[2] and Kamil Önder[2,3]

[1]Upper Austria University of Applied Sciences, Softwarepark 11, 4232 Hagenberg, Austria
[2]Paracelsus Medical Private University, Department of Dermatology, Müllner Hauptstraße 48, 5020 Salzburg, Austria
[3]Department of Cell Biology, University of Salzburg, Salzburg, Austria

Email: Christian Frech*- christian.frech@fh-hagenberg.at; Michael Kommenda - michael.kommenda@fh-hagenberg.at; Viktoria Dorfer - viktoria.dorfer@fh-hagenberg.at; Thomas Kern - thomas.kern@fh-hagenberg.at; Helmut Hintner - h.hintner@salk.at; Johann W. Bauer - jo.bauer@salk.at; Kamil Önder - k.oender@salk.at;

*Corresponding author

Abstract

Background: Protein-protein interaction (PPI) data sets generated by high-throughput experiments are contaminated by large numbers of erroneous PPIs. Therefore, computational methods for PPI validation are necessary to improve the quality of such data sets. Against the background of the theory that most extant PPIs arose as a consequence of gene duplication and divergence, we investigated if traditional homology-based concepts for PPI validation can be further improved by a more comprehensive search for homologs.

Results: To validate a putative PPI we combine FASTA and PSI-BLAST to perform a sequence-based search for pairs of interacting homologous proteins within a large, integrated PPI database. A normalized scoring scheme that incorporates both the quality and quantity of all observed matches allows us (1) to consider also tentative paralogs and orthologs in the analysis and (2) to easily combine the search results obtained by FASTA and PSI-BLAST. ROC curves illustrate the high efficacy of this approach.

Conclusions: The duplication-divergence model of PPI evolution suggests that for true PPIs many homologous PPIs exist, not only among close relatives but also among remote homologs. We demonstrated that a validation technique that consequently exploits this idea is very efficient and improves over traditional homology-based

concepts. In particular, our insights might be useful in cases where traditional homology-based techniques are not an option owing to a lack of assured paralogs or orthologs.

Background

Physical interactions between proteins, commonly referred to as protein-protein interactions (PPIs), occur at every level of cell function to elaborate the organism's phenotype. The study of PPIs is therefore of great interest and is helping to reveal basic molecular mechanisms of many diseases. High-throughput screening methods have given insight into hundreds of thousands of potential PPIs in several organisms. However, a major disadvantage of high-throughput approaches is their high rate of *false-positive* PPIs, i.e. erroneously reported PPIs that do not occur *in vivo* [1–7].

The development and implementation of computational methods for PPI validation is therefore an important goal in bioinformatics today. Common approaches include determining intersections between different high-throughput PPI data sets [3], incorporating protein annotation data [5, 8], analyzing expression profiles [4, 9–12], investigating topological criteria of PPI networks [13–17], and inspecting patterns of co-evolution [18].

Another established *in silico* technique to validate a pair of physically interacting proteins is to search for homologs that also interact; if found, the confidence of the questioned PPI is increased. The original *interolog* concept suggests to examine PPIs among *bona fide orthologous* proteins in other species, i.e. functionally conserved proteins that evolved from a common ancestor [19, 20]. However, large-scale application of this method for PPI validation is strongly hampered by the limited coverage of most interactomes and the small number of known *bona fide* orthologs [21]. Another, first practical approach involved the inspection of PPIs among *paralogous* proteins, i.e. homologous proteins that evolved by gene duplication and are found within the same species [4]. However, sensitivity remains a major problem because in most organisms assured paralogs with known interactions are scarce. The strategy we follow here inspects homologous proteins independent of species boundaries and functional conservation (Figure 1). Several papers applied this 'all-inclusive' approach to homology-based PPI validation [8, 22, 23]. Also techniques developed for PPI prediction, a relatively more well-studied bioinformatics problem, successfully utilized this idea, for example Brown *et al.* [24] or Jonsson *et al.* [25].

The aim of the present paper is threefold. First, we draw the reader's attention to the duplication-divergence hypothesis of PPI evolution, i.e. the idea that extant PPIs primarily originated from gene duplications, the homologs diverging over time. Although this hypothesis has been an acknowleged theory for some time now we think that its implications for homology-based PPI validation have not been fully recognized so far. In particular, the implication we stress and explore here is that not only interactions among conserved homologs with high sequence similarity convey useful information about a PPI's biological relevance but also interactions among distant, putative homologs.

Second, we emphasize the high efficacy of homology-based validation when carried out on large PPI data sets. We compiled a comprehensive data set of known physical binary PPIs from six PPI source databases, comprising 135,276 PPIs from 20 different organisms. This is, to the best of our knowledge, the largest collection of PPIs that has been used so far in this kind of analysis.

Third, we propose an improved, sequence-based technique for homology-based PPI validation. Unlike previously published, mostly binary validation schemes that deem a questioned PPI as biologically relevant as soon as a single homologous PPI is found, we follow a similar approach as Jonsson *et al.* [25] and compute a confidence score that takes into account both the quality and quantity of all identified homologous PPIs. By assigning higher scores to high-quality hits and lower scores to low-quality hits we are able to extend our analysis from reliable homologs to highly putative paralogs and orthologs with E-values up to 10. In addition, we propose a normalization of obtained scores, which allows us to combine search results from multiple homology search strategies, in our case FASTA and PSI-BLAST. Based on ROC curves we show that with our approach it is possible to improve over traditional methods for PPI validation.

Results and Discussion

Duplication-Divergence Hypothesis of PPI Evolution

Gene duplication is a ubiquitous mechanism in molecular evolution and the principal source of biological innovation, producing new proteins and novel functional domains [26–30]. Here, we follow the idea that the duplication of genetic material coupled with subsequent divergence is also the dominant mechanism for the development of novel PPIs [31]. This hypothesis is supported by both theoretical models [32–34] and empirical evidence [35–40]. A brief review of these papers can be found in our Supplementary Material.

Duplication-divergence models of PPI evolution propose a simple and yet plausible idea of how evolution might have formed PPI networks over millions of years, namely by repeated duplication of genetic material

followed by subsequent divergence. Under this model, a PPI can be considered to be true if *bona fide* homologous PPIs can be discerned, i.e. PPIs that descended from a common ancestral PPI. Figure 2 illustrates this idea.

Implications of the Duplication-Divergence Model for Homology-Based PPI Validation

The duplication-divergence model of PPI evolution suggests that for homology-based PPI validation the actual similarity between homologous proteins is of minor importance: so long as two interacting proteins are each truly homologous to another pair of interacting proteins, there is a good chance that their PPIs are homologous as well and therefore biologically relevant (Figure 2). Note that though from this evolutionary perspective homology is conclusive for PPI validation it is generally not for homology-based PPI *prediction*, because most duplicated PPIs are eventually lost. Thus the inference of PPIs by homology is either only reliable if two proteins are very similar [41, 42] or if an inferred PPI is also confirmed by complementary data [24]. Consequently, we consider homology-based PPI validation and homology-based PPI prediction as different problem domains.

Weak Homologous Interactions - Signal or Noise?

If the duplication-divergence model of PPI evolution is correct we would expect to find many homologous PPIs not only among closely related proteins, but also among weak paralogs and orthologs. In fact, the existence of weak homologous PPIs should be an important characteristic of biologically relevant PPIs. We analyzed if this is actually the case. A second question we addressed here is whether such a characteristic, if it exists, could be useful for PPI validation.

For both our Gold Standard Positive (GSP) and Gold Standard Negative (GSN) data sets we used PSI-BLAST with an E-value threshold of 10 to search for homologous PPIs and determined their distribution within different E-value windows. Figure 3 shows the results. It reveals two important differences between GSP PPIs and GSN PPIs. First, there is an increased probability for GSP PPIs to have paralogous and orthologous PPIs at all. Second, significantly more GSP PPIs than GSN PPIs have large numbers of paralogous and orthologous PPIs (>10). Most interestingly both differences are observable up to high E-value windows.

Not surprisingly, the first characteristic, the existence of at least one homologous PPI, is a highly reliable signal for GSP PPIs when sequence similarity is high. For example, almost every fifth PPI taken out of the GSP data set (18%) has a homologous PPI with an E-value lower than 10^{-100} (Figure 3A). By contrast,

the existence of such 'high-quality' homologs is extremely unlikely for a GSN PPI (0.25%). Although the strength of this signal drops with a reduction in sequence similarity, it remains intact up to high E-values: within the last E-value window (ranging from 3 to 10) the probability of observing a homologous PPI for a GSP PPI remains still twice as high (63%) as for a GSN PPI (30%).

The distribution of homologous PPIs reveals the second interesting characteristic of GSP PPIs: they tend to accumulate large numbers of homologs. According to Figure 3A, from E-value 10^{-20} upwards about 25% of all GSP PPIs have more than 10 homologous PPIs in every E-value window; for GSN PPIs, this percentage never exceeds 8%. Thus, in many cases the existence of a large number of homologous PPIs seems to be more conclusive than the mere existence of any homologous PPI, especially when sequence similarity is low: whereas about twice as many GSP PPIs than GSN PPIs have at least one homologous PPI within the last E-value window, almost four times as many GSP PPIs (21.4%) than GSN PPIs (5.8%) have between 10 and 100 homologs, and more than five times as many GSP PPIs (6.8%) than GSN PPIs (1.3%) have more than 100 homologs.

Both characteristics are observed independently of the fact whether only paralogous (Figure 3B) or only orthologous PPIs (Figure 3C) are investigated, although for lower numbers of homologous PPIs we obtain stronger signals on the paralogous data set than on the orthologous data set. This is consistent with the finding that PPIs seem to be more conserved within species than across species [41]. Interestingly very large numbers (>100) of homologous PPIs are only observed within the orthologous data set, most likely due to an increased number of gene duplications in higher eukaryotes.

From our analysis we conclude (1) that the existence of large numbers of weak homologous PPIs is in fact in many cases a distinguishing characteristic of biologically relevant PPIs and (2) that therefore weak homologs should be taken into account by homology-based PPI validation techniques because they convey useful information. Our scoring scheme is an attempt to make use of this insight.

Overall Performance

The Receiver Operating Characteristic (ROC) curves in Figure 4 illustrate the overall performance of our scoring scheme for the *MIPS*, the *Small Scale* and the *Multiple Evidence* gold standard data sets. The y-axis shows the True-Positive Rate (TPR or *sensitivity*), i.e. the percentage of GSP PPIs that were correctly confirmed as biologically relevant. The x-axis represents the False-Positive Rate (FPR or 1-*specificity*), i.e. the percentage of GSN PPIs that were erroneously confirmed as biologically relevant. By varying the threshold of the score above which a PPI is confirmed as biologically relevant, we observe

different FPRs and TPRs (a more detailed introduction to ROC curves can be found in our Supplementary Material).

For example, on the *MIPS* and the *Multiple Evidence* data sets, we observe a TPR of more than 70% at an FPR of 10%. With an increased threshold we observe a TPR of 80% at an FPR of 20% for the same two data sets. This performance compares very well with other, not homology-based validation techniques for which TPRs and FPRs have also been established on gold standard data sets from yeast [11–14, 18]. This underscores the high efficacy of homology-based PPI validation, especially when carried out on rich PPI data sets as done in this study.

Interestingly, the *Small Scale* gold standard performed worst, which might reflect differences in the quality of our data sets. We consider the *Multiple Evidence* gold standard to be the highest quality data set because detection of a PPI with different experimental methods is a very reliable indicator of a PPI's biological relevance [3]. Indeed, this data set achieves the best performance up to a TPR of 70%. The *MIPS* gold standard contains PPIs audited by human experts and is thus very trustworthy as well, although we observe a slightly poorer performance using this data set. PPIs of the *Small Scale* gold standard are not reviewed manually and thus its reliability might reflect the quality of the automated text mining tools that are frequently used to extract the PPIs from the scientific literature. Because these tools are error-prone [43], the *Small Scale* gold standard might contain more spurious PPIs than the other two data sets.

Note that although we have a skewed class distribution in our gold standard data sets, i.e. our *Random* GSN data set is about 50 times larger than each GSP data set, this does not affect the overall ROC curve [44]. In fact, we observed the same overall ROC curve on a balanced data set where the number of randomly chosen GSNs roughly equaled the number of GSPs (data not shown). Further notice that ROC curves as shown in Figure 4 are ideal to illustrate the overall performance of a classifier but do not make suggestions about which specific score threshold should be actually applied in order to classify a PPI as true or false. This decision always depends on the TPR and FPR one is willing to accept (refer to Supplementary Figure 1 for selected score thresholds and their associated TPRs and FPRs).

Comparison of Homology-Based Validation Schemes

We were interested in knowing the performance of previously published homology-based validation methods on our data set and if our scoring scheme can actually improve over these methods. Previous methods involve simpler, binary selection processes in which a PPI is deemed to be biologically relevant as

soon as a single homologous PPI is found below a certain E-value threshold [4, 8, 23]. Figure 5 illustrates the results. Note that we did not compare our scoring scheme with homology-based PPI *prediction* techniques because these techniques generally include non-homology-based criteria in order to assess the biological significance of a PPI, which makes a direct comparison difficult.

The FASTA-based binary validation scheme shows remarkably high specificity on our data set, even at high E-values. For example, inclusion of homologous PPIs with E-values between 10^{-4} and 10 resulted in an increase of the TPR of almost 25% (from 48.5% to 72.1%) whereas the FPR grew by only 13.5%, remaining below 16%. This is remarkable if we consider the low sequence similarity and the probable existence of many spurious hits at E-values up to 10. By contrast, the PSI-BLAST-based binary validation scheme is less specific (even at low E-values) but much more sensitive: almost 87% of our GSP PPIs had at least one homologous PPI identified by PSI-BLAST (E-value $\leq$10). In addition to FASTA and PSI-BLAST we evaluated the use of BLAST for homology detection as well (data not shown). In comparison to FASTA we observed no noticeable difference in performance apart from a slight decrease in maximum sensitivity (about 2% lower than with FASTA).

Our scoring scheme, represented by the blue curve (squares), combines evidence from homologous PPIs found by FASTA and PSI-BLAST and clearly outperforms both of the individual binary validation schemes. For example, at a TPR of 70% our approach produces 4% fewer false-positives than the FASTA-based binary approach, and about 6% fewer false-positives than the PSI-BLAST-based binary validation scheme.

We also tested specific parameter settings suggested by previous homology-based validation schemes. Saeed and Deane [23] used BLAST with an E-value up to 10^{-4} to identify homologous PPIs and evaluated a TPR of 63% at an FPR of 7%. If we transfer this setting to FASTA and apply it to our data sets we observe a TPR of 48.5% at an FPR of 2.3%. Our scoring scheme, by contrast, produces only 1.5% false-positives at the same level of sensitivity (48.5%). The difference in FPR increases for higher levels of sensitivity, which illustrates the additional value arising from the incorporation of multiple methods for homology detection and from the incorporation of weak homologs.

Patil and Nakamura [8] used PSI-BLAST with an E-value up to 10^{-8} and reported a TPR of 89.7% at an FPR of 37.1% for their gold standards. If we use the same parameter setting on our data sets, we observe a significantly worse TPR of 70.1% at an FPR of 16.1%. Again, our scoring scheme outperforms this result and achieves a reduced FPR of 10% at the same level of sensitivity (70.1%). Concerning the work of Patil and Nakamura it is also noteworthy that our exclusively sequence-based scoring scheme produces a

superior overall ROC curve than their Bayesian network approach, which incorporates three genomic features (sequence, structure and annotation information). Again this underscores the potential efficacy of homology-based validation.

We found no published PSI-BLAST parameters in the paper from Deane *et al.* [4] and therefore did not compare our results with theirs.

Contribution of Weak Homologs

Finally, we address the question if and to what extent weak homologs contribute to the overall classification performance of our scoring scheme. Is it actually beneficial to include homologous PPIs with high E-values (>1) for PPI validation, i.e. do weak homologs indeed contribute positively in terms of increased sensitivity and/or increased specificity? To answer this question, we evaluated the classification performance of our scoring scheme with and without the inclusion of weak homologs. Figure 6 shows the results.

The inclusion of weak homologous PPIs generally contributes positively to the overall classification performance. For example, by restricting our analysis to homologous PPIs with an E-value below 10^{-10}, we see a maximum TPR of 69% at an FPR of 14.5%. When we extend our analysis and include also homologs with E-values up to 1, we achieve the same sensitivity at a significantly reduced FPR of 10%. Another increase of the E-value threshold up to 10 leads to a further reduction of the FPR by 1%.

Note that a similar effect cannot be observed for the classic, binary validation schemes, where for a less stringent E-value threshold, an increase in sensitivity is always accompanied by a decrease in specificity (Figure 5). This underlines the additional value of our scoring approach: it finds significant evidence for biologically relevant PPIs among weak homologs without the compromise of an increased rate of false alarms. Although the additional benefit resulting from the inclusion of weak homologs with an E-value above 1 is rather low for this data set, we expect it to increase for data sets where high-quality homologs are not at hand. Yeast is comparably well investigated, with approximately 50% of its estimated 40,000 to 75,000 PPIs already known [45]. As a consequence, most of its biologically relevant PPIs have homologous PPIs among high-quality paralogs, and matches among weak homologs add little extra value to the overall score. This situation is different from most other organisms where interactome coverage is far below 50% and where weak paralogs and orthologs are often the only possibility to validate a questioned PPI.

Conclusions

Knowledge of PPIs is key to understanding cell function. Although high-throughput PPI detection techniques are now making it possible to catalogue all PPIs of a cell, the notoriously high error rates of these methods are a major obstacle to achieving this ambitious goal. Computational methods that can efficiently separate the PPI wheat from the chaff are therefore highly desirable.

We think that recent insights into the evolution of PPIs, in particular the duplication-divergence hypothesis, might be crucial to this endeavor. Nature is a tinkerer, not an inventor [46]. For PPIs this means that new PPIs are primarily adapted from pre-existing PPIs rather than invented *de novo*. Consequently, most true-positive PPIs must have a long evolutionary track record with many homologous PPIs, not only among highly similar homologs but also among distant relatives. This important characteristic of biologically relevant PPIs should, in principle, allow successful discrimination between true and false PPIs.

Homology-based validation techniques therefore seem very promising, but curiously they have not gained much attention so far. Literature searches revealed just five papers that proposed a homology-based PPI validation technique on a large scale, only three of which presented a critical assessment of the method's performance. Presumably this reflects the fact that homology-based validation requires having at hand a set of PPIs among homologous proteins, when few such PPIs have been known. However, with more and more PPIs now being reported from high-throughput experiments, this limitation is no longer factor.

In this paper, we made use of a large PPI data set to re-assess the potential of homology-based PPI validation. We showed that the classic, binary validation technique is already very efficient but can be further improved by using multiple methods for homology detection and more remote homologs to complement close homologs.

We expect our findings to be most relevant in situations where interactions among assured paralogs or orthologs are not at hand and thus traditional homology-based validation via high-quality interologs is not an option. Existing PPI databases could use our or a similar homology-based technique to significantly reduce their number of false-positive PPIs without the risk of losing too many biologically relevant ones, especially within the better explored model organisms. Other prospective applications of our findings include the elucidation of physically interacting proteins from known protein complexes or the validation of *in silico* predicted PPIs in cases where homology was not used as a criterion for prediction in advance.

Prospective improvements may involve more sophisticated methods for homology detection (e.g. Profile-HMMs), identification of PPI-mediating protein features prior to homology detection to refine the

selection of homologs, and an assessment of the statistical significance (P-values) of computed scores to obtain an intuitive measure of a PPI's validity.

Methods

Database Search for Homologous PPIs

The identification of *bona fide* homologous PPIs is difficult. On the one hand, the modular architecture of proteins implies that a protein actually has not just one distinct evolutionary history, but one for each biological feature it contains. Given that in general the protein features (e.g. the domains) that mediate a specific PPI are not known, it is difficult to determine which of these evolutionary histories is relevant. Hence, all features of a protein need to be investigated, which increases the risk of determining false homologous PPIs. On the other hand, if we wish to include also weak paralogs and orthologs in our analysis to be able to unveil gene duplication events that happened far in the evolutionary past, the methods for homology detection become misleading and produce spurious hits—another source of false homologous PPIs.

To maximize both sensitivity and specificity despite these difficulties, we opted for a large-scale, sequence-based screening procedure in combination with a novel scoring scheme. Given a putative physical interaction between two proteins, we use both FASTA [47] and PSI-BLAST [48] to search for homologous PPIs. FASTA supplies more reliable results for closely related proteins, while PSI-BLAST is more sensitive for remote relationships [49]. To further increase the sensitivity of our method, we consider local sequence similarities with E-values up to 10. This produces many spurious hits, and thus the traditional homology-based PPI validation technique that simply checks for the existence of a single homologous PPI becomes misleading (compare Figure 3). We therefore follow a similar approach as Jonsson *et al.* [25] and use a simple scoring scheme that weights each match according to its sequence similarity: low E-values score high, and high E-values score low. High-scoring PPIs can result from a few high-quality hits but also from numerous low-quality hits. If the duplication-divergence model of PPI evolution is correct, both characteristics, i.e. the existence of few high-quality homologous PPIs and a large number of low-quality homologous PPIs, should be observable for biologically relevant PPIs.

Homologous PPIs are searched within a subset of the Protein Interaction and Molecule Search (PRIMOS) database[1], release BETA-2.7/2007-04 [50]. This subset consisted of 135,276 redundancy-removed, physical binary PPIs between 42,288 proteins from 20 organisms (Supplementary Table 1). According to Figure 1,

[1] http://primos.fh-hagenberg.at

we considered a PPI to be homologous if we found a PPI between two homologous proteins. Parameter settings used for FASTA and PSI-BLAST can be found in our Supplementary Material. Note that for performance reasons we restricted the homology searches to proteins with known PPIs, i.e. homologous proteins without any known PPI in the PRIMOS database were not reported as homologous.

Gold Standard Data Sets

To assess both the performance of previously published homology-based validation techniques and the usefulness our approach we defined four gold standard positive (GSP) and one gold standard negative (GSN) data sets with PPIs from S. *cerevisiae*. Yeast is relatively well-studied, which allows us to be rather stringent in the selection of our GSP data sets. In addition, yeast has already been used numerous times in similar studies, which eases the comparison with previous results.

Our *MIPS* GSP comprises 1,541 physical binary PPIs obtained from the Comprehensive Yeast Genome Database (CYGD) [51]. This database is considered as a high-quality resource for yeast PPIs and is frequently used as a gold standard reference set. To further improve the quality of this data set we decided to exclude PPIs reported by high-throughput experiments (see Supplementary Material). The *Multiple Evidence* GSP consists of 393 PPIs that have been reported by at least two experimental methods and in at least two publications. As an additional criterion, we considered only publications imported from a single PPI source database. For example, if a PPI was reported from a publication contained in DIP and MINT, this PPI would have been excluded from this data set. If DIP had been the only source database for this PPI it would have been included. This filter criterion reduces the risk that we regard a PPI as homologous when it is actually a duplicate. The third GSP, the *Small Scale* data set, was compiled from yeast PPIs reported by small-scale experiments and contains 902 PPIs. We identified these PPIs by considering only publications with a maximum of three reported PPIs. Again, to minimize the risk of duplicate PPIs, only publications imported from one primary PPI database have been considered (same procedure as for the *Multiple Evidence* GSP). Finally we defined a *Combined* GSP data set that contains all PPIs from the previous three GSP data sets (2,723 PPIs in total).

The GSN PPI data set, called *Random*, was generated by randomly selecting 50,000 protein pairs out of 7,058 yeast proteins[2] that were not found interacting within our database. Although a randomly selected data set will not be completely free of real PPIs such a data set has the advantage that it is free of a

[2]obtained from ftp://ftp.expasy.org/databases/uniprot/current_release/knowledgebase/complete/uniprot_sprot.fasta.gz, downloaded on March 29, 2007

selection bias, for example towards protein pairs with different molecular functions [23]. However, the amount of real PPIs within a randomly selected GSN should be low at about 0.25% [52].

Our three primary GSP data sets overlap only to a low degree: just 8 PPIs are common to all three data sets, 25 between *MIPS* and *Small Scale*, 86 between *Small Scale* and *Multiple Evidence*, and 10 between *MIPS* and *Multiple Evidence*.

Scoring Scheme

The score $\mathrm{S}(a, b)$ of a queried interaction between two proteins a and b is defined as

$$\mathrm{S}(a,b) = \sum_{o \in \mathrm{O}} \sum_{\substack{(h_a, h_b) \in \\ \mathrm{H_a}(o) \times \mathrm{H_b}(o)}} \begin{cases} \mathrm{sim}(h_a)\mathrm{sim}(h_b) & h_a \mathrm{interactswith} h_b \\ 0 & \mathrm{otherwise} \end{cases} \tag{1}$$

where O is the set of organisms with known experimental PPIs in the PRIMOS database, $\mathrm{H}_a(o)$ and $\mathrm{H}_b(o)$ denote the sets of proteins from organism o that are homologous to protein a and b, respectively. If homolog h_a and homolog h_b have an experimentally observed interaction stored in our database, we add a score proportional to their sequence similarity to an overall sum.

Note that the computation of $\mathrm{S}(a, b)$ excludes homologous PPIs where the two proteins are from different organisms. Homologous PPIs from the same organism with one protein identical to one of the source PPI partners are allowed. In this case, the E-value of the identical protein was assumed to be 0. Furthermore, if we find two identical homologous PPIs in an organism, i.e. two pairs (h_{a_1}, h_{b_1}) and (h_{a_2}, h_{b_2}) where $h_{a_1} = h_{b_2}$ and $h_{b_1} = h_{a_2}$, then we count only the homologous PPI with the lower E-value and ignore the other. The E-value of a homologous PPI (h_a, h_b) is defined as $\max(\mathrm{evalue}(h_a), \mathrm{evalue}(h_b))$.

The similarity measure $\mathrm{sim}(x)$ of a homologous protein x is defined as

$$\mathrm{sim}(x) = \begin{cases} 300 & \mathrm{evalue}(x) = 0 \\ -\log_{10}(\frac{\mathrm{evalue}(x)}{100}) & \mathrm{otherwise} \end{cases} \tag{2}$$

where $\mathrm{evalue}(x)$ is the E-value of homolog x reported by FASTA and PSI-BLAST, respectively (note that FASTA and PSI-BLAST scores are computed independently, see below). Since we allowed a maximum E-value of 10, division by 100 ensures that the negative logarithm returns a positive value over the full range of possible E-values.

The score $\mathrm{S}(a, b)$ is normalized by the maximum possible score

$$\mathrm{S}_{\mathrm{norm}}(a,b) = \frac{\mathrm{S}(a,b)}{\mathrm{S}_{\mathrm{max}}(a,b)} \tag{3}$$

where $S_{max}(a, b)$ is defined as $S(a, b)$ with all h_a assumed as interacting with all h_b. This scales the score to values ranging from 0 (minimum score) to 1 (maximum score).

We independently compute two normalized scores, one with the homologs identified by FASTA and one with the homologs identified by PSI-BLAST. The final score is then defined as the arithmetic mean of both normalized scores:

$$S_{final}(a, b) = \frac{S_{norm}^{FASTA}(a, b) + S_{norm}^{PSIBLAST}(a, b)}{2} \quad (4)$$

Authors contributions

CF devised the method as well as its biological background (duplication-divergence hypothesis), designed and conducted the data analysis, and drafted the manuscript. MK and VD provided the data for analysis, were involved in many fruitful discussions, and revised the draft manuscript. TK supported in algorithm design and coordinated the project. HH and JB piloted the underlying PRIMOS system and contributed with biomedical knowhow. KÖ initiated the project, contributed with ideas throughout development, and revised the draft manuscript.

Acknowledgements

This research was supported by the Austrian Research Promotion Agency (FFG) under the FHplus program and has been financed by the Austrian government (BMVIT and BMBWK) as well as by our co-financing partner Salzburger Landeskliniken GmbH.

The authors would like to thank all partners and colleagues that contributed to this project with their work, especially Wolfgang Straßer and Doris Siegl for the development of the underlying PRIMOS system. Any opinions, findings, and conclusions or recommendations in this paper are those of the authors and do not necessarily represent the views of the research sponsors.

References

1. Mrowka R, Patzak A, Herzel H: **Is there a bias in proteome research?** *Genome Res* 2001, **11**(12):1971–1973, [http://www.genome.org/cgi/content/full/11/12/1971].
2. Legrain P, Wojcik J, Gauthier JM: **Protein–protein interaction maps: a lead towards cellular functions.** *Trends Genet* 2001, **17**(6):346–352, [http://dx.doi.org/10.1016/S0168-9525(01)02323-X].
3. von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, Bork P: **Comparative assessment of large-scale data sets of protein-protein interactions.** *Nature* 2002, **417**(6887):399–403, [http://dx.doi.org/10.1038/nature750].
4. Deane CM, Salwinski L, Xenarios I, Eisenberg D: **Protein interactions: two methods for assessment of the reliability of high throughput observations.** *Mol Cell Proteomics* 2002, **1**(5):349–356, [http://www.mcponline.org/cgi/content/full/1/5/349].

5. Sprinzak E, Sattath S, Margalit H: **How reliable are experimental protein-protein interaction data?** *J Mol Biol* 2003, **327**(5):919–923.

6. Gilchrist MA, Salter LA, Wagner A: **A statistical framework for combining and interpreting proteomic datasets.** *Bioinformatics* 2004, **20**(5):689–700, [http://dx.doi.org/10.1093/bioinformatics/btg469].

7. Edwards AM, Kus B, Jansen R, Greenbaum D, Greenblatt J, Gerstein M: **Bridging structural biology and genomics: assessing protein interaction data with known complexes.** *Trends Genet* 2002, **18**(10):529–536. [We have used a small test set of structure-based interactions to assess the quality of several protein-interaction datasets, and have quantified significant sources of error.].

8. Patil A, Nakamura H: **Filtering high-throughput protein-protein interaction data using a combination of genomic features.** *BMC Bioinformatics* 2005, **6**:100, [http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1127019].

9. Jansen R, Greenbaum D, Gerstein M: **Relating whole-genome expression data with protein-protein interactions.** *Genome Res* 2002, **12**:37–46, [http://dx.doi.org/10.1101/gr.205602].

10. Kemmeren P, van Berkum NL, Vilo J, Bijma T, Donders R, Brazma A, Holstege FCP: **Protein interaction verification and functional annotation by integrated analysis of genome-scale data.** *Mol Cell* 2002, **9**(5):1133–1143.

11. Deng M, Sun F, Chen T: **Assessment of the reliability of protein-protein interactions and protein function prediction.** *Pac Symp Biocomput* 2003, :140–151.

12. Tirosh I, Barkai N: **Computational verification of protein-protein interactions by orthologous co-expression.** *BMC Bioinformatics* 2005, **6**:40, [http://www.pubmedcentral.nih.gov/articlerender.fcgi?tool=pubmed&pubmedid=15740634].

13. Saito R, Suzuki H, Hayashizaki Y: **Construction of reliable protein-protein interaction networks with a new interaction generality measure.** *Bioinformatics* 2003, **19**(6):756–763.

14. Goldberg DS, Roth FP: **Assessing experimentally derived interactions in a small world.** *Proc Natl Acad Sci U S A* 2003, **100**(8):4372–4376, [http://www.pnas.org/cgi/content/full/100/8/4372].

15. Bader JS, Chaudhuri A, Rothberg JM, Chant J: **Gaining confidence in high-throughput protein interaction networks.** *Nat Biotechnol* 2004, **22**:78–85, [http://www.nature.com/nbt/journal/v22/n1/full/nbt924.html].

16. Chen J, Hsu W, Lee ML, Ng SK: **Discovering reliable protein interactions from high-throughput experimental data using network topology.** *Artif Intell Med* 2005, **35**(1-2):37–47, [http://dx.doi.org/10.1016/j.artmed.2005.02.004].

17. Pei P, Zhang A: **A topological measurement for weighted protein interaction network.** *Proc IEEE Comput Syst Bioinform Conf* 2005, :268–278.

18. Tan SH, Zhang Z, Ng SK: **ADVICE: Automated Detection and Validation of Interaction by Co-Evolution.** *Nucleic Acids Res* 2004, **32**(Web Server issue):W69–W72, [http://dx.doi.org/10.1093/nar/gkh471].

19. Matthews LR, Vaglio P, Reboul J, Ge H, Davis BP, Garrels J, Vincent S, Vidal M: **Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs".** *Genome Res* 2001, **11**(12):2120–2126, [http://www.genome.org/cgi/content/full/11/12/2120].

20. Walhout AJ, Sordella R, Lu X, Hartley JL, Temple GF, Brasch MA, Thierry-Mieg N, Vidal M: **Protein interaction mapping in C. elegans using proteins involved in vulval development.** *Science* 2000, **287**(5450):116–122.

21. Remm M, Storm CE, Sonnhammer EL: **Automatic clustering of orthologs and in-paralogs from pairwise species comparisons.** *J Mol Biol* 2001, **314**(5):1041–1052, [http://dx.doi.org/10.1006/jmbi.2000.5197].

22. Suthram S, Shlomi T, Ruppin E, Sharan R, Ideker T: **A direct comparison of protein interaction confidence assignment schemes.** *BMC Bioinformatics* 2006, **7**:360, [http://www.pubmedcentral.nih.gov/articlerender.fcgi?tool=pubmed&pubmedid=16872496].

23. Saeed R, Deane C: **An assessment of the uses of homologous interactions.** *Bioinformatics* 2007, [http://dx.doi.org/10.1093/bioinformatics/btm576].

24. Brown KR, Jurisica I: **Online predicted human interaction database.** *Bioinformatics* 2005, **21**(9):2076–2082, [http://dx.doi.org/10.1093/bioinformatics/bti273].

25. Jonsson PF, Cavanna T, Zicha D, Bates PA: **Cluster analysis of networks generated through homology: automatic identification of important protein communities involved in cancer metastasis.** *BMC Bioinformatics* 2006, **7**:2, [http://dx.doi.org/10.1186/1471-2105-7-2].

26. Zhang J: **Evolution by gene duplication: an update**. *Trends in Ecology and Evolution* 2003, **18 No. 6**:292–298, [http://dx.doi.org/10.1016/S0169-5347(03)00033-8].

27. Gough J, Karplus K, Hughey R, Chothia C: **Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure.** *J Mol Biol* 2001, **313**(4):903–919, [http://dx.doi.org/10.1006/jmbi.2001.5080].

28. Todd AE, Orengo CA, Thornton JM: **Evolution of function in protein superfamilies, from a structural perspective.** *J Mol Biol* 2001, **307**(4):1113–1143, [http://dx.doi.org/10.1006/jmbi.2001.4513].

29. Murzin AG: **How far divergent evolution goes in proteins.** *Curr Opin Struct Biol* 1998, **8**(3):380–387.

30. Ohno S: *Evolution by gene duplication*. Springer-Verlag 1970.

31. Levy ED, Pereira-Leal JB: **Evolution and dynamics of protein interactions and networks.** *Curr Opin Struct Biol* 2008, [http://dx.doi.org/10.1016/j.sbi.2008.03.003].

32. Evlampiev K, Isambert H: **Modeling protein network evolution under genome duplication and domain shuffling.** *BMC Syst Biol* 2007, **1**:49, [http://dx.doi.org/10.1186/1752-0509-1-49].

33. Vázquez A, Flammini A, Maritan A, Vespignani A: **Modeling of Protein Interaction Networks**. *Complexus* 2003, **1**:38–44.

34. Pastor-Satorras R, Smith E, Solé RV: **Evolving protein interaction networks through gene duplication.** *J Theor Biol* 2003, **222**(2):199–210.

35. Light S, Kraulis P, Elofsson A: **Preferential attachment in the evolution of metabolic networks.** *BMC Genomics* 2005, **6**:159, [http://dx.doi.org/10.1186/1471-2164-6-159].

36. Pereira-Leal JB, Teichmann SA: **Novel specificities emerge by stepwise duplication of functional modules.** *Genome Res* 2005, **15**(4):552–559, [http://dx.doi.org/10.1101/gr.3102105].

37. Amoutzias GD, Robertson DL, Oliver SG, Bornberg-Bauer E: **Convergent evolution of gene networks by single-gene duplications in higher eukaryotes.** *EMBO Rep* 2004, **5**(3):274–279, [http://dx.doi.org/10.1038/sj.embor.7400096].

38. Teichmann SA, Babu MM: **Gene regulatory network growth by duplication.** *Nat Genet* 2004, **36**(5):492–496, [http://dx.doi.org/10.1038/ng1340].

39. van Noort V, Snel B, Huynen MA: **The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model.** *EMBO Rep* 2004, **5**(3):280–284, [http://dx.doi.org/10.1038/sj.embor.7400090].

40. Eisenberg E, Levanon EY: **Preferential attachment in the protein network evolution.** *Phys Rev Lett* 2003, **91**(13):138701.

41. Mika S, Rost B: **Protein-protein interactions more conserved within species than across species.** *PLoS Comput Biol* 2006, **2**(7):e79, [http://www.pubmedcentral.nih.gov/articlerender.fcgi?tool=pubmed&pubmedid=16854211].

42. Yu H, Luscombe NM, Lu HX, Zhu X, Xia Y, Han JDJ, Bertin N, Chung S, Vidal M, Gerstein M: **Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs.** *Genome Res* 2004, **14**(6):1107–1118, [http://www.genome.org/cgi/content/full/14/6/1107].

43. Jose H, Vadivukarasi T, Devakumar J: **Extraction of protein interaction data: a comparative analysis of methods in use.** *EURASIP J Bioinform Syst Biol* 2007, :53096, [http://dx.doi.org/10.1155/2007/53096].

44. Fawcett T: **ROC Graphs: Notes and Practical Considerations for Researchers**. *Machine Learning* 2004.

45. Hart GT, Ramani AK, Marcotte EM: **How complete are current yeast and human protein-interaction networks?** *Genome Biol* 2006, **7**(11):120, [http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Abstract&list_uids=17147767].

46. Jacob F: **Evolution and tinkering.** *Science* 1977, **196**(4295):1161–1166.

47. Pearson WR, Lipman DJ: **Improved tools for biological sequence comparison.** *Proc Natl Acad Sci U S A* 1988, **85**(8):2444–2448.

48. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25**(17):3389–3402.

49. Park J, Karplus K, Barrett C, Hughey R, Haussler D, Hubbard T, Chothia C: **Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods.** *J Mol Biol* 1998, **284**(4):1201–1210, [http://dx.doi.org/10.1006/jmbi.1998.2221].

50. Straßer W, Siegl D, Önder K, Bauer J: **InSilico Proteomics System: Integration and Application of Protein and Protein-Protein Interaction Data using Microsoft .NET**. *Journal of Integrative Bioinformatics* 2006, **3(2)**.

51. Güldener U, ünsterk ötter MM, Oesterheld M, Pagel P, Ruepp A, Mewes HW, ümpflen VS: **MPact: the MIPS protein interaction resource on yeast.** *Nucleic Acids Res* 2006, **34**(Database issue):D436–D441, [http://nar.oxfordjournals.org/cgi/content/full/34/suppl_1/D436].

52. Ben-Hur A, Noble WS: **Choosing negative examples for the prediction of protein-protein interactions.** *BMC Bioinformatics* 2006, **7 Suppl 1**:S2, [http://dx.doi.org/10.1186/1471-2105-7-S1-S2].

Figures

Figure 1 - Homology-Based PPI Validation

Concept of homology-based PPI validation: based on an experimentally observed physical interaction between two proteins, X and Y (the questioned 'source' PPI), homologs of both proteins are identified (generally by local sequence alignments). These homologs include both paralogs from within the same species and orthologs from other species. An interaction between a homolog of X and a homolog of Y is called a homologous PPI. A homologous PPI can be merely hypothetical, i.e. there is no evidence that it actually exists (faint lines), or it can already have been observed experimentally (thick lines). If an experimentally observed homologous PPI is found, confidence in the questioned source PPI is increased.

Figure 2 - Duplication-Divergence Model of PPI Evolution

A simplified gene tree illustrating the emergence of new PPIs under the duplication-divergence model of PPI evolution. In an ancestral species, the gene encoding a self-interacting protein, A, is duplicated. From the resulting genes A_1 and A_2, A_1 at some point loses its capability for self-interaction. Subsequent speciation forms the rat (R) and mouse (M) lineages, which evolve differently: in the mouse lineage, gene M_1 is duplicated again, in the rat lineage R_2 is duplicated. One of the R_2 duplicates loses its capability for homodimerization due to deleterious mutations. The common evolutionary origin of all resulting PPIs has an important consequence: each of these PPIs provides evidence for any other, regardless of whether the homologous PPIs are functionally conserved or are observed within the same or different species.

Figure 3 - Number of Homologous PPIs

Percentage of GSP PPIs (*Combined* data set, left bars) and GSN PPIs (*Random* data set, right bars) with a certain number of homologous PPIs (A), paralogous PPIs (B), and orthologous PPIs (C). We investigated eight distinct E-value windows (x-axis, not cumulative) and used PSI-BLAST to determine the number of homologous PPIs within each of these windows. Each bar is composed of four distinct groups: the percentage of PPIs with a single identified homologous PPI, the percentage with 2 to 10 homologous PPIs, the percentage with 11 to 100 homologous PPIs, and the percentage with more than 100 identified homologous PPIs.

Figure 4 - Overall Performance

Overall performance of our scoring scheme on the *MIPS*, the *Small Scale*, and the *Multiple Evidence* gold standard data sets. Our *Random* data set served as the gold standard negative. Each data point of the curves corresponds to a pair of true positive and false positive rates, defined as the fraction of GSP PPIs and GSN PPIs that achieved a score above a sliding threshold. The threshold ranged from 1 to 10^{-5} in this figure (values not shown).

Figure 5 - Comparison of Homology-Based Validation Schemes

Performance of our scoring scheme (PRIMOS Score) in comparison to two conventional approaches (named 'FASTA Binary' and 'PSI-BLAST Binary' here), where a PPI is deemed as biologically relevant as soon as a single homologous PPI is found below a certain E-value. GSP PPIs comprised all PPIs from our *Combined* data set, the *Random* data set was used for the GSN PPIs. For the two binary schemes, we used FASTA and PSI-BLAST, respectively, to identify homologous PPIs and calculated the TPRs and FPRs as the fraction of GSP PPIs and GSN PPIs that had at least one homologous PPI below a sliding E-value threshold, ranging from 10^{-300} to 10 in this figure. Black rectangle: parameter settings from Saeed and Deane [23]. Black circle: parameter setting used by Patil and Nakamura [8].

Figure 6 - Contribution of Weak Homologs

Contribution of weak homologous PPIs to the overall classification performance of our scoring scheme. The gray ROC curve (squares) represents the original performance of our scoring scheme (considers all homologous PPIs with an E-value up to 10). The red ROC curve (crosses) illustrates the performance of our scoring scheme when only homologs with an E-value up to 1 are examined, and the blue curve

(triangles) ignores all homologs with an E-value above 10^{-10}. GSP PPIs were taken from our *Combined* data set, GSN PPIs comprised all PPIs from the *Random* data set.

Additional Files

Additional file 1 — Supplementary Information