Protein–Protein Interactions More Conserved within Species than across Species

Sven Mika[1,2,*], Burkhard Rost[2,3,4]

1 Department of Biochemistry and Molecular Biophysics, Columbia University, New York, New York, United States of America, **2** Columbia University Center for Computational Biology and Bioinformatics, Irvine Cancer Center, New York, New York, United States of America, **3** Institute of Physical Biochemistry, University Witten/Herdecke, Witten, Germany, **4** NorthEast Structural Genomics Consortium, New York, New York, United States of America

Experimental high-throughput studies of protein–protein interactions are beginning to provide enough data for comprehensive computational studies. Today, about ten large data sets, each with thousands of interacting pairs, coarsely sample the interactions in fly, human, worm, and yeast. Another about 55,000 pairs of interacting proteins have been identified by more careful, detailed biochemical experiments. Most interactions are experimentally observed in prokaryotes and simple eukaryotes; very few interactions are observed in higher eukaryotes such as mammals. It is commonly assumed that pathways in mammals can be inferred through homology to model organisms, e.g. the experimental observation that two yeast proteins interact is transferred to infer that the two corresponding proteins in human also interact. Two pairs for which the interaction is conserved are often described as interologs. The goal of this investigation was a large-scale comprehensive analysis of such inferences, i.e. of the evolutionary conservation of interologs. Here, we introduced a novel score for measuring the overlap between protein–protein interaction data sets. This measure appeared to reflect the overall quality of the data and was the basis for our two surprising results from our large-scale analysis. Firstly, homology-based inferences of physical protein–protein interactions appeared far less successful than expected. In fact, such inferences were accurate only for extremely high levels of sequence similarity. Secondly, and most surprisingly, the identification of interacting partners through sequence similarity was significantly more reliable for protein pairs within the same organism than for pairs between species. Our analysis underlined that the discrepancies between different datasets are large, even when using the same type of experiment on the same organism. This reality considerably constrains the power of homology-based transfer of interactions. In particular, the experimental probing of interactions in distant model organisms has to be undertaken with some caution. More comprehensive images of protein–protein networks will require the combination of many high-throughput methods, including *in silico* inferences and predictions.http://www.rostlab.org/results/2006/ppi_homology/

Citation: Mika S, Rost B (2006) Protein-protein interactions more conserved within species than across species. PLoS Comput Biol 2(7): e79. DOI: 10.1371/journal.pcbi.0020079

Introduction

Experiments Peek at Complete Protein–Protein Networks

The faster large-scale sequencing projects determine the alphabet of life, the higher the pressure to determine some of the actual processes that make life what it is. The understanding of functional relations among all proteins is essential to understanding how cells work. Recent breakthroughs in experimental high-throughput techniques have begun to peek at complete protein–protein interaction networks of entire organisms (Table S1). One central method is to use yeast two-hybrid (Y2H) assays [1] that are based on a genially simple idea: first, separate two domains (activation and DNA-binding) of a transcription factor that activates a reporter gene, then merge each of the two domains to a different protein (A and B) [2,3]. If A and B interact, the two transcription domains will merge, and thereby activate the reporter gene that will be detected. The difficulty of using Y2H is in mastering the details of the experimental setup. Other high-throughput methods to detect protein–protein interactions, such as phage-display assays [4], tandem affinity purifications (TAP) [5,6], co-immunoprecipitation, and affinity chromatography [2,7–9], are also commonly used. An important advantage of using Y2H over these other high-throughput techniques is the ability to measure physical interactions between proteins as opposed to pure functional associations. Also, Y2H experiments work with physiological conditions, i.e., conditions that resemble those in eukaryotic cells [2,3,10,11]. Ito et al. [12] and Uetz et al. [13] first scanned large fractions of the yeast proteome for protein–protein interactions. Others added further interactions: Ho et al. [14] used mass spectrometry and Gavin et al. [15] used TAP. Protein networks in the fly (*Drosophelia melanogaster*) have been targeted through three different Y2H studies [11,16,17], in the worm (*Caenorhabditis elegans*) through one [18], and a large subset of about 1,500 human protein network relations were detected through TAP [19]. These data bear deeper insights into cellular processes.

Editor: Andrey Rzhetsky, Columbia University, United States of America

Received November 18, 2005; **Revised** May 18, 2006; **Published** July 21, 2006

A previous version of this article appeared as an Early Online Release on May 18, 2006 (DOI: 10.1371/journal.pcbi.0020079.eor).

DOI: 10.1371/journal.pcbi.0020079

Abbreviations: AAA, ATPases associated with various cellular activities; HVAL, measure for sequence similarity; PIDE, percentage sequence identity; PPI, physical protein–protein interaction; PSI-BLAST, position-specific iterative basic local alignment search tool; TAP, tandem affinity purification; Y2H, yeast two-hybrid

* To whom correspondence should be addressed. E-mail: mika@rostlab.org

Synopsis

The IntAct database contains about ten large-scale data sets of protein–protein interactions. Each set contains thousands of experimentally observed pair interactions. Most pairs were observed in yeast (*Saccharomyces cerevisiae*), fly (*Drosophila melanogaster*), and worm (*Caenorhabditis elegans*). These interactions are often perceived as model organisms in the sense that one can infer that two mouse proteins interact if one experimentally observes the two corresponding proteins in worm to interact. Here, the authors analyzed in detail how the sequence signals of physical protein–protein interactions are conserved. It is a common assumption that protein–protein interactions can easily be inferred through homology transfer from one model organism to another organism of interest. Here, the authors demonstrated that such homology transfers are only accurate at unexpectedly high levels of sequence identity. Even more surprisingly, homology transfers of protein–protein interactions are significantly more reliable for protein pairs from the same species than for two protein pairs from different organisms. The observation that interactions were much more conserved within than across species was valid for all levels of sequence similarity, i.e. for very similar as well as for more diverged interologs.

Today's Data Are Incomplete and Not Fully Reliable

Y2H systems are not 100% accurate; they, for instance, identify many putative interactions that cannot be confirmed by other studies. One reason for false positives (interactions incorrectly postulated) is that the two proteins A and B may activate the reporter gene directly without having to interact [3]. The Margalit group has estimated the false positive rate in high-throughput Y2H assays to be about 50% [20]; the Eisenberg group has arrived at the same estimate through measuring the reliability of interactions in the Database of Interacting Proteins [21]. Y2H experiments also do not achieve complete coverage, i.e., they miss many interactions. Conversely, false negatives (missed interactions) might result from the particular experimental setup (which may prevent the interaction between A and B) or from problems in the assembly of the two transcriptional domains (activation and DNA-binding) needed for Y2H. These problems do not prevent Y2H from evolving as one of the major experimental probes for interactions; they do, however, imply that today's data sets are neither complete nor fully accurate [20,22]. One of the strong arguments in favor of large-scale Y2H experiments is that they are more systematic and much less driven by happenstance than hypothesis-driven, detailed experiments.

Known Interactions Are Expanded through Homology-Based Inference

Evolutionary connections help explain the rapid success of molecular biology: we can study a particular protein in a simple bacterium and learn about the function of the same protein in multicellular eukaryotes. This idea enables us to use model organisms to predict protein structure [23–25], subcellular localization [26], enzymatic activity [27–29], and other aspects of protein function [30–34]. The same principle is frequently applied to the extension of interactions (Figure 1): Assume that two proteins A and B are experimentally observed to bind in organism o, and that alignment methods identify related protein pairs in organism o (A′-B′) and in organism p (A″-B″). Can we infer that the pairs A′-B′ and A″-B″ also interact with each other? The Vidal group [10] has investigated how yeast interactions detected by Ito [35] and Uetz [13] map to interactions in worm. They concluded that at BLAST E-values $<10^{-10}$, only 16%–30% of the yeast interactions are transferable [36]; similar results were reported by the Gerstein group [37]. Although homology inference is common practice, no large-scale study has ever estimated levels of accuracy and coverage for physical interactions. A particular aspect of this question relates to paralogs and orthologs. Two proteins are often considered as paralogs when they originate from the same organism and differ in function. Paralogs are assumed to have arisen from gene duplication followed by the specialization and drifting away of one of the copies, while the other copy has maintained its original function. Orthologs, on the other hand, are described as two proteins with largely identical function and a common ancestor that reside in different organisms [37–39]. Applied to homology-based inference of interactions, a common assumption is that interactions are more conserved between orthologs than between paralogs [40–42], i.e., interactions are more conserved between than within organisms. If true, model organisms would be ideal for the study of interactions.

Focus on Transient Physical Interactions (PPIs)

One important difference between Y2H and TAP is that while Y2H aims at the detection of physically interacting proteins, TAP identifies large groups of proteins that are

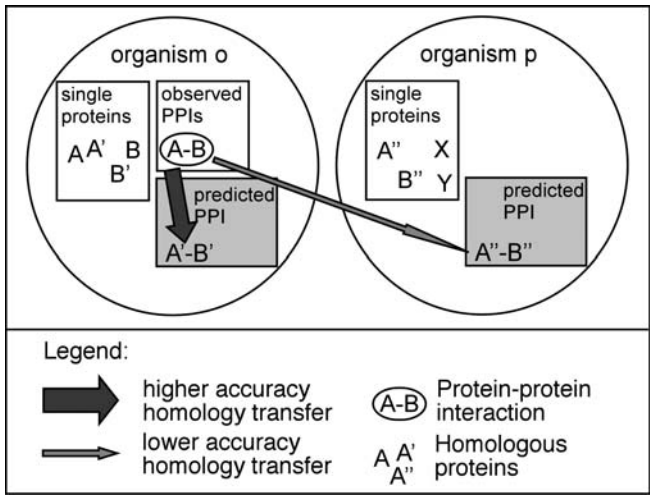

Figure 1. Concept of Homology Inference and Interologs

Interologs are two pairs of protein interactions that fulfill the following conditions: (A interacts with B) + (A is similar to A′) + (B is similar to B′) → (A′ interacts with B′). All quadruples (A, B, A′, B′) for which this relation is true are referred to as interologs [37,79]. To illustrate our analysis, we have to extend this simple relation. Assume that a physical protein–protein interaction (PPI) between proteins A and B is observed in organism o. If A and B are both sequence similar (above a certain threshold) to two other proteins A′ and B′ in the same organism o, we should be able to infer the physical interaction between A′ and B′. Note that both pairs, A/A′ as well as B/B′, have to be above the particular similarity threshold for us to be able to make this inference. Thus, we neither use an average similarity of both pairs (A/A′ and B/B′) nor a minimum similarity for just one pair (A/A′ or B/B′). Now let us assume that we have another pair of proteins A″ and B″ in another organism p, and that both are as similar to A and B as are A′ and B′, respectively. One of our findings was that homology transfers A-B → A′-B′ were more reliable than those from A-B → A″-B″.

DOI: 10.1371/journal.pcbi.0020079.g001

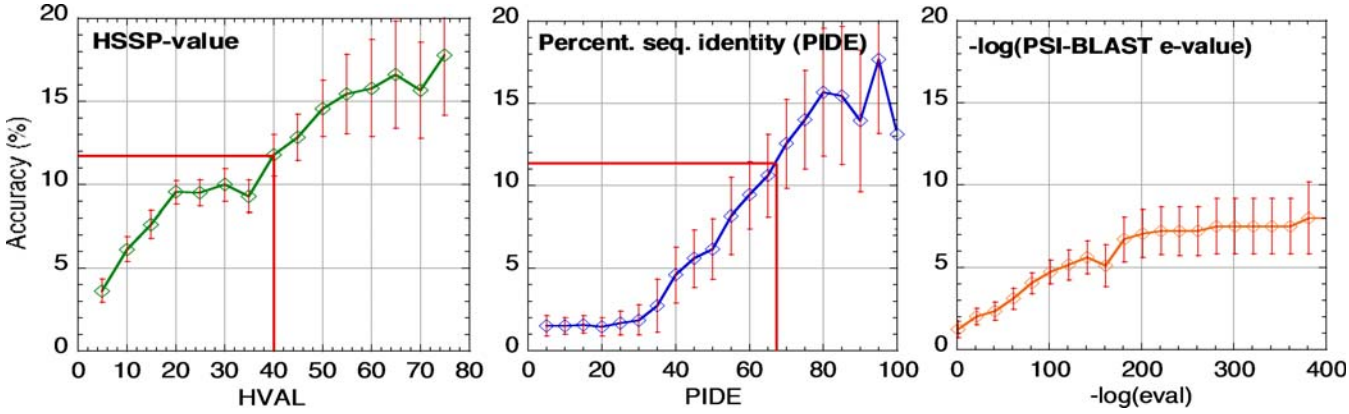

Figure 2. Sequence Conservation of PPIs

The performance of homology transfer was evaluated with the data sets in Experiment 1 (Table 4). Each panel plots the conservation (accuracy of homology transfer) using a different measure for sequence similarity: HVAL (Equation 1), PIDE (percentage pairwise sequence identity), and the PSI-BLAST E-value. It is surprising that even at high similarity thresholds (PIDE $>$ 50; HVAL $>$ 30), accuracy remained low and never reached levels of 20%. This behavior was partially explained by our overlap analysis: for low overlap (Equations 2 and 3) between datasets, we expect a low accuracy. Numbers at HVAL = 40 (which equals a PIDE of 68 at an alignment length of 100 residues) were marked with red lines. HVAL = 40 is the point, where the overlap-values (Equation 3) for two identical datasets seem to indicate a zone of $>$ 70% data consistency (see Table 3). Error bars for the three plots were calculated by bootstrapping over the PPIs in the source datasets (see Methods section).

DOI: 10.1371/journal.pcbi.0020079.g002

associated, for instance, through a common pathway [43]. Most high-throughput techniques resemble TAP in the sense that they reveal association rather than physical interaction. To illustrate this difference, assume we hypothesized that co-expressed proteins interact physically, and we wanted to use this hypothesis to predict physical interactions directly from co-expression data. Assume further that six proteins are strung together in a linear pathway (1 binds 2, 2 binds 3, etc.), and that all six are co-expressed. Of the 15 [N*(N − 1)/2] possible interactions, only 5 (N − 1) are physical, i.e., only 33% of the co-expressed proteins interact. Since most pathways involve many more than six interactions this example is likely to significantly underestimate the actual problem. In other words, even if all physically interacting proteins were co-expressed, predictions of interactions based on such association alone would still be more often wrong than right. This significantly constrains the way in which we can use association-type data to analyze physical interactions. In order to emphasize our focus on physical interactions, we used the abbreviation PPI for transient physical protein–protein interactions (as opposed to functional associations as measured by TAP-like data, and as opposed to permanent physical interactions between, e.g., two different domains or two different chains of the same protein [44]).

Coping with the Dilemma of Incomplete Data Sets

How can we evaluate accuracy and coverage of homology transfer (Figure 1) of interactions if the data are incomplete? An extreme stance is to simply not assess the performance at all. The rationale is simple: assume a method inferred that A″ and B″ in Figure 1 interacted without any experimental evidence for this interaction. May be the inference was wrong; it also may just have been a new *in silico* discovery not yet identified by experiments. If the set of all interactions were complete, the absence of an observation would imply noninteraction. Although there is currently no such complete set, we challenge that the performance of homology transfer has to be estimated somehow to render a tool that is controllable in the context of genome annotation pipelines. Here, we took the opposite radical stance by treating all interactions that have not been observed as nonexisting. While this is obviously wrong, we assume that today's incompleteness is not systematic. If true, our results will simply underestimate the quantities that we measured, but will correctly capture relative values (such as that homology transfer is half as accurate at ~40% sequence identity as at ~60%, Figure 2). We also did not merge data sets that measure functional association (e.g., TAP) with those that measure physical interaction (e.g., Y2H). Instead, we regarded only physical interactions as positives.

Here, we presented the analysis of PPI in, to our knowledge, the largest data set investigated thus far. We defined and measured the overlap between different data sets, and analyzed the expected levels of accuracy and coverage for homology-based inference of PPIs depending on the level of sequence similarity. The most surprising finding originated from differentiating between intraspecies and interspecies inferences ($o \neq p$ in Figure 1), namely that PPIs are more conserved within than between organisms.

Results/Discussion

Different Experiments Overlap Very Little

If we want to homology infer PPIs between organisms, we first have to measure the overlap within organisms and then between organisms. We introduced such a measure (Equation 2 and Equation 3, see Materials and Methods) and applied it to assessing the overlap between datasets in IntAct [45]. A large overlap value implies high agreement between two experimental sets of interactions. Our definition of overlap takes into account that two data sets may not have used the same proteins thereby rendering a score that is, in principle, independent of the size of common subsets (see Materials and Methods section). The scores are straightforward when comparing different datasets within the same organism (Equation 2) because we only have to identify identical pairs of proteins. As noted before [22,46–49], the data sets overlap

Table 1. Identity-Based Overlap (Equation 2) between Original Experimental Y2H Datasets from Fly and Yeast

Datasets	Overlap[a]		
Saccharomyces cerevisiae (yeast)	Ito [34]	Uetz [13]	
Ito [34]	**100**	27.0	
Uetz [13]	27.0	**100**	
Drosophilia melanogaster (fly)	Giot [17]	Stanyon [16]	Formstecher [11]
Giot [17]	**100**	3.3	5.4
Stanyon [16]	3.3	**100**	4.3
Formstecher [11]	5.4	4.3	**100**

[a] Overlap values are measured between two experimental data sets that have been filtered to account for the different sets of proteins used (Methods). All values compiled according to Equation 2 in percentages.
DOI: 10.1371/journal.pcbi.0020079.t001

Table 2. Homology-Based Overlap (Equation 3) between Original Experimental Y2H Datasets from Fly and Yeast

Datasets	Overlap[a]		
Saccharomyces cerevisiae (yeast)	Ito [34]	Uetz [13]	
Ito [34]	**70.2**	37.7	
Uetz [13]	37.7	**84.8**	
Drosophilia melanogaster (fly)	Giot [17]	Stanyon [16]	Formstecher [11]
Giot [17]	**53.5**	4.3	4.2
Stanyon [16]	4.3	**76.6**	7.5
Formstecher [11]	4.2	7.5	**73.2**

[a] All values compiled according to Equation 3 in percentages; the minimal sequence similarity required to consider proteins from a different organism to be similar was HVAL $>$ 20 (Equation 1) corresponding to 49% percentage sequence identity for 100 residue alignments. Overlap values for equal datasets can be smaller than 100% since homology rather than direct sequence matching is used (Equation 3). Here, we used a very weak constraint of HVAL $>$ 20 (corresponding to about 50% sequence identity for alignments over 100 residues).
DOI: 10.1371/journal.pcbi.0020079.t002

maximally for about 30% of all PPIs in yeast (*Saccharomyces Cerevisiae*) and much less for PPIs in fly (*Drosophila Melanogaster*, Table 1). Interspecies comparisons are trickier because we now have to identify the corresponding homologous pairs in the other organism. Equation 3 solves this problem by counting homologous instead of identical pairs of proteins; it is applicable to intraspecies and interspecies comparisons. A consequence of counting homologous rather than identical protein pairs is that the same data set no longer overlaps 100% with itself (Table 2), because the interaction between A and B may be detected while that between the homologs A′ and B′ may not be. The application of Equation 3 to the intraspecies comparison for yeast and fly datasets yielded similar results as the application of Equation 2 to the same datasets (Table 1). The overlap between different yeast datasets seems to be generally higher than that between different fly datasets. Finally, we merged datasets of different large-scale experiments for each organism and compared these pseudo-complete PPIs between organisms by using Equation 3 (Table 3). As expected the overlap between organisms was increased with increasing thresholds in what was considered homologous (Table 3; HSSP-value (HVAL)$>$40 highest, HVAL$>$0 lowest, Equation 1; note that the HSSP value (homology derived secondary structure of proteins) is an empirical measure for sequence similarity that empirically embeds the simple fact that high levels of sequence similarity are less meaningful for short than they are for long alignments). This increase in overlap was achieved by finding fewer matches (Table 3, empty cells). Conversely, the overlap was very low at levels of sequence similarity that mark the twilight zone of sequence-structure inference [25], i.e., the line above which most pairs of proteins have largely similar structure (HVAL$>$0, Table 3). In other words, overall fold similarity does not suffice to infer similarity in interactions.

Automatic Homology Transfer of PPIs Is Very Limited

We generated a homology performance plot (see Materials and Methods section) by comparing an unbiased, nonredundant data set (no two pairs of proteins in the set had significant sequence similarity (see Materials and Methods section) against the redundant set with all PPIs (note that we removed identical pairs even in this set, Table 4, Experiment 1). When using the observed PPI between two proteins (A-B), we applied the same sequence similarity threshold to identify both homologs (A/A′, B/B′) to infer the PPI between A′-B′. Pairs such as A-B′ or A′-B were not counted because those pairs could only be detected within the same organism and not across two species. Not surprisingly, the accuracy of homology transfer was proportional to sequence similarity (Figure 2). However, accuracy dropped rapidly already at very high levels of sequence similarity (e.g., at ~80% pairwise sequence identity, and below position-specific iterative basic local alignment search tool expectation values [PSI-BLAST E-values] $< 10^{-150}$). Closer inspection of the HSSP formula (Equation 1) reveals that the curves for HSSP values and percentage sequence identity were very similar to each other. The problem with E-values largely originated from including short alignments, i.e., many of the proteins identified at very significant E-values ($E < 10^{-50}$) might have been aligned to only small fractions of the source protein. This is a known limitation of E-values that cannot easily be normalized away because PPI interfaces may be rather short (i.e., even alignments of 20 residues in very long proteins may correctly reflect binding similarity). Although the small overlap between experimental data sets (Table 3) suggested that these estimates for accuracy at a given similarity threshold were most likely overpessimistic, the overlap scores also showed that at HVAL $>$ 40, the consistency of the data was above 70% (Table 3). Therefore, our estimates at such high thresholds might be approximately correct; if so, the accuracy of homology transfer for high similarity (HVAL $>$ 40, Percentage sequence IDEntity (PIDE) $>$ 70) were just over 10% (Figure 2). Clearly, our findings suggested that automatic homology-based inferences of PPIs have to be taken with extreme caution.

Homology Transfer Is Better within than between Organisms

Arguably [40–42], homology transfer is expected to be slightly better between organisms than within organisms. Instead, we observed the extreme opposite (Figure 3): at all levels of sequence similarity, and for all organisms with sufficient data, homology-inference was significantly more accurate for pairs of homologs from the same organism

Table 3. Homology-Based Overlap (Equation 3) between Merged Datasets for Different Similarity Thresholds

Datasets	Overlap		
	Yeast (*Saccharomyces cerevisiae*)	Fly (*Drosophilia melanogaster*)	Worm (*Caenorhabditis elegans*)
HVAL > 0			
Yeast	**11.3**	0.5	0.8
Fly	0.5	**1.5**	0.8
Worm	0.8	0.8	**7.9**
HVAL > 20			
Yeast	**65.5**	9.2	13.2
Fly	9.2	**44.9**	5.1
Worm	13.2	5.1	**69.7**
HVAL > 40			
Yeast	**82.6**	—	—
Fly	—	**75.5**	13.8
Worm	—	13.8	**88.8**

A — in the table means that the overlap cannot be calculated due to the nonexistence of any shared homologous proteins between the two sets at the given HVAL (Equation 1). Note that for proteins of ~100 residues HVAL > 40 correspond to about 73% pairwise sequence identity, HVAL > 20 to > 53%, and HVAL > 0 to > 33%.

DOI: 10.1371/journal.pcbi.0020079.t003

Table 4. Datasets Used for Homology Performance Plots

Experiment (Figure)	Datasets	
	Organism o[a]	Organism p[b]
1(2)	All	All
2(3)	All fly	All fly
3(3)	All nonfly	All fly
4(3)	All worm	All worm
5(3)	All nonworm	All worm
6(3)	All yeast	All yeast
7(3)	All nonyeast	All yeast

Organisms o and p are equal for some experiments. Datasets of o have to be nonredundant and can be either small-scale or high-throughput Y2H datasets (no TAP-like data). Datasets of organism p are redundant and have to be Y2H generated in order to guarantee a complete interaction matrix. TAP-like interactions were not used as true positives. Every single graph in Figure 3 shows the results of two experiments from Table 4 (grouped into organisms). Note that for all listed experiments, comparisons between identical datasets were omitted. For example, for experiment 6 in Table 4, this means that interactions from *yeast-Ito-2001* (organism o) will not be compared to any other interactions from this dataset in organism p (which in this case is equal to organism o).

[a] Nonredundant; No TAP-like data; PPIs

[b] Redundant; High-Throughput; TAP, tandem affinity purification; PPI, protein–protein interaction

DOI: 10.1371/journal.pcbi.0020079.t004

(intraspecies) than for pairs of homologs between different organisms (interspecies). In other words, if we experimentally observed the interaction between A and B in yeast, and if we found another pair of similar proteins A′ and B′ in yeast (not A-B′ or A′-B), as well as another pair A″ and B″ in fruit fly, then the interactions between A′ and B′ would be much more likely than those between A″ and B″. Consequently, yeast would be a rather poor model organism for the interaction network in fly.

Table 4 and Figures 2 and 3 clearly establish our main messages that intraspecies homology transfer is more accurate than interspecies transfer and that homology transfer is accurate only at unexpectedly high levels of sequence similarity. These results were stable with respect to different ways of processing the data for the experimental interactions. Changes that influenced the outcome insignificantly included the following alternatives.

Results Were Stable with Respect to Details in Filtering Data

(1) Different sampling of intraspecies vs. interspecies: We allowed transfers of the type A-B to A′-B or A-B to A-B′ (see Materials and Methods section). The performance became significantly better for intraspecies PPI transfers, thus further widening the gap between intraspecies and interspecies transfers (Figure S2A). (2) Inclusion of transfers within the same data set: we included homology transfers within the same experimental dataset (see Materials and Methods section). The effect was very similar to those observed for different sampling (see #1), i.e., the gap was widened between intraspecies and interspecies inferences (Figure S2B). (3) We used TAP-like data (Table S1) as a constraint for the negatives. To illustrate this, assume that TAP pulled down a complex of six proteins. While we cannot infer that all 15 possible interactions are physical, all could be. Therefore, we ignored a false positive prediction (i.e., we did not count it) if we could find the interaction in those 15 TAP protein–protein pairs. The accuracy slightly increased for both yeast versus yeast (intraspecies) comparisons as well as for nonyeast versus yeast (interspecies) comparisons (Figure S2C). Note that yeast is the only organism with available TAP-like data. (4) We used a redundant dataset (instead of a nonredundant, bias-reduced set) from organism o (Figure 7) to hunt for interologs in organism p (Figure 7). The main message indicated by the results for this latter experiment stays the same as in our original procedure (see Materials and Methods section): Intraspecies comparisons are more accurate than interspecies comparisons. Because there were more samples in the dataset for organism o (Figure 7) and thus higher counts, the errors slightly decreased (Figure S2D).

Examples

In the following, we presented a few representative examples that illustrate these points with more details than it is possible through averages over large data sets. Both show how homology transfer fails across species while it succeeds within an organism (Ao-Bo observed, A′o-B′o observed, A″m-B″m not observed).

Example 1: same family, different ancestors, different PPI. The two peroxins *PEX1* and *PEX6* are known to functionally and physically interact in both human [50] and yeast [51–53] (Figure 4A). A particular mutation in human *PEX1* disrupts the interaction with *PEX6*, and appears directly linked to the Zellweger Syndrome, an autosomal, recessive peroxisome biogenesis disorder, in which the growth of the myelin sheath (the fatty cover of nerve cells in the brain) is strongly affected. Patients usually suffer from visual disturbances, high iron and copper blood levels, and enlarged livers [53]. Both proteins *PEX1* and *PEX6* belong to the ATPases associated with various cellular activities (AAA) family and are involved in the

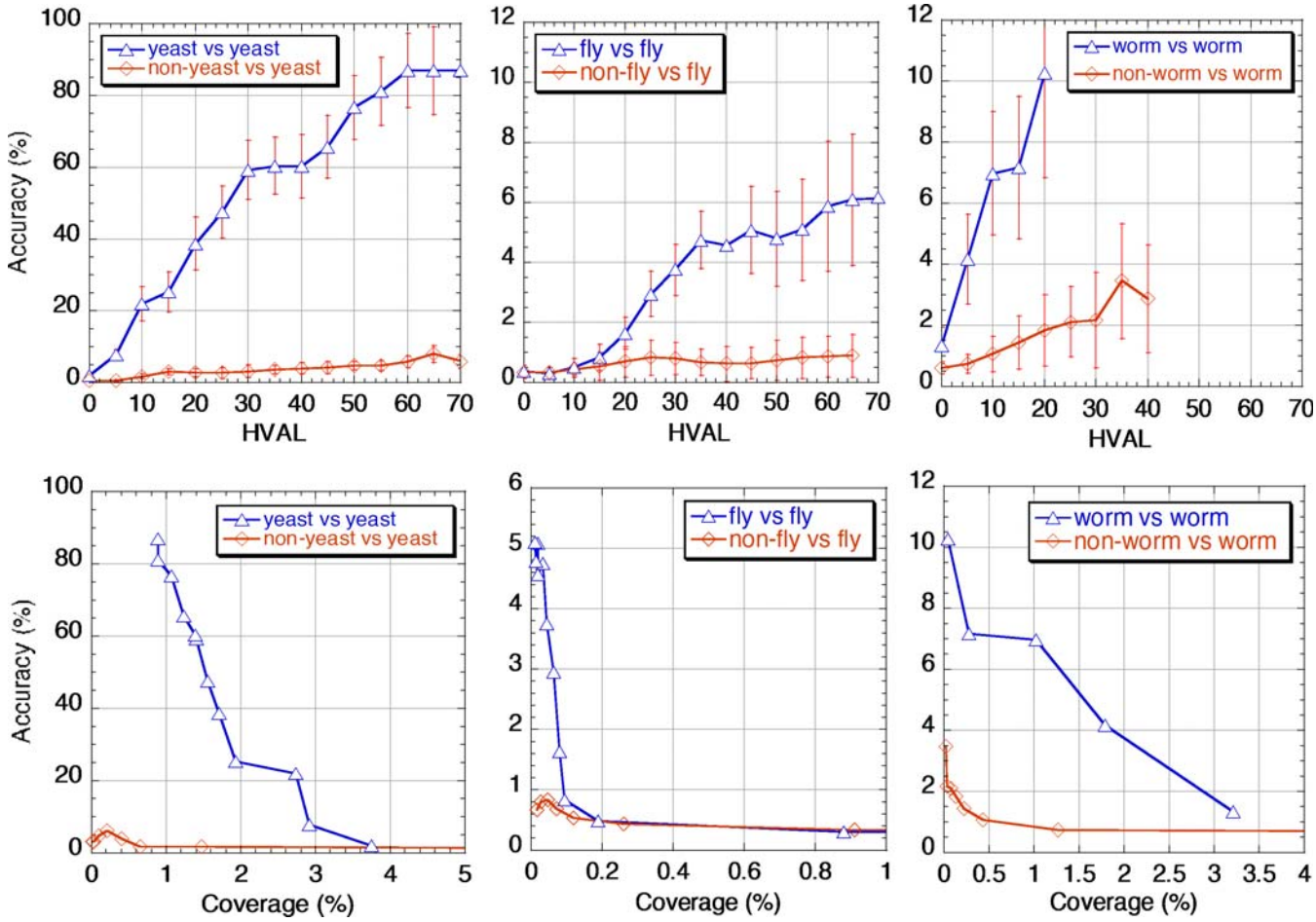

Figure 3. Performance of Homology Transfer

Plots compiled for experiments 2–7 in Table 4. Each of the upper three graphs stands for one particular organism o and shows two plots: (1) Use all known PPIs (large-scale and small-scale) of organism o to find Y2H large-scale detected PPIs in the same organism (but from different experiment, blue line). (2) Use all PPIs (large-scale and small-scale) of all other organisms (not o) to find PPIs detected by Y2H in o (red line). Only organisms with available Y2H datasets in IntAct were chosen in order to be able to create complete interaction matrices for the target datasets (yeast, worm, and fruit fly). All error bars were calculated through bootstrapping over the source PPIs (100 times, Methods). Some lines end at certain thresholds because the counts for true positives and false positives were too low (< 30 true or false positives) to calculate accuracy (Equation 4, see Materials and Methods, often also referred to as specificity or precision). Figure S1 shows the correlation between the size of the error bars and the counts of true positives at each HSSP-value cutoff. The three bottom plots show ROC-like curves, where accuracy is plotted versus coverage for the exact same data as for the three upper plots. The figures demonstrate that for all levels of similarity, the accuracy of intraspecies predictions of PPIs is significantly higher than for predictions across two organisms.

DOI: 10.1371/journal.pcbi.0020079.g003

import of proteins into the peroxisome [52,53]. Thereby, the complex of *PEX1* and *PEX6* is associated with the cytoplasmic side of the peroxisomal membrane [51]. Searching for proteins that are sequence-similar to *PEX1* and *PEX6* within yeast at an HVAL > 20 (Equation 1, see Materials and Methods) brought up two *26S protease regulatory subunits, 6A* and *6B* (proteins A′o and B′o); experts have also classified both these yeast proteins as AAA ATPases (Figure 4A). The interaction between these two yeast proteins was surprisingly found in all Y2H large scale protein–protein interaction scans [13–15,35]. Using the same threshold (HVAL > 20) the closest proteins in fly were the *26S protease subunit 4* and the *NEM-sensitive fusion protein 2* (*NSF2*) (Figure 4A). The latter—*NSF2*— is a special form of the *NEM-sensitive fusion protein 1* (*NSF1*) and is fly-specific in the sense that it does not exist in yeast, worm, or human [54–56]. An interaction between *26S protease subunit 4* and *NSF2* was not found in any of our PPI *drosophila* datasets, nor has it been reported in the literature. *NSF2* is, among other things, responsible for exocytose through vesicle fusion by disassembling the postfusion SNARE protein complexes [54,57]. Like the other *PEX1* and *PEX6* relatives discussed so far, *NSF2* is also an ATPase [54]. A detailed phylogenetic analysis of all proteins in the AAA family has suggested three major subfamilies, one with NSF homologs (*NSF1* and *2*), one with the *26S protease subunits*, and a third with *p97/Cdc48p* homologs [56]. Most importantly these three subfamilies apparently did not arise from a common ancestor but rather, they evolved independently during speciation [56].

This particular example illustrated how yeast may generally be a rather poor model organism for more complex species such as fly, worm or vertebrates. Proteins from these higher eukaryotes have to perform many different tasks in often highly specialized cell types (e.g., nerve cells). This might have lead to an evolutionary pressure to build new protein-interaction networks from the available protein building

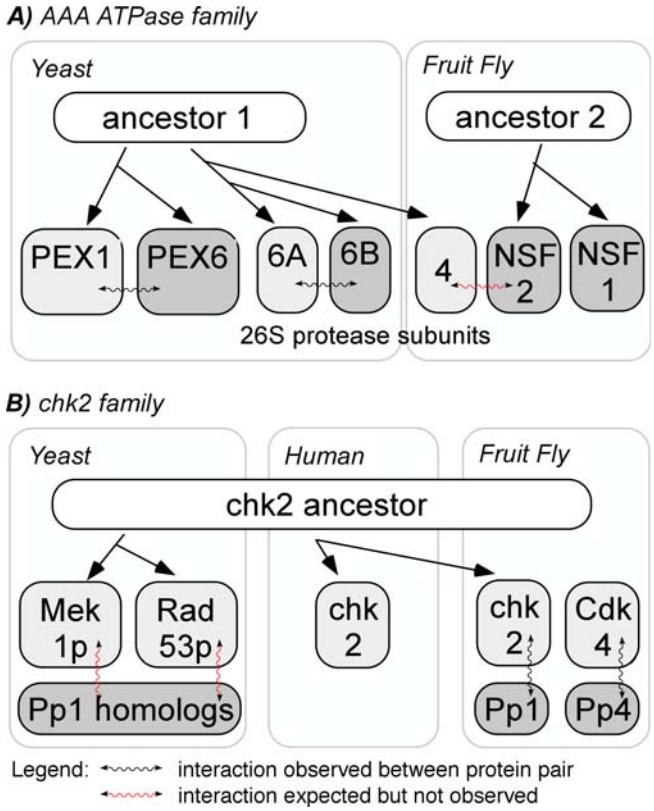

Figure 4. Interspecies Failure and Intraspecies Success of Homology Transfer

(A) Same family, different ancestors, different PPI: Two yeast peroxisomal proteins (*PEX1* and *PEX2*) are closely related through their common ancestor protein and their function as AAA ATPases to the two yeast *26S protease regulatory subunits 6A* and *6B*. In the fruit fly, gene duplication of a second ancestor protein (the *NSF* ancestor) led to two distinct *NSF* proteins (*NSF1* and *2*). Since the ancestors for the NSFs (*NSF1* and *2*) and for the *26S protease subunits* were two different proteins, we conclude that despite their common biochemical function as ATPases, the different cellular functions of NSFs and 26S protease subunits also led to a distinct behavior with respect to protein–protein interactions. Therefore, neither *NSF1* nor *NSF2* were observed to bind to the *26S protease subunit 4*.

(B) Same pathway, different functions, different binding: Evolutionary plasticity in the *chk2* family led to a diverse range of functions of these proteins while staying in the same pathway. For example *Rad53p* in yeast is a main player in the cell cycle checkpoint during mitosis, whereas *Mek1p* acts in the same position during meiosis. Also, *drosophila chk2* and human *chk2* act at different times during the cell cycle different from *Mek1p* and *Rad53p*. No *drosophila Pp1* homolog in yeast was found to interact with either *Mek1p* or *Rad53p*, even though *drosophila Pp1* was shown to bind to *drosophila* chk2.

DOI: 10.1371/journal.pcbi.0020079.g004

blocks (e.g., ATPase function). Thus, by only slightly altering the existing sequences, new binding properties were added to these proteins, while others were lost. A similar argument could be used to explain a likely poor homology transfer between fly and human or worm and human.

Example 2: same pathway, different functions, different binding properties. The *drosophila Ser/Thr protein phosphatase 4* (*Pp4*) and the *cyclin dependent kinase 4* (*Cdk4*) were found in our small-scale dataset for *drosophila* PPIs. At HVAL>20, we found two sequence-similar proteins in fly, namely *Ser/Thr protein phosphatase alpha 2* (*Pp1*) similar to *Pp4*, and *chk2* similar to *Cdk4*; both these fly proteins (*Pp1* and *chk2*) have been shown to interact [16]. Fly *chk2* as well as its sequence relatives in

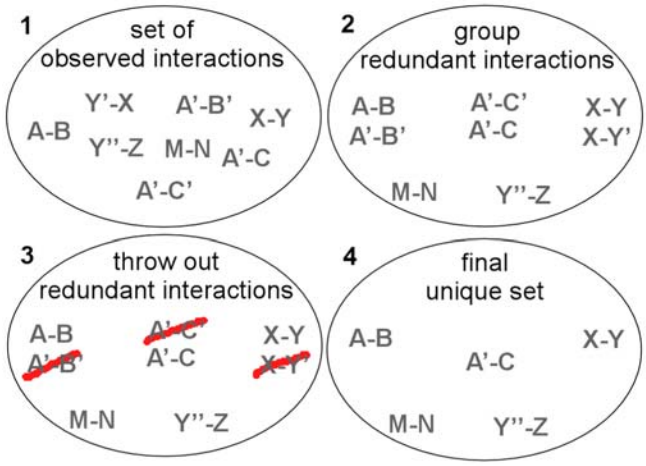

Figure 5. Creating Sequence-Unique PPI sets

(1) Starting with a dataset of PPIs, we first cluster the data according to sequence similarity (apply a certain homology threshold) into sequence similar PPIs (2). Note here that the interactions A′-B′ and A′-C′ do not fall into the same cluster because B′ and C′ are unrelated. Thus, for two interactions (e.g., A-B and A′-B′) to be considered similar by our algorithm, both interacting proteins (A and B) have to be homologous to the two proteins of the other interaction (A has to be similar to A′ and B has to be similar to B′). (3) We randomly throw out all redundant interactions in each cluster so that only one PPI remains as a representative of each cluster. (4) Those representatives constitute the final unique dataset of PPIs.

DOI: 10.1371/journal.pcbi.0020079.g005

yeast (*Mek1p* and *Rad53p*) and human are involved in cell-cycle checkpoints, which are signal transduction pathways that control the cell cycle and prevent the cell from further replication if the DNA double strand breaks, the DNA is incompletely replicated, or in case of other DNA damages [58–60]. A checkpoint can halt an ongoing mitosis or meiosis or even terminate it and induce apoptosis. A phylogenetic analysis of the *chk2* family members found that fly *chk2* and its yeast and human homologs stem from the same ancestor (Figure 4B). Nevertheless, it is also known that this family of proteins has a rather strong evolutionary plasticity in terms of the particular tasks of its members [60,61]. For example in yeast, *Mek1p* only controls the meiotic pachytene checkpoint by making sure that only homologous chromosomes recombine with each other [61], whereas yeast *Rad53p* controls mitotic cell replication and does not seem to be required for meiotic checkpoint control at all [60]. Also, the timing within the cell cycle is different for yeast *Rad53p* and its *drosophila* ortholog *chk2* [60]. This plasticity in the chk2 family might explain why many yeast proteins homologous to *drosophila Pp1* were not found to interact with either *Rad53p* or *Mek1p*.

Sequence-Based Homology Transfer Is Limited Although Binding Sites Are Partially Conserved in Three-Dimensional (3-D) Structure

Recently, the Sali group analyzed the conservation of protein–protein binding sites on homologous and structurally aligned protein surfaces. They found that the differences in the localization of binding sites between homologous proteins are significantly smaller than the differences expected at random [62]. On the one hand, this result is similar to what we found for higher levels of similarity (Figure 3). On the other hand of very little similarity the difference

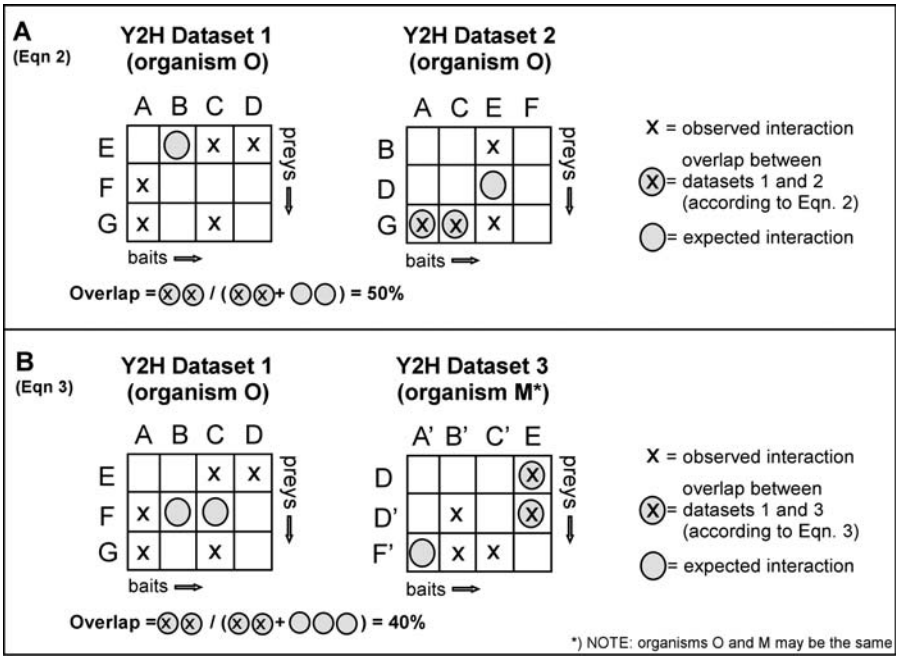

Figure 6. Ways of Calculating the Overlap between Two Y2H Datasets

(A) Identity-based overlap between Datasets 1 and 2 according to Equation 2. Note that we can only calculate this score if both datasets are from the same organism. Starting with the observed interaction C-E in Dataset 1, we are trying to find the exact same interaction in Dataset 2. The following situations might occur: (a) C and E are also observed to interact in Dataset 2. (b) C and E are not observed to interact in Dataset 2. (c) It is impossible for C and E to be interacting in Dataset 2 due to either of these two reasons: (i) Either C or E are not part of Dataset 2 or (ii) C and E are either both used as preys or both used as baits in Dataset 2. Repeating the above procedure for all other observed interactions in Datasets 1 and 2, we finally calculate the identity-based overlap by dividing the number of common interactions found in Datasets 1 and 2 by the total number of expected interactions (observed and not-observed).

(B) The same procedure as described above is applied to the two Datasets 1 and 3, which are now allowed to be from different organisms. The only difference to Equation 2 (A) is the usage of homology for comparing two PPIs instead of a binary decision scheme (PPIs identical or not-identical). Thus, starting with the interaction D-E from Dataset 1, we try to find possible homologous interactions (not only the identical PPI) in Dataset 3. The only two options in this example are D-E and D′-E (Dataset 3), which in our example are both observed in Dataset 3. Iterating through all observed interactions of Datasets 1 and 3 and summing up the expected interactions and the overlapping homologous interactions, we can then calculate the homology-based overlap (Equation 3). Note that any results from Equation 2 are not comparable to any results from Equation 3.

DOI: 10.1371/journal.pcbi.0020079.g006

between the 3-D–based results and ours lie most likely in the additional constraints implicitly used by the Sali group, namely that we know the 3-D structures and that we can focus in our alignment on all residues in the binding site. Using only sequence information, we cannot do this because binding residues close in 3-D may be separated considerably in sequence, thereby diluting the pattern of conservation picked up by alignment methods. However, for most PPIs from IntAct, we can neither label the binding site, nor do we have 3-D structural information. Therefore, we are limited to having to measure overall sequence similarity. If we were able to predict binding sites [63–66], we might improve homology transfer considerably.

Conclusions

As demonstrated again by our overlap measure, today's datasets of PPIs are still rather inconsistent (Tables 1–3). The discrepancies were significantly smaller between yeast than between fly datasets (Tables 1 and 2). This finding also explains the much higher accuracy for intrayeast as opposed to intrafly or intraworm transfer. Why datasets of yeast appear more consistent than those of fly datasets remains speculation. One reason might be that measurements of protein–protein interactions are performed within yeast (Y2H) and are thus more precise for yeast proteins than for other species' proteins, since those might behave differently in the unfamiliar yeast cell. Although incomplete and not fully consistent, PPI datasets are finally large enough to validate quantitative analyses. In particular, this enables a large-scale assessment of the performance of automated homology transfer for PPIs. Assuming that today's errors are largely nonsystematic, estimates for the performance of homology transfer will provide correct qualitative pictures, albeit the actual numbers will be overpessimistic. In the extreme regimen of comparing very similar pairs of proteins, we could establish that data sets appeared very consistent (Figure 2). Consequently, our estimates for the performance of homology transfer were likely to be relatively reliable in this regimen. Nevertheless, even for very high similarity, automated homology transfer was often mistaken; it approached random when approaching the sequence-structure twilight zone, i.e. the region in which sequence similarity no longer implies 3-D similarity (Figure 3). Although many interactions observed in one organism were not observed in another, similar interactions in the same organism (at similar levels of sequence similarity) were often observed (Figure 3). Consequently, our results challenge that using homology to transfer a protein–protein interaction from one organism to another is more difficult and less accurate than a transfer

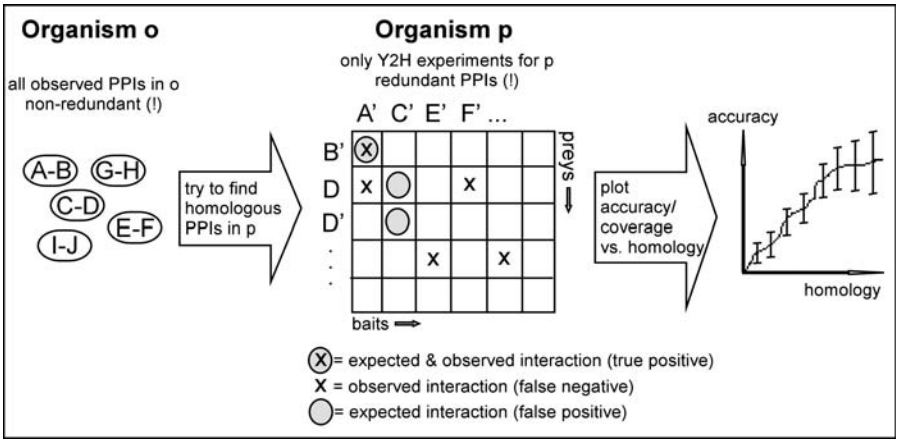

Figure 7. Evaluating Homology Inference of PPIs

Starting with the entirety of observed interactions in any organism o (Y2H plus small scale experiments), we first reduce the sequence redundancy from this dataset as described in Figure 3. Then we try to find homologs in the organism p for each of the unique PPIs of organism o. Since we want to be able to conclude that every nondetected interaction in organism p does actually not exist in real life, we need to have a complete interaction matrix (baits × preys) for organism p. Thus, we are forced to exclude all small-scale data from the organism p dataset and remain with a merger of all (redundant) Y2H interactions for this organism. For each interaction A-B from organism o, we can face any of the following situations: (a) A homologous interaction A′-B′ can be found in organism p, (b) no homologous interaction can be found in p, or (c) It is impossible to detect an interaction of type A′-B′ in p because of one of the following two reasons: (i) either A′ or B′ are missing in the dataset for p or (ii) Both A′ and B′ are either preys or both are baits in the dataset for organism p. The latter case (c.ii) is illustrated by the interaction E-F in organism o, which cannot be detected in organism p only because E′ and F′ are both used as preys in the experiment. No counts for false positives are made for those cases. Adding the numbers of true positives (expected and observed PPIs), false positives (expected but not observed) and false negatives (observed interaction only in organism p) allows us to calculate accuracy and coverage for each homology threshold used to infer interactions (Equation 4). It is important to note that in the case where o = p, comparisons between two identical experimental PPI-sets are ignored (e.g. A-B in o's set "*yeast-Ito-2001*" is not used to predict A′-B′ in p's set "*yeast-Ito-2001*"; o = p = yeast).

DOI: 10.1371/journal.pcbi.0020079.g007

within the same species. This implies that distant model organisms have a limited value to unravel protein networks. We showed that these results are stable even when making major changes to the ways in which we analyzed the experimental data. Whether we used high- or low-confidence data, whether we allowed for same-set PPI transfers or not, whether we reduced bias or not, whether or not we filtered the negatives by TAP-like data about putative physical interactions, whether or not we restricted our analysis to limited inferences per family, we always observed the same: PPIs are more conserved within than across species. This discrepancy between intraspecies and interspecies conservation of interologs was valid for all levels of sequence similarity. Finally, we tested the ability of homology transfers to predict another functional annotation and then compared the performances of interspecies versus intraspecies comparisons thereof. We chose subcellular localization as an easily extractable and available protein feature. By using a list of proteins annotated for subcellular localizations from UniProt [67], we could show that there is no significant difference in performances for interspecies versus intraspecies homology transfers for this particular feature.

Materials and Methods

Data sets. Several publicly available databases such as GRID [68], BIND [69], MINT [70], and DIP [71,72] gather information about interacting proteins in different organisms. For our analysis, we used the IntAct database [45], a protein–protein interaction resource maintained at the European Bioinformaics Institute (EBI) in Cambridge (http://ebi.ac.uk/intact/). IntAct uses the PSI format (extended markup language (XML)-tagged) to store data [73], fly [12–15], fly [11,16,17], worm [18] and human [19] as well as about 30 so called small-scale datasets, which are collections of results from many detailed experiments for different organisms. The largest small-scale dataset is that of human with about 38,000 interactions. Concerning the high-throughput datasets, IntAct carries detailed information about which proteins were used as baits and which proteins were used as preys, so that a complete interaction matrix can easily be reconstructed from these sets. Table S1 contains all protein–protein interaction datasets deposited in IntAct at the moment along with links to these datasets (small-scale and large-scale). The Giot [17], Ito [35], and Li [18] datasets contain some information about the level of confidence that was assigned to each interaction. For these three sets, we excluded everything from our analysis that either had a confidence-value of less than 0.4 (Giot: values range from 0 to 1) or those that were not in a so called "core" dataset of trusted interactions (Ito and Li divide their sets into core and full or core and noncore subsets, where core means a higher confidence in the measured interaction). Note that for the initial submission of this manuscript we had compiled all results for unfiltered data sets, i.e., we had included all experimental interactions; the results were qualitatively identical to those given here (data not shown).

True positives and false negatives: focus on Physical Interactions = PPIs. Technically, we realized our goal of exclusively focusing on PPIs through the particular way of labeling positives and negatives. We labeled as positives (true PPIs) only those pairs that were identified by experiments that target the detection of physical interactions (only Y2H experiments).

We then also assumed that these data for each organism was complete, i.e., we labeled all pairs as negatives that were not detected by Y2H.

Measuring sequence similarity/homology. The term homology usually implies an evolutionary relation in the sense of having a common ancestor. Strictly speaking, we cannot measure homology. Instead, alignment methods measure sequence similarity in some way or other. In our work the ranges of similarity were so high that the pairs of proteins were most likely homologous. We used BLAST and PSI-BLAST [74] to align all protein sequences in IntAct against each other (standard procedure [75]: 3 iterations at $E<10^{-10}$ against filtered database of all proteins to build clean profiles, then one run with frozen profile against unfiltered database at $E < 10^{-3}$, freeze profile again and run against all IntAct proteins). Then we extracted the PSI-BLAST E-values for each alignment, as well as the percentage of sequence identity (PIDE) and the distance to the HSSP curve, i.e.

the HSSP-value [25,76,77] (HVAL). The HVAL is defined as:

$$HVAL_{(PIDE,L)} = PIDE - \begin{cases} 100 & \text{for } L \leq 11 \\ 480 \cdot L^{-0.32 \cdot \{1+\exp(-L/1000)\}} & \text{for } L \leq 450 \\ 19.5 & \text{for } L > 450 \end{cases} \quad (1)$$

where L was the number of residues aligned between two proteins, and PIDE the percentage of pairwise identical residues. HSSP values consider both pairwise sequence identity and alignment length: the higher the value the more similar two proteins. Values around 0 typically imply that two proteins have similar 3-D structures and correspond to about 22% pairwise sequence identity at alignment lengths above 250 residues.

Nonredundant data sets. We removed bias from PPI datasets by the following procedure (Figure 5). (1) Move down a list L of PPIs starting with pair A-B. (2) Group all interactions in this list into clusters of similar PPIs. Consider two distinct PPIs as similar only if both partners of the first interaction are homologs to the respective protein in the second interaction. For instance, let A′ be a homolog of A, and B′ be a homolog of B. Then all interactions A′-B, A′-B′, and A-B′ will fall into the same group as the interaction A-B. Note that this also means that any interaction A-C will not end up in this group if C is not a homolog of B. Here, we used a very conservative criterion for homolog, namely HVAL $>$ 0 (Equation 1). This threshold is conservative in the sense that it will also remove nonredundant pairs, i.e., many proteins that are actually not homologs. (3) Reduce each group formed in step 2 to one single representative PPI. (4) Continue working with the final unique (nonredundant) dataset.

Identity- and homology-based overlap between datasets. We defined two procedures resembling the Jaccard correlation to measure the overlap between two different datasets of PPIs in IntAct. Equation 2 defines the first measure; for clarity we refer to this measure as the identity-based overlap. This measure can only be applied to two PPI sets from the same organism.

$$overlap(M,N) = \frac{PPI(MandN)}{PPI(MandN) + PPI(MxorN)} \quad (2)$$

where *PPI*(*MandN*) is the number of PPIs that were detected in both sets (common PPIs) and *PPI*(*MxorN*) is the number of PPIs that were only detected in one of the two datasets (exclusive or). Figure 6A describes this procedure. Note that only those interactions contributed to the count of *PPI*(*MxorN*) that could possibly have been detected in both datasets. For example, if the PPI A-B is detected in dataset 1, but not in dataset 2, we only increase *PPI*(*MxorN*) by one, if A and B were both included in dataset 2. In other words, we completely ignored interactions A-B in one dataset, if either A, or B (or both) were not present in the other dataset. Given this definition (Equation 2), an overlap value of 0.5 means that every second PPI of dataset 1 is not present in dataset 2. Inversely, every second PPI from dataset 2 cannot be found in dataset 1. Furthermore, applying Equation 2 to calculate the overlap of one dataset with itself always results in 1 (100% overlap).

The second measure capturing an overlap between two interaction datasets was applicable to any two datasets, even if they were from different organisms. We referred to this measure as the homology-based overlap. It was defined as follows (Figure 6B):

$$overlap(M,N,h) = \frac{PPI(MandN)^{(h)}}{PPI(MandN)^{(h)} + PPI(MxorN)^{(h)}} \quad (3)$$

where $PPI(MandN)^{(h)}$ is the number of homologous PPIs reported in both datasets considering a homology threshold of HVAL $>$ h. Assume again that A is homolog of A′ and B of B′. If the interaction A-B is in dataset 1 and the interaction A′-B′ is in dataset 2, the count for $PPI(MandN)^{(h)}$ will increase by one. The quantities $PPI(MandN)^{(h)}$ and $PPI(MxorN)^{(h)}$ are similar to those in Equation 2 with the simple caveat that we substituted identical pairs with homologous pairs, because there are no identical pairs between two different organisms. Unlike for Equation 2, when using Equation 3 to measure the overlap between a dataset and itself, the result usually happens to be $<$ 1 ($<$ 100%). For an explanation consider the following example. Assume that our dataset contains the interaction A(bait)-B(prey) along with another protein A′ (bait, homologous to A) that is not found to interact with B. The absence of A′-B will increase the count of $PPI(MxorN)^{(h)}$ by one, thereby yielding a self overlap $<$1. On the one hand, for very high levels of similarity (say A and A′ have 99% pairwise sequence identity), the reduction from 1 can be interpreted as a reflection of the limitation of experimental accuracy. On the other hand, for low levels of similarity, the reduction is related to the fact that PPIs are simply not conserved between distant relatives. Note that we also investigated overlap when replacing HVAL (Equation 1) by PSI-BLAST E-values as a measure for sequence similarity. While the resulting numbers differed slightly, the trends that we reported remained the same (data not shown).

Homology performance curves. For given levels of sequence similarity, we monitored and plotted the accuracy of inferring PPIs through homology from one dataset to another. The procedure is described in Figure 7.

The resulting curves can be interpreted as the degree to which PPIs are evolutionarily conserved. In a more technical sense, the curves reflect the performance of homology transfer of PPIs (Figure 1). The HVAL (Equation 1) determined the minimal similarity between A and A′, as well as between B and B′. Other ways of considering two pairs of interacting proteins as related, for instance the arithmetic or geometric average of both HVALs (A/A′ and B/B′), led to a slightly worse performance of our homology inferences, i.e. the curves were similar albeit lower overall (data not shown). Note that each large-scale Y2H data set (Table S1) should, by experimental design, contain a complete interaction matrix (preys $\times$ baits) that is, ideally, both fully correct and comprehensive for all the proteins tested in that experiment. Consider an interaction A-B from any dataset (small-scale or large-scale) of an organism o; if we find the homologs A′ and B′ in a large-scale dataset of another organism p, we can transfer the interaction property from A-B to A′-B′. In other words, by looking at the PPI between A and B (A-B), we simply predict that A′ and B′ also interact. Because of the complete interaction matrix that we are looking at for organism p, we can now also say whether this prediction was actually right or wrong. In particular, the prediction is correct, if we find the interaction A′-B′ in p and wrong if we do not find it in p plus A′ and B′ are on different axes of the interaction matrix (A′ = prey, B′ = bait or vice versa). In order to compare the performance of homology transfers across two organisms (o $\neq$ p) to the one for intraorganism transfers (o = p), we have to allow p and o to be the same. Therefore, in order to be able to compare results from both types of experiments (intraspecies versus interspecies), we have to apply the following restrictions to comparisons within the same species (o = p): Transfers from an interaction A-B to another PPI of the type A-B′ or A′-B (one protein identical, the other homologous) are not allowed since these cases are only observable in intraspecies predictions but not in interspecies transfers. Additionally for intraspecies predictions, we required that A-B and the predicted interaction (A′-B′) stem from different datasets (different Y2H experiments) in order to ignore possible homology-based assumptions about two PPIs within the same dataset. The problem here is that in case a research group found an interaction (e.g., A-B) through a Y2H scan, would they work harder to also find an interaction A′-B′ (A′ = homolog to A, B′ = homolog to B) or A′-B rather than an unrelated interaction (e.g., M-N).

Accuracy and coverage. We measured the accuracy (Acc) and coverage (Cov) for the inference (prediction) of interacting protein pairs by the standard formulas:

$$\text{Acc} = \frac{\text{TP}}{\text{TP} + \text{FP}}; \qquad \text{Cov} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (4)$$

where TP are the true positives (i.e., physical interactions that are experimentally observed [e.g., by Y2H, note TAP-like relations are not included here] and that are also correctly inferred by homology). FP are the false positives (i.e., the pairs inferred through homology but not observed by Y2H experiments). Finally, FN are the false negatives (i.e., the physical interactions that have been observed but were not identified). We monitored levels of accuracy and coverage as a function of the sequence similarity between the proteins of known and those of unknown annotations. There is a trade-off between these two: the more restrictive the sequence similarity threshold, the more interactions will be inferred (higher coverage) at the expense of reduced accuracy; and the higher the threshold, the more will be right (high accuracy) at the expense of few inferences (low coverage).

Error estimate. The error in the estimates of accuracy and coverage were determined by bootstrapping [78] over the protein–protein interactions in the source datasets. In particular, we picked n interactions at random from the non-redundant source dataset and compiled the averages over a larger set with possibly many replicas of the same incidence. The levels of accuracy/coverage for different thresholds in sequence similarity were then calculated according to the procedure described above (Figure 7). For the bootstrapping, these two steps had been repeated 100 times before the standard deviation (sigma) for all levels of accuracy were calculated.

Supporting Information

Table S1. Large-Scale Protein–Protein Interaction Datasets from IntAct

Found at DOI: 10.1371/journal.pcbi.0020079.st001 (74 KB DOC).

Figure S1. Number of true positive counts versus HVAL

Each curve shows the accuracy (red) as shown in Figure 3 and the number of true positives counted at a certain HSSP-value cutoff (green)

Found at DOI: 10.1371/journal.pcbi.0020079.sg001 (72 KB PDF).

Figure S2. Results Are Stable with Respect to Variations in the Experimental Setup

(A) Different sampling of intra- versus inter-species: we allowed transfers of the type A-B to A'-B or A-B to A-B' (see Materials and Methods section). The performance became significantly better for intra-species PPI-transfers, thus further widening the gap between intra- and inter-species transfers.
(B) Inclusion of transfers within the same data set: we included homology transfers within the same experimental dataset (see Materials and Methods section). The effect was very similar to those observed for different sampling (#1), i.e. widening the gap between intra- and inter-species inferences.
(C) Using TAP-like data (Table S1) as a constraint for the negatives. To illustrate this, assume that TAP pulled down a complex of six proteins. While we cannot infer that all 15 possible interactions are physical, all could be. Therefore, we ignored a false positive prediction (did not count it) if we could find the interaction in those 15 TAP protein-protein pairs. The accuracy slightly increased for both yeast versus yeast (intra-species) comparisons as well as for non-yeast versus yeast (inter-species) comparisons. Note that yeast is the only organism with available TAP-like data.
(D) We used a redundant dataset (instead of a non-redundant, bias-reduced set) from organism o (Figure 7) to hunt for interologs in organism p (Figure 7). The main message indicated by the results for this latter experiment (#4) stays the same as in our original procedure (see Materials and Methods section): intra species comparisons are more accurate than inter-species comparisons. Due to more samples in the dataset for organism o (Figure 7) and thus higher counts, the errors slightly decreased.

Found at DOI: 10.1371/journal.pcbi.0020079.sg002 (153 KB PDF).

Acknowledgments

Thanks to Jinfeng Liu, Hans-Erik Aronson, Kristen McFadden, and Paul Glick (all from Columbia University) for computer assistance. Thanks to the anonymous reviewers for their helpful criticism. Furthermore, thanks in particular to Amos Bairoch (Swiss Institute of Bioinformatics, Geneva, Switzerland), Rolf Apweiler (European Bioinformatics Institute, Hinxton, United Kingdom), Phil Bourne (San Diego University, San Diego, California, United States), David Eisenberg (University of California—Los Angeles, Los Angeles, California, United States), and their crews for maintaining excellent databases and to all experimentalists who enabled this work by publishing their PPI results in PubMed/MedLine.

Author contributions. SM conceived and designed the experiments. SM performed the experiments. SM analyzed the data. BR contributed reagents/materials/analysis tools. SM and BR wrote the paper.

Funding. This work was supported by the grants R01-GM63029-01 from the National Institute of Health.

Competing interests. The authors have declared that no competing interests exist.

References

1. Fields S, Song O (1989) A novel genetic system to detect protein–protein interactions. Nature 340: 245–246.
2. Causier B (2004) Studying the interactome with the yeast two-hybrid system and mass spectrometry. Mass Spectrom Rev 23: 350–367.
3. Legrain P, Wojcik J, Gauthier JM (2001) Protein–protein interaction maps: A lead towards cellular functions. Trends Genet 17: 346–352.
4. Willats WG (2002) Phage display: Practicalities and prospects. Plant Mol Biol 50: 837–854.
5. Puig O, Caspary F, Rigaut G, Rutz B, Bouveret E, et al. (2001) The tandem affinity purification (TAP) method: A general procedure of protein complex purification. Methods 24: 218–229.
6. Rigaut G, Shevchenko A, Rutz B, Wilm M, Mann M, et al. (1999) A generic protein purification method for protein complex characterization and proteome exploration. Nat Biotechnol 17: 1030–1032.
7. Aebersold R, Mann M (2003) Mass spectrometry-based proteomics. Nature 422: 198–207.
8. Bauer A, Kuster B (2003) Affinity purification-mass spectrometry. Powerful tools for the characterization of protein complexes. Eur J Biochem 270: 570–578.
9. Lin D, Tabb DL, Yates JR III (2003) Large-scale protein identification using mass spectrometry. Biochim Biophys Acta 1646: 1–10.
10. Walhout AJ, Vidal M (2001) Protein interaction maps for model organisms. Nat Rev Mol Cell Biol 2: 55–62.
11. Formstecher E, Aresta S, Collura V, Hamburger A, Meil A, et al. (2005) Protein interaction mapping: A *Drosophila* case study. Genome Res 15: 376–384.
12. Ito T, Tashiro K, Muta S, Ozawa R, Chiba T, et al. (2000) Toward a protein–protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. Proc Natl Acad Sci U S A 97: 1143–1147.
13. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, et al. (2000) A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. Nature 403: 623–627.
14. Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, et al. (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. Nature 415: 180–183.
15. Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, et al. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 415: 141–147.
16. Stanyon CA, Liu G, Mangiola BA, Patel N, Giot L, et al. (2004) A *Drosophila* protein-interaction map centered on cell-cycle regulators. Genome Biol 5: R96.
17. Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, et al. (2003) A protein interaction map of *Drosophila melanogaster*. Science 302: 1727–1736.
18. Li S, Armstrong CM, Bertin N, Ge H, Milstein S, et al. (2004) A map of the interactome network of the metazoan *C. elegans*. Science 303: 540–543.
19. Bouwmeester T, Bauch A, Ruffner H, Angrand PO, Bergamini G, et al. (2004) A physical and functional map of the human TNF-α/NF-κB signal transduction pathway. Nat Cell Biol 6: 97–105.
20. Sprinzak E, Sattath S, Margalit H (2003) How reliable are experimental protein–protein interaction data? J Mol Biol 327: 919–923.
21. Deane CM, Salwinski L, Xenarios I, Eisenberg D (2002) Protein interactions: Two methods for assessment of the reliability of high throughput observations. Mol Cell Proteomics 1: 349–356.
22. Sprinzak E, Sattath S, Margalit H (2003) How reliable are experimental protein–protein interaction data? J Mol Biol 327: 919–923.
23. Abagyan RA, Batalov S (1997) Do aligned sequences share the same fold? J Mol Biol 273: 355–368.
24. Brenner SE, Chothia C, Hubbard TJP (1998) Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. Proc Natl Acad Sci U S A 95: 6073–6078.
25. Rost B (1999) Twilight zone of protein sequence alignments. Protein Eng 12: 85–94.
26. Nair R, Rost B (2002) Sequence conserved for sub-cellular localization. Protein Sci 11: 2836–2847.
27. Todd AE, Orengo CA, Thornton JM (2001) Evolution of function in protein superfamilies, from a structural perspective. J Mol Biol 307: 1113–1143.
28. Devos D, Valencia A (2001) Intrinsic errors in genome annotation. Trends Genet 17: 429–431.
29. Rost B (2002) Enzyme function less conserved than anticipated. J Mol Biol 318: 595–608.
30. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO (1999) Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. Proc Natl Acad Sci U S A 96: 4285–4288.
31. Pawlowski K, Jaroszewski L, Rychlewski L, Godzik A (2000) Sensitive sequence comparison as protein function predictor. Pac Symp Biocomput 5: 42–53.
32. Thornton JM (2001) From genome to function. Science 292: 2095–2097.
33. Koonin EV, Wolf YI, Karev GP (2002) The structure of the protein universe and genome evolution. Nature 420: 218–223.
34. Rost B, Liu J, Nair R, Wrzeszczynski KO, Ofran Y (2003) Automatic prediction of protein function. Cell Mol Life Sci 60: 2637–2650.
35. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, et al. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc Natl Acad Sci U S A 98: 4569–4574.
36. Matthews L, Vaglio P, Reboul J, Ge H, Davis B, et al. (2001) Identification of potential interaction networks using sequence-based searches for conserved protein–protein interactions or "interologs." Genome Res 11:: 2120–2126.
37. Yu H, Luscombe N, Lu H, Zhu X, Xia Y, et al. (2004) Annotation transfer between genomes: Protein–protein interologs and protein-DNA regulogs. Genome Res 14: 1107–1118.

38. Chen R, Jeong S (2000) Functional prediction: Identification of protein orthologs and paralogs. Protein Sci 9: 2344–2353.
39. Tatusov R, Koonin E, Lipman D (1997) A genomic perspective on protein families. Science 278: 631–637.
40. Tirosh I, Barkai N (2005) Computational verification of protein–protein interactions by orthologous co-expression. BMC Bioinformatics 6: 40.
41. Lehner B, Fraser AG (2004) A first-draft human protein-interaction map. Genome Biol 5: R63.
42. Bhardwaj N, Lu H (2005) Correlation between gene expression profiles and protein–protein interactions within and across genomes. Bioinformatics 21: 2730–2738.
43. Bowers PM, Cokus SJ, Eisenberg D, Yeates TO (2004) Use of logic relationships to decipher protein network organization. Science 306: 2246–2249.
44. Ofran Y, Rost B (2003) Analysing six types of protein–protein interfaces. J Mol Biol 325: 377–387.
45. Hermjakob H, Montecchi-Palazzi L, Lewington C, Mudali S, Kerrien S, et al. (2004) IntAct: An open source molecular interaction database. Nucleic Acids Res 32: D452–D455.
46. Liu J, Rost B (2001) Comparing function and structure between entire proteomes. Protein Sci 10: 1970–1979.
47. Aloy P, Russell RB (2002) The third dimension for protein interactions and complexes. Trends Bioch Sci 27: 633–638.
48. von Mering C, Krause R, Snel B, Cornell M, Oliver S, et al. (2002) Comparative assessment of large-scale data sets of protein–protein interactions. Nature 417: 399–403.
49. Salwinski L, Eisenberg D (2003) Computational methods of analysis of protein–protein interactions. Curr Opin Struct Biol 13: 377–382.
50. Tamura S, Matsumoto N, Imamura A, Shimozawa N, Suzuki Y, et al. (2001) Phenotype–genotype relationships in peroxisome biogenesis disorders of PEX1-defective complementation group 1 are defined by Pex1p–Pex6p interaction. Biochem J 357: 417–426.
51. Kiel JA, Hilbrands RE, van der Klei IJ, Rasmussen SW, Salomons FA, et al. (1999) *Hansenula polymorpha Pex1p* and *Pex6p* are peroxisome-associated AAA proteins that functionally and physically interact. Yeast 15: 1059–1078.
52. Titorenko VI, Smith JJ, Szilard RK, Rachubinski RA (2000) Peroxisome biogenesis in the yeast *Yarrowia lipolytica*. Cell Biochem Biophys 32 (Spring): 21–26.
53. Matsumoto N, Tamura S, Fujiki Y (2003) The pathogenic peroxin Pex26p recruits the Pex1p-Pex6p AAA ATPase complexes to peroxisomes. Nat Cell Biol 5: 454–460.
54. Stewart BA, Mohtashami M, Rivlin P, Deitcher DL, Trimble WS, et al. (2002) Dominant-negative NSF2 disrupts the structure and function of *Drosophila* neuromuscular synapses. J Neurobiol 51: 261–271.
55. Mohtashami M, Stewart BA, Boulianne GL, Trimble WS (2001) Analysis of the mutant *Drosophila N*-ethylmaleimide sensitive fusion-1 protein in comatose reveals molecular correlates of the behavioural paralysis. J Neurochem 77: 1407–1417.
56. Pullikuth AK, Gill SS (1999) Identification of a Manduca sexta NSF ortholog, a member of the AAA family of ATPases. Gene 240: 343–354.
57. Stewart BA, Pearce J, Bajec M, Khorana R (2005) Disruption of synaptic development and ultrastructure by *Drosophila* NSF2 alleles. J Comp Neurol 488: 101–111.
58. Xu J, Du W (2003) *Drosophila chk2* plays an important role in a mitotic checkpoint in syncytial embryos. FEBS Lett 545: 209–212.
59. Brodsky MH, Weinert BT, Tsang G, Rong YS, McGinnis NM, et al. (2004) *Drosophila melanogaster* MNK/Chk2 and p53 regulate multiple DNA repair and apoptotic pathways following DNA damage. Mol Cell Biol 24: 1219–1231.
60. Masrouha N, Yang L, Hijal S, Larochelle S, Suter B (2003) The Drosophila *chk2* gene *loki* is essential for embryonic DNA double-strand-break checkpoints induced in S phase or G2. Genetics 163: 973–982.
61. Meier B, Ahmed S (2001) Checkpoints: Chromosome pairing takes an unexpected twist. Curr Biol 11: R865–R868.
62. Korkin D, Davis FP, Sali A (2005) Localization of protein-binding sites within families of proteins. Protein Sci 14: 2350–2360.
63. Fariselli P, Olmea O, Valencia A, Casadio R (2001) Progress in predicting inter-residue contacts of proteins with neural networks and correlated mutations. Proteins Suppl 5: 157–162.
64. Fariselli P, Olmea O, Valencia A, Casadio R (2001) Prediction of contact maps with neural networks and correlated mutations. Protein Eng 14: 835–843.
65. Sprinzak E, Margalit H (2001) Correlated sequence-signatures as markers of protein–protein interaction. J Mol Biol 311: 681–692.
66. Ofran Y, Rost B (2003) Predicted protein–protein interaction sites from local sequence information. FEBS Lett 544: 236–239.
67. Apweiler R, Bairoch A, Wu C, Barker W, Boeckmann B, et al. (2004) UniProt: The Universal Protein knowledgebase. Nucleic Acids Res 32: D115–D119.
68. Breitkreutz B, Stark C, Tyers M (2003) The GRID: The General Repository for Interaction Datasets. Genome Biol 4: R23.
69. Alfarano C, Andrade CE, Anthony K, Bahroos N, Bajec M, et al. (2005) The Biomolecular Interaction Network Database and related tools 2005 update. Nucleic Acids Res 33: D418–D424.
70. Zanzoni A, Montecchi-Palazzi L, Quondam M, Ausiello G, Helmer-Citterich M, et al. (2002) MINT: A Molecular INTeraction database. FEBS Lett 513: 135–140.
71. Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, et al. (2004) The Database of Interacting Proteins: 2004 update. Nucleic Acids Res 32: D449–D451.
72. Xenarios I, Salwinski L, Duan XJ, Higney P, Kim SM, et al. (2002) DIP, the Database of Interacting Proteins: A research tool for studying cellular networks of protein interactions. Nucleic Acids Res 30: 303–305.
73. Hermjakob H, Montecchi-Palazzi L, Bader G, Wojcik J, Salwinski L, et al. (2004) The HUPO PSI's molecular interaction format.—A community standard for the representation of protein interaction data. Nat Biotechnol 22: 177–183.
74. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res 25: 3389–3402.
75. Przybylski D, Rost B (2002) Alignments grow, secondary structure prediction improves. Proteins 46: 197–205.
76. Mika S, Rost B (2003) UniqueProt: Creating representative protein sequence sets. Nucleic Acids Res 31: 3789–3791.
77. Sander C, Schneider R (1991) Database of homology-derived structures and the structural meaning of sequence alignments. Proteins 9: 56–68.
78. Efron B, Halloran E, Holmes S (1996) Bootstrap confidence levels for phylogenetic trees. Proc Natl Acad Sci U S A 93: 13429–13434.
79. Kemmer D, Huang Y, Shah S, Lim J, Brumm J, et al. (2005) Ulysses—An application for the projection of molecular interactions across species. Genome Biol 6:: R106. Epub 2 December 2005.