Integration and Prediction of PPI Using Multiple Resources from Public Databases

Ramón Aragues*, Javier García-García*, Baldo Oliva §

Structural Bioinformatics Lab. (GRIB). Universitat Pompeu Fabra-IMIM. Barcelona Research Park of Biomedicine (PRBB). 08003-Barcelona, Catalonia, Spain.

§Corresponding author: Baldo Oliva, Structural Bioinformatics Lab. (GRIB). Universitat Pompeu Fabra-IMIM. Barcelona Research Park of Biomedicine (PRBB).

Received June 24, 2008; Accepted July 16, 2008; Published July 17, 2008

Citation: Ramón A, Javier GG, Baldo O (2008) Integration and Prediction of PPI Using Multiple Resources from Public Databases. J Proteomics Bioinform 1: 166-187.

Copyright: © 2008 Ramón A, etal. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abstract

Background: The analysis and usage of biological data is hindered by the spread of information across multiple repositories and the difficulties posed by different nomenclature systems and storage formats. In particular, the study and use of protein-protein interactions is one area where there is an important need for data integration. Without good integration strategies, it is difficult to assess how much interaction data is available and its properties.

Results: We present a data integration approach for protein-protein interactions. This integrative approach has been implemented into PIANA, a protein-protein interaction software framework under the GNU Public License (http://sbi.imim.es/piana). We find that the integrated network of interactions shows properties very similar to those observed in previously reported protein interaction networks. We also find that interaction prediction methods find interactions for many proteins for which experimental methods have not produced any information.

Conclusions: PIANA´s approach to protein interaction data integration solves many of the nomenclature issues common to systems dealing with biological data. The concept presented here can be extended to other types of biological data. The integration of all available protein interaction data is fundamental to obtaining a comprehensive picture of the interactions taking place in the cell.

Keywords: Protein-protein interaction; database integration; protein identifiers

Introduction

The completion of genome sequencing projects stimulated the development of high-throughput experimental methods aimed at functional characterization of the discovered genes. In particular, the identification of protein-protein interactions has been accelerated by the development of new technologies such as two-hybrid assays (Parrish et al., 2006; Rual et al., 2005; Stelzl et al., 2005) and affinity purifications followed by mass spectrometry (Gavin et al., 2006; Krogan et al., 2006; Puig et al., 2001). Thus, a vast amount of protein-protein interaction data has been collected, including proteome-scale interactome maps for yeast (Ito et al., 2001; Uetz et al., 2000), fly (Giot et al., 2003) and worm (Li et al., 2004), and a partial map for human (Rual et al., 2005; Stelzl et al., 2005). In addition to providing insights about biological systems (Barabasi et al., 2004; Cusick et al., 2005), protein interaction maps can be used to infer the function of proteins (Sharan et al., 2007), detect remote homologs (Espadaler et al., 2005a) and to identify the binding sites of a protein (Kim et al., 2006).

However, interaction data is spread across multiple repositories and codified using various nomenclature systems (Mathivanan et al., 2006). In consequence, experimental biologists face difficulties when trying to find all known interactions for their proteins of interest, and the computational analysis and usage of protein interaction data is usually constrained to using a partial subset of all available knowledge. For example, any comprehensive search of interactions for a particular protein must include at least seven databases of protein-protein interactions: the Database of Interacting Proteins (DIP) (Salwinski et al., 2004), the MIPS database of interactions (Pagel et al., 2005), the Molecular
INTerations database (MINT) (Chatr-aryamontri et al., 2007), IntAct (Kerrien et al., 2007), the Biomolecular Interactions Database (BIND) (Alfarano et al., 2005), the BioGrid (Stark et al., 2006) and Human Protein Reference Database (HPRD) (Peri et al., 2003).

Besides, each database uses different strategies for identifying proteins, and translations between synonym identifiers (i.e. identifiers linked to the same protein sequence) are required before any manual search or automatic processing. Moreover, there are methods for predicting protein interactions that can be used when no experimental interactions have been detected for a protein, but results from these methods are usually spread across multiple websites, each in its own format.

There are efforts to standardize and harmonize protein interaction data. HUPO-PSI (Hermjakob, 2006) has developed a schema that enables the description of interactions between a wide range of molecular types, thus facilitating the access and data exchange between different research groups. The IMEx consortium (Orchard et al., 2007) is a group of major public interaction data providers sharing curation effort and exchanging completed records on molecular data following the HUPO standard exchange format. In consequence, the rate of data curation and data sharing between different repositories has been improved, but integration is still not completed. For example, HUPO PSI-MI 2.5 format allows the identification of interactors by unique identifiers from different databases, but the guidelines implemented do not include a strategy for naming proteins, which leaves unresolved many of the integration issues.

The issue of protein nomenclature has been addressed by internationally recognized scientific organizations like HGNC (Wain et al., 2002) and SGD (Christie et al., 2004), but they do not cover all species and do not map all database identifiers. IPI (Kersey et al., 2004) offers a non-complete redundant data set with cross-references with external identifiers. PIANA (Protein Interactions And Network Analysis), a framework for creating, managing and analyzing protein-protein interactions (Aragues et al., 2006). Here, we describe the PIANA approach to protein nomenclature and its strategy to protein-protein interaction data integration. Furthermore, we describe the properties of the experimental interaction network obtained for all species by integrating interactions from DIP (Salwinski et al., 2004), MIPS (Pagel et al., 2005), MINT (Chatr-aryamontri et al., 2007), IntAct (Kerrien et al., 2007), BIND (Alfarano et al., 2005), BioGrid (Stark et al., 2006) and HPRD (Peri et al., 2003). We also describe the properties of the interaction networks obtained from different methods of protein-protein interaction prediction. We conclude by discussing potential enhancements to the integration approach here described.

Materials and Methods

Interaction Networks Based on ProteinIDs and other External Identifiers

Interaction networks are built using proteinIDs as nodes
(see sections ‘Mapping protein identifiers’ and ‘Protein-protein interactions integration’). When translating the nodes of the network to external protein identifiers (process referred as ‘unifying the network’), there are two possibilities: 1) one proteinID corresponds to a single external identifier and 2) different proteinIDs correspond to the same identifier, and thus, nodes and interactions are merged. Therefore, the same PIANA proteinID network will correspond to different unified networks, depending on the external identifier. Statistics in this article have been obtained after unifying the networks by NCBI geneID. Although geneIDs only cover 42% of proteinIDs, the cardinality proteinID:externalIdentifier is the highest (Table 1), and therefore geneID is the best suited identifier type for obtaining an unbiased view of the integrated protein interaction network. Protein sequences of unknown geneID were unified using UniProt accessions.

### Table 1: Protein identifiers statistics.

Summary of the most relevant protein identifier types, calculated from a total of 6,476,028 distinct sequenceIDs in the database. Columns are: identifier type, number of distinct identifiers, the proportion of proteinIDs with respect to external identifier correspondences, the proportion of external identifiers with respect to proteinIDs, and the percentage of proteinIDs covered by the external identifier. Primary gene symbols are those gene symbols that have been established as the official gene name by nomenclature authorities such as HUGO (Wain et al., 2002) or FlyBase (Crosby et al., 2007).

| Identifier Type       | Number of distinct identifiers | External Identifier: ProteinID | ProteinID: External Identifier | % proteinIDs covered |
|-----------------------|--------------------------------|--------------------------------|--------------------------------|----------------------|
| Uniprot Accession     | 4639397                        | 1:1 99.74                      | 1:1 96.98                      | 68%                  |
|                       |                                | 1:2 0.07                       | 1:2 0.98                       |                      |
| NCBI Accession        | 10760685                       | 1:1 98.44                      | 1:1 24.85                      | 96%                  |
|                       |                                | 1:2 0.83                       | 1:2 26.78                      |                      |
| NCBI geneID           | 2416561                        | 1:1 90.77                      | 1:1 98.41                      | 42%                  |
|                       |                                | 1:2 1.87                       | 1:2 0.17                       |                      |
| Gene Symbol           | 4143090                        | 1:1 81.10                      | 1:1 79.58                      | 80%                  |
|                       |                                | 1:2 3.95                       | 1:2 6.57                       |                      |
| Primary Gene Symbol   | 1207358                        | 1:1 90.34                      | 1:1 97.70                      | 42%                  |
|                       |                                | 1:2 5.27                       | 1:2 0.13                       |                      |

Methods for the Prediction of Protein Interactions

We used predictions of protein-protein interactions obtained by four different methods: (i) Gene fusion, in which two proteins are predicted to interact if their corresponding genes appear fused in another genome (Enright et al., 1999); (ii) Phylogenetic profiles, in which similarity of phylogenetic profiles is interpreted as being indicative of two proteins need to be simultaneously present to perform a given function together (Pellegrini et al., 1999); (iii) Distant conservation of sequence patterns and structure relationships, in which structural similarities among domains of known interacting proteins and conservation of pairs of sequence patches involved in protein–protein interfaces are used to predict putative protein interaction pairs (Espadaler et al., 2005b); and (iv) Structural interologs, in which interactions are transferred between proteins with the same structural domains (Aragues et al., 2006). Interactions for the two first methods were retrieved from STRING (von Mering et al., 2007) by querying the database for interactions with a score higher than 0.7 for that particular methodology. Interactions for (iii) were obtained from the work of Espadaler et al. (Espadaler et al., 2005b). Interactions for (iv) were predicted by transferring experimental interactions in PIANA between proteins with a domain within the same SCOP family.

Results

Overview

PIANA (Protein Interactions And Network Analysis) (Aragues et al., 2006) is a software framework capable of
(i) integrating multiple sources of information into a single relational database (see database design on additional file 1); (ii) creating and analyzing protein interaction networks; and (iii) mapping multiple types of biological data onto protein interaction networks. PIANA code and documentation are freely available under an open source license for local installation and modification (http://sbi.imim.es/piana). The data warehousing approach and software architecture of PIANA are shown in Figure 1 (see additional file X for details). The PIANA database is accessed by the Graph library through a database interface, which is also used by the PIANA library to create, manage and analyze protein-protein interaction networks. The whole process can be controlled from a user interface module.

Mapping Protein Identifiers

PIANA handles an extensive set of protein identifiers types: UniProt entries and accessions; gene symbols; NCBI gi, geneID, Unigene and accession numbers; ENSEMBL; RefSeq; PDB; and FastA formatted sequences. PIANA internally identifies proteins with proteinIDs (integers). Each proteinID is linked to a pair [aminoacid sequence, taxonomy id], so there is a unique identifier for each protein sequence for a given organism. This allows PIANA to use the <sequence, species> of the protein as an inter-lingua between the external identifiers provided by the main repositories of genes and proteins. Therefore, one external protein identifier (e.g. UniProt entry THR_HUMAN) can be associated to one or more proteinIDs (e.g. 11483), which are in turn linked to other external identifiers that are also used to represent that protein (e.g., gene symbol ‘f2’ and Unigene ‘Hs.410092’). Consequently, along the different processes involved in inputting/outputting PIANA, external identifiers are ‘translated’ to proteinIDs, the desired operations are performed, and finally, if needed, proteinIDs are returned into the external identifier expected by the user (Figure 2). This strategy reduces the ambiguity and processing problems to the minimum: there is no need for continuously translating between distinct types of protein identifiers, since all information has been previously stored by assigning it to specific proteinIDs. Furthermore, codifying interactions in terms of proteinIDs allows PIANA to capture a larger number of interactions than platforms based on third party protein identifiers.

Figure 1: PIANA architecture

A set a parsers inserts information from external repositories into the PIANA database.

[Diagram of PIANA architecture]

J Proteomics Bioinform Volume 1(4):166-187(2008) - 169
ISSN:0974-276X  JPB, an open access journal
PIANA uses proteinIDs as an interlingua between external identifiers. PIANA keeps all information in terms of proteinIDs (an integer that uniquely identifies a protein sequence of a given taxonomy). User inputs are immediately translated to proteinIDs. Once this translation has been performed, all operations are performed at the sequence level, reducing ambiguities and synonyms conversions to a minimum.

Moreover, PIANA uses a number of techniques to assure the quality and completeness of the identifiers used as input/output: 1) inferring correspondences between identifiers and sequences even in the case that no external database explicitly contained the cross-reference: if one database identifies sequence A with identifier id1 and another database uses identifier id2 to sequence A, PIANA infers that id1 is equivalent to id2; 2) uniqueness of output protein identifiers: if two proteinIDs are linked to the same external identifier, those proteins are considered to be the same, and hence, merged into a single network node; 3) avoiding gene name ambiguities: thanks to integrating the species of the protein into the internal identifier, gene names are not confused even if the same symbol is used for several species; and 4) using representative protein identifiers: (i) PIANA will use the identifier labeled as ‘preferred’ by the source database (e.g. official gene symbol) unless the user says the contrary; and (ii) any input identifiers given by the user are prioritized over other identifiers in the PIANA database.

Protein Sequences Integration

Sequence and taxonomy data was obtained from Uniprot Swissprot and Trembl (The Uniprot Consortium, 2007), NCBI GenBank (Benson et al., 2007) and NCBI Blast nr (Maglott et al., 2007) databases (see additional file 2 for the complete list of protein sequence repositories used). Unexpectedly, UniProt Swiss-Prot (i.e. curated sequences) and UniProt TrEMBL (i.e. predicted sequences) have a significant overlap (additional file 3). Moreover, the overlap between TrEMBL and GenBank is lower than anticipated. Cross-references between external identifiers and
proteinIDs were obtained from multiple third-party repositories (see additional file 2). Table 1 shows the coverage provided by the main protein external identifiers for all proteinIDs (i.e. pair [protein sequence, taxonomy]) in the PIANA database.

MProtein-Protein Interactions Integration

Each interaction described in a third-party database is ‘translated’ to one or more interactions between proteinIDs. For example, if the external database contains an interaction between proteins A and B, with A corresponding to two proteinIDs (e.g 1 and 2) and B to one proteinID (e.g. 3), two interactions (1-3 and 2-3) will be inserted into the PIANA database. Both interactions will be described in the PIANA database as coming from that specific external database and labeled with the method used to detect the interaction between A and B. For example, HPRD describes an interaction between Entrez Gene 217 (mitochondrial ALDH) and Entrez Gene 3336 (heat shock protein). According to the correspondences in the PIANA database, Entrez Gene 217 corresponds to 13 different proteinIDs, and Entrez Gene 3336 corresponds to 12 proteinIDs. Therefore, PIANA will internally store the interaction between those two proteins as 156 different interactions. This methodology allows PIANA to give full control to the user: 1) interactions can be retrieved from any type of identifier; 2) a network can be created for a given external database (e.g. use only interactions from IntAct) and/or a specific method (e.g. do not use interactions detected in two hybrids assays) and/or a species (e.g. only interested in human interactions); 3) PIANA outputs can be set to use any type of protein identifier and therefore, interactions between proteinIDs are transformed to non-redundant interactions between protein identifiers (Methods). Consequently, describing interactions in terms of protein sequences instead of external identifiers provides a true integration of all known interactions into a single network, while keeping record of the source databases and detection methods associated with the interaction. Currently, PIANA can integrate interactions from DIP (Salwinski et al., 2004), MIPS (Page et al., 2005), MINT (Chatr-aryamontri et al., 2007), IntAct (Kerrien et al., 2007), BIND (Alfarano et al., 2005), BioGrid (Stark et al., 2006), HPRD (Perti et al., 2003), STRING (von Mering et al., 2007), interactions predicted by distant conservation of sequence patterns and structure relationships (Espadaler et al., 2005b), interactions transferred between proteins based on orthology (Yu et al., 2004) and, in general, any interaction data that is in tabulated or PSI-MI (Hermjakob, 2006) formats. See additional file 2 for the detailed description of interaction repositories that have been used in this work. Furthermore, data does not have to be integrated indiscriminately without differentiating high-throughput versus small-scale experiments and literature annotation. Therefore, PIANA allows users to define subsets of interactions based on the source repository and detection methods employed. For example, a subset of reliable interactions can be extracted by requiring them to be in at least two different repositories.

Experimental Interactions

The integrated set of experimental interactions consisted of 4,055,698 interactions between 113,785 different proteinIDs. When grouping proteinIDs by their associated NCBI geneID (Methods), there were 405,808 interactions for 53,143 proteins, an average of 7.63 interactions per protein.

Interactions Distribution

The experimental interactions in the PIANA integrated database have been obtained from 7 different repositories, belong to 736 different species, and were detected using 106 different experimental methods. As shown on Table 2, the species with the largest number of experimental interactions are yeast (111,535 interactions) and human (110,457 interactions). Most interactions were found in just one database and were detected by just one method (Figure 3). The high correlation between the number of methods and databases is explained by the fact that most interactions appear in just one external repository, and these repositories usually label interactions with a single detection method. We calculated the overlap between 7 repositories with experimental information in terms of interactions (Table 3A) and proteins (Table 3B). BioGrid (Stark et al., 2006) is the repository with the highest number of interactions (216,370) and with the highest number of unique interactions (163,700). The two repositories that show the greatest overlap are MINT and IntAct (61% of interactions and 82% of proteins in MINT are also in IntAct) while the lowest overlap was between HPRD and DIP (only 4% of interactions and 9% of proteins in HPRD are also in DIP). Most low overlaps in terms of interactions are explained by the low overlap in terms of proteins. Therefore, data integration is required in order to obtain an interaction network that covers most proteins and interactions.

We were interested in analyzing the distribution of interactions in terms of the detection method employed. We examined the overlaps between different detection methods in terms of interactions (Table 4A) and proteins (Table 4B). We observed that high-throughput methods account for most of the known interactions (126,136 for affinity methods and 103,334 for yeast two hybrid assays). The overlap between the interactions detected by the different methods is low, even in cases where the overlap at the protein level is high. For example, while 51% of proteins with interactions from affinity methods also had interactions detected by yeast two-hybrid methods, only 9% of interactions from yeast two-hybrid were also detected by affinity methods. Therefore, in order to maximize the number of known interactions for
a protein, multiple experimental detection methods should be employed.

| Species                        | Number of interactions | Number of proteins |
|--------------------------------|------------------------|--------------------|
| Saccharomyces cerevisiae       | 111535                 | 6493               |
| Homo sapiens                  | 110457                 | 36900              |
| Drosophila melanogaster       | 90562                  | 11605              |
| Escherichia coli              | 16097                  | 3467               |
| Caenorhabditis elegans        | 9184                   | 3959               |
| Mus musculus                  | 4776                   | 3495               |
| Plasmodium falciparum 3D7     | 2723                   | 1279               |
| Other species                 | 17021                  | 9724               |

Table 2: Number of protein interactions per species.
The number of interactions and proteins with at least one known interaction are shown for species with more than 2000 interactions.

Figure 3: Distribution of interactions in PIANA across different source databases and detection methods. Most interactions were found in just one database and were detected just by one method. Unspecific detection method names were not taken into account (e.g., experimental, in-vitro, in-vivo).
| A) | Total | Unique | IntAct | DIP | MINT | HPRD | MIPS | BioGrid |
|----|-------|--------|--------|-----|------|------|------|--------|
| 104339 | 42284 | 36867 (38%) | 47119 (61%) | 8729 (23%) | 87 (10%) | 25355 (22%) | 27269 (44%) | 25994 (82%) |
| 97377 | 46115 | 37210 (48%) | 825 (2%) | 34 (4%) | 19870 (20%) | 28143 (45%) | 24709 (32%) | 21036 (82%) |
| 77419 | 16392 | 8925 (23%) | 107 (13%) | 107 (13%) | 24709 (32%) | 26406 (42%) | 20550 (54%) | 21036 (82%) |
| 38372 | 10978 | 8321 (23%) | 312 (38%) | 312 (38%) | 2839 (7%) | 2839 (7%) | 2839 (7%) | 21036 (82%) |
| 833 | 389 | 87 (10%) | 34 (4%) | 34 (4%) | 24709 (32%) | 26406 (42%) | 20550 (54%) | 21036 (82%) |
| 216370 | 163700 | 21036 (82%) | 2839 (7%) | 2839 (7%) | 2839 (7%) | 2839 (7%) | 2839 (7%) | 21036 (82%) |
| 62444 | 24925 | 15611 (75%) | 11240 (53%) | 7263 (77%) | 439 (52%) | 11240 (53%) | 7263 (77%) | 21036 (82%) |

| Total | IntAct | DIP | MINT | HPRD | MIPS | BioGrid |
|-------|--------|-----|------|------|------|--------|
| 104339 | 97377 | 77419 | 38372 | 833 | 216370 | 62444 |
| Total | 21036 | 21036 | 21036 | 21036 | 21036 | 21036 |

| B) | Total | Unique | IntAct | DIP | MINT | HPRD | MIPS | BioGrid |
|----|-------|--------|--------|-----|------|------|------|--------|
| 30998 | 6273 | 15297 (74%) | 21415 (82%) | 5445 (58%) | 532 (63%) | 11893 (57%) | 13849 (66%) | 25993 (82%) |
| 20794 | 2761 | 15611 (75%) | 15611 (75%) | 834 (9%) | 219 (26%) | 7862 (38%) | 12567 (60%) | 25993 (82%) |
| 25993 | 2008 | 783 (8%) | 783 (8%) | 783 (8%) | 550 (65%) | 11240 (53%) | 13580 (65%) | 25993 (82%) |
| 9461 | 783 | 440 (52%) | 440 (52%) | 440 (52%) | 440 (52%) | 7263 (77%) | 2507 (27%) | 25993 (82%) |
| 848 | 115 | 219 (26%) | 219 (26%) | 219 (26%) | 219 (26%) | 439 (52%) | 411 (49%) | 25993 (82%) |
| 21036 | 5337 | 21036 | 21036 | 21036 | 21036 | 21036 | 21036 | 21036 |
| 21011 | 4781 | 13849 (66%) | 13849 (66%) | 13849 (66%) | 13849 (66%) | 13849 (66%) | 13849 (66%) | 25993 (82%) |

| Total | IntAct | DIP | MINT | HPRD | MIPS | BioGrid |
|-------|--------|-----|------|------|------|--------|
| 30998 | 20794 | 25993 | 9461 | 848 | 21036 | 21011 |
| Total | 21036 | 21036 | 21036 | 21036 | 21036 | 21036 |

**Table 3:** Pairwise overlaps of protein interactions and proteins for seven interaction repositories. For each repository, cells show the overlap with other repositories in terms of (A) interactions and (B) proteins. In parenthesis, the percentage that the overlap represents over the repository from the pair with less interactions or proteins is shown. Unique interactions and proteins are those only appearing in that repository. This table reflects the overlaps in the interaction network unified by NCBI geneID identifiers.
### A)

|                  | Total     | Affinity methods | Yeast two-hybrid | Phenotypic | 3D structure | Array technologies | Dosage |
|------------------|-----------|------------------|------------------|------------|--------------|--------------------|--------|
| 126136           |           |                  |                  |            |              |                    |        |
| 103334           |           |                  |                  |            |              |                    |        |
| 72159            |           |                  |                  |            |              |                    |        |
| 6525             |           |                  |                  |            |              |                    |        |
| 4627             |           |                  |                  |            |              |                    |        |
| 3914             |           |                  |                  |            |              |                    |        |
| 3104             |           |                  |                  |            |              |                    |        |

|                  | Total     | Affinity methods | Yeast two-hybrid | Phenotypic | 3D structure | Array technologies | Dosage |
|------------------|-----------|------------------|------------------|------------|--------------|--------------------|--------|
|                  | 126136    |                  | 103334           | 72159      | 6525         | 4627               | 3914   |

### B)

|                  | Total     | Affinity methods | Yeast two-hybrid | Phenotypic | 3D structure | Array technologies | Dosage |
|------------------|-----------|------------------|------------------|------------|--------------|--------------------|--------|
| 20043            |           |                  |                  |            |              |                    |        |
| 33246            |           |                  |                  |            |              |                    |        |
| 5619             |           |                  |                  |            |              |                    |        |
| 5737             |           |                  |                  |            |              |                    |        |
| 907              |           |                  |                  |            |              |                    |        |
| 2004             |           |                  |                  |            |              |                    |        |
| 1083             |           |                  |                  |            |              |                    |        |

|                  | Total     | Affinity methods | Yeast two-hybrid | Phenotypic | 3D structure | Array technologies | Dosage |
|------------------|-----------|------------------|------------------|------------|--------------|--------------------|--------|
|                  | 20043     |                  | 33246            | 5619       | 5737         | 907                | 2004   |

**Table 4**: Pairwise overlaps of protein interactions and proteins for seven detection methods.

For each detection method, cells show the overlap with other methods in terms of (A) interactions and (B) proteins. In parenthesis, the percentage that the overlap represents over the method from the pair with less interactions or proteins is shown. This table reflects the overlaps in the interaction network unified by NCBI geneID identifiers.
Properties of the Experimental Integrated Protein Interaction Network

Well-documented observations about protein interaction networks are confirmed when analyzing the integrated experimental interaction networks of different species. Moreover, the integrated network shows the modular functional organization of the proteome reported by previous works (Gavin et al., 2006). In particular, proteins tend to interact with proteins of the same Gene Ontology (GO) (Harris et al., 2004) biological process (Table 5). Furthermore, 95% of the interacting proteins in the integrated network have the same cellular component according to GO. In addition, the following properties were observed for the yeast protein interaction network (Table 6): (i) yeast hubs (proteins with 5 or more interactions) are more likely to be essential (Giaever et al., 2002) than non-hubs (22% of hubs are essential versus only 5% of non-hubs), although this might be a reflection of hubs usually having multiple interfaces (Kim et al., 2006); (ii) approximately 59% of the interactions have the same cell localization according to (Lee et al., 2002); (iii) approximately 60% of the interactions reported are found coexpressed during the yeast cell cycle according to (Cho et al., 1998).

| Number of interactions with GO information | Number of interactions with common GO term |
|-------------------------------------------|------------------------------------------|
| a) Cellular Component                      | 221213                                   |
|                                           | 209254 (95%)                             |
| b) Biological Process                      | 266245                                   |
|                                           | 227254 (85%)                             |
| c) Molecular Function                       | 270008                                   |
|                                           | 141323 (52%)                             |

Table 5: Commonalities in localization, molecular function and biological process of experimentally detected interacting proteins.

This table shows the fraction of experimentally detected interacting proteins with the following properties: a) co-localized according to GO cellular component terms; b) same biological process according to GO biological process terms; and c) same molecular function according to GO molecular function terms. An interaction was considered to respect the GO restriction if both interacting proteins shared a GO term when retrieving GO parents up to level 3 (Harris et al., 2004). In parenthesis, the percentage of interactions where both interacting proteins share a GO term is shown. Interactions were used for the study only if both proteins had at least one GO term assigned. Interactions where a protein interacts with itself were discarded for this study.
Interaction partners are co-localized 37684 out of 72661 (52%)

Interaction partners are co-expressed 1524 out of 2576 (59%)

Essential yeast hubs 930 out of 4229 (22%)

Essential non yeast hubs 44 out of 886 (5%)

Table 6: Properties of the yeast protein interaction network obtained by integrating multiple sources with PIANA.

Yeast co-localization data was obtained from the work of Lee and co-workers (Lee et al., 2002). Yeast co-expression data was obtained from the work of Cho et al (Cho et al., 1998). Yeast essentiality data was obtained from the work by Giaever et al (Giaever et al., 2002). A yeast protein was considered a hub if it had 5 or more interaction partners. The interactions and proteins were included in the study for those cases in which information was available. Interactions where a protein interacts with itself were discarded for this study.

Protein Function Prediction from the Experimental Integrated Network

Recently, it has been shown that the number of common interaction partners between two proteins can be used to annotate proteins (Brun et al., 2003; Samanta et al., 2003). We have studied the use of this heuristic to predict molecular functions and biological processes as defined by GO (Harris et al., 2004), by calculating the percentage of shared GO terms between proteins with common interaction partners (Figure 4). As expected, we observe that the interactions of a protein in the integrated network can be used to predict its function and the biological processes in which it intervenes. For example, proteins with 10-20 interaction partners in common share 90% of their GO biological process terms. Moreover, the accuracy of the predictions based on the integrated network is similar to that obtained when solely using the subset of interactions from DIP (Salwinski et al., 2004), while the number of annotated proteins is much higher (additional file 4).

Predicted Interaction Networks

We were interested in assessing protein interaction predictions and evaluating the similarities between the predicted interaction network and the experimental interaction network. In particular, we studied 4 different types of predictions (Methods): (i) Gene fusion events (Enright et al., 1999) as predicted by STRING (von Mering et al., 2007); (ii) Phylogenetic profiles (Pellegrini et al., 1999) as predicted by STRING (von Mering et al., 2007); (iii) Distant conservation of sequence patterns and structure relationships as described by Espadeler et al. (Espadeler et al., 2005b); and (iv) Structural interologs predicted by PIANA (Aragues et al., 2006). We calculated the overlap between the different experimental and prediction methods in terms of interactions (Table 7A) and proteins (Table 7B), observing a high overlap between prediction methods based on genomes analyses (i.e. gene fusion events and phylogenetic profiles) and a very low overlap between all other prediction methods. This minimal overlap between interaction predictions is explained by the different types of input data used by each method and the type of proteins for which the methods are capable of predicting interactions. For example, the method based on structural interologs predicts interactions for proteins with known 3D structure, while STRING predictions from gene fusion events were mainly applied to prokaryotes. Most proteins with known 3D structure are eukaryotes (Berman et al., 2000), and therefore, the two methods rarely predict similar interactions. Moreover, there is low overlap between predicted interactions and those obtained by experimental high throughput methods, both in terms of interactions and proteins. These results indicate that different methods identify interactions for different proteins. For example, there are many species for which no yeast two-hybrid experiments have been carried out, while many predictions can be ‘transferred’ to those species on the basis of genomes analysis, resulting in a low overlap at the interaction and protein level between the two methods.
Figure 4. Function prediction based on common interaction partners in the integrated experimental network.

The percentage of shared GO terms is plotted as a function of the number of common interaction partners.
Table 7: Pairwise overlaps of protein interactions and proteins for four interaction prediction methods, two types of high-throughput methods (yeast two hybrid assays and affinity purification methods), and curated data (in vitro and in vivo).

For each method, cells show the overlap with other methods in terms of (A) interactions and (B) proteins. In parenthesis, the percentage that the overlap represents over the method from the pair with less interactions or proteins is shown. This table reflects the overlaps in the interaction network unified by NCBI geneID identifiers.

We evaluated whether interacting proteins according to different prediction methods tended to share biological process, molecular function and cellular component according to GO (Table 8). We observed that the method that better captures functional relationships between proteins is the one based on gene fusion events (Methods): 85% of the predicted interacting pairs belong to the same biological process. Moreover, all prediction methods detected a sensible number of colocalized proteins. For example, 87% of interacting proteins according to the prediction method based on structural interologs had the same cellular location.

**Discussion**

We presented the data integration approach of PIANA, a software framework designed for creating, managing and analyzing protein-protein interaction networks. PIANA was created to address nomenclature and integration issues common in protein interaction repositories and network visualization tools. Moreover, the modular approach of PIANA makes it a useful resource for bioinformaticians wishing to avoid the low-level details related to working with protein interaction networks.

Many areas of biological research are hampered by the difficulties found in accessing all biological information available. In particular, protein-protein interactions analysis is usually biased by the input sources of data. PIANA is one of the very few protein interaction platforms where all interactions from all external databases can be found for a protein of interest, regardless of the type of identifier used as input or the name given to the protein by the researcher that submitted the interactions. We presented a detailed analysis of the protein-protein interactions in the integrated network, in terms of their distribution across different databases and detection methods. We showed that most interactions appear in just one database and the overlap in terms of interactions is below 50% between most repositories, reinforcing the need for tools that unify all known interactions into a single network. Moreover, this integrated network has been shown to agree with properties previously reported about protein-protein interaction networks retrieved from just one database/detection method, such as its capability of predicting the function of proteins. Besides, the overlap between different experimental and prediction methods for protein-protein interaction identification was low, both in terms of interactions and proteins for which at least one interaction has been described. Despite this low overlap, interaction prediction approaches such as those based on gene fusion events and structural interologs were successful at identifying pairs of proteins within the same GO biological process. However, more in-depth studies are undertaken to evaluate the ability of annotating proteins based on interaction predictions (Espadaler et al., 2008).
Our analysis of protein interaction data in the public domain is similar to the studies of Herzel et al. (Futschik et al., 2007) and Pandey and coworkers (Mathivanan et al., 2006). However, our study includes protein interactions for all species, as well as predicted interactions from diverse methods. Moreover, we have analyzed the overlap between diverse experimental and prediction methods. The main conclusions from the studies in (Futschik et al., 2007) and (Mathivanan et al., 2006) are confirmed for interactions for organisms other than human. However, we found a higher overlap between the different interaction repositories, probably due to recent efforts in data exchange. Moreover, the total number of interactions in the experimental human integrated network is 110,457, compared to the 154,000-369,000 interactions estimated by Marcotte and coworkers (Hart et al., 2006).

PIANA’s approach to data integration is a good equilibrium between reliability and flexibility, while giving a good coverage of the information available. Two potential improvements to the current integration approach are: (i) the implementation of more sophisticated gene name disambiguation (Schijvenaars et al., 2005; Xu et al., 2007); and (ii) the capability of detecting highly similar protein sequences (e.g. via sequence alignments) and thus, transferring interactions and identifiers between similar proteins. The data integration techniques described here could also be of help for areas other than protein-protein interactions, such as gene expression studies or regulatory networks.

### Conclusions

Our approach to data integration is based on using the...
sequence of proteins as an interlingua between the different identifiers. This strategy allows PIANA, our protein-protein interaction software platform to integrate data from multiple sources into a single interaction network, while allowing the user to control which interactions are used in the analyses. The low overlap found between the different repositories of interaction data reinforces the need for integration tools. Moreover, we found that the integrated network of interactions shows properties similar to those previously reported for partial interaction networks. Finally, we observed that interaction predictions are not as accurate as experimentally detected interactions in tasks such as protein annotation. However, prediction methods can help experimental methods to cover a larger portion of the interactome space.

Authors’ Contributions

RA designed PIANA and wrote the manuscript. JGG and RA implemented the code and performed analyses. BO conceived of the PIANA project and provided scientific guidance. JGG and BO helped draft the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We thank members of the UPF-IMIM SBI lab and P. Boixeda for their helpful comments. R.A is supported by a grant from the Spanish Ministerio de Ciencia y Tecnología (MCyT, BIO2002-03609). J.GG. is supported by a FI grant from the Catalonia Agència de Gestió d’Ajuts Universitaris i de Recerca del Departament d’Innovació, Empresa I Universitats de la Generalitat de Catalunya. The work has been supported by grants from the Spanish Ministerio de Educación y Ciencia (MEC, BIO2005-00533, PROFIT PSE-010000-2007-1 and FIT-350300-2006-40/41/42).

References

1. Aittokallio T, Schwikowski B (2006) Graph-based methods for analysing networks in cell biology. Brief Bioinform 7: 243-255.

2. Alfarano C, Andrade CE, Anthony K, Bahroos N, Bajec M, Bantoft K, Betel D, Bobeckho B, Boutilier K, Burgess E, Buzadzija K, Cavero R, D Abreo C, Donaldson I, Dorairajoo D, Dumontier MJ, Dumontier MR, Earles V, Farrall R, Feldman H, Gardeman E, Gong Y, Gonzaga, R, Grytsan V, Gryz E, Gu V, Haldorsen E, Halupa A, Haw R, Hrvojic A, Hurrell L, Isserlin R, Jack F, Juma F, Khan A, Kon T, Konopinsky S, Le V, Lee E, Ling S, Magidin M, Moniakis J, Montojo J, Moore S, Muskat B, Ng I, Paraiso JP, Parker B, Pintilie G, Pirone R, Salama JJ, Sgro S, Shan T, Shu Y, Siew J, Skinner D, Snyder K, Stasiuk R, Strumpf D, Tuckam B, Tao S, Wang Z, White M, Willis R, Wolting C, Wong S, Wrong A, Xin C, Yao R, Yates B, Zhang S, Zheng K, Pawson T, Ouellette BF, Hogue CW (2005) The Biomolecular Interaction Network Database and related tools 2005 update. Nucleic Acids Res 33: D418-424.

3. Aragues R, Jaeggi D, Oliva B (2006) PIANA: protein interactions and network analysis. Bioinformatics 22: 1015-1017.

4. Barabasi AL, OltvaiZN (2004) Network biology: understanding the cell’s functional organization. Nat Rev Genet 5: 101-113.

5. Benson DA, Karsch MI, Lipman DJ, Ostell J, Wheeler DL (2007) GenBank. Nucleic Acids Res 35: D21-25.

6. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. Nucleic Acids Res 28: 235-242.

7. Breitkreutz BJ, Stark C, Tyers M (2003) Osprey: a network visualization system. Genome Biol 4: R22.

8. Brun C, Chevenet F, Martin D, Wojcik J, Guenoche A, Jacq B (2003) Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network. Genome Biol 5: R6.

9. Cerami EG, Bader GD, Gross BE, Sander C (2006) cPath: open source software for collecting, storing, and querying biological pathways. BMC Bioinformatics 7: 497.

10. Cline MS, Smoot M, Cerami E, Kuchinsky A, Landys N, Workman C, Christmas R, Avila-Campilo I, Creech M, Gross B, Hanspers K, Isserlin R, Kelley R, Killcoyne S, Lotia S, Maere S, Morris J, Ono K, Pavlovic V, Pico AR, Vailaya A, Wang PL, Adler A, Conklin BR, Hood L, Kuiper M, Sander C, Schmuelevich I, Schwikowski B, Warner GJ, Ideker T, Bader GD (2007) Integration of biological networks and gene expression data using Cytoscape. Nature protocols 2: 2366-2382.

11. Crosby MA, Goodman JL, Strellets VB, Zhang P, Gelbart WM (2007) FlyBase: genomics by the dozen. Nucleic acids research 35: D486-491.

12. Cusick ME, Klitgord N, Vidal M, Hill DE (2005) Interactome: gateway into systems biology. Hum Mol Genet 14 Spec No 2: R171-181.
13. Chatraryamontri A, Ceol A, Palazzi LM, Nardelli G, Schneider MV, Castagnoli L, Cesareni G (2007) MINT: the Molecular INTeraction database. Nucleic Acids Res 35: D572-574.

14. Chaurasia G, Iqbal Y, Hanig C, Herzel H, Wanker EE, Futschik ME (2007) UniHI: an entry gate to the human protein interactome. Nucleic acids research 35: D590-594.

15. Cho RJ, Campbell MJ, Winzeler EA, Steinmetz L, Conway A, Wolfsberg TG, Gabrielian AE, Landsman D, Lockhart DJ, Davis RW (1998) A genomewide transcriptional analysis of the mitotic cell cycle. Molecular cell 2: 65-73.

16. Christie KR, Weng S, Balakrishnan R, Costanzo MC, Dolinski K, Dwight SS, Engel SR, Feierbach B, Fisk DG, Hirschman JE, Hong EL, Issel-Tarver L, Nash R, Sethuraman A, Starr B, Theesfeld CL, Andrade R, Binkley G, Dong Q, Lane C, Schroeder M, Botstein D, Cherry JM (2004) Saccharomyces Genome Database (SGD) provides tools to identify and analyze sequences from Saccharomyces cerevisiae and related sequences from other organisms. Nucleic acids research 32: D311-314.

17. Enright AJ, Iliopoulos I, Kyprides NC, Ouzounis CA (1999) Protein interaction maps for complete genomes based on gene fusion events. Nature 402: 86-90.

18. Espadaler J, Aragues R, Eswar N, Marti-Renom MA, Querol E, Aviles FX, Sali A, Oliva B (2005a) Detecting remotely related proteins by their interactions and sequence similarity. Proc Natl Acad Sci USA 102: 7151-7156.

19. Espadaler J, Eswar N, Querol E, Aviles FX, Sali A, Marti-Renom MA, Oliva B (2008) Prediction of enzyme function by combining sequence similarity and protein interactions. BMC bioinformatics 9: 249.

20. Espadaler J, Romero IO, Jackson RM, Oliva B (2005b) Prediction of protein-protein interactions using distant conservation of sequence patterns and structure relationships. Bioinformatics 21: 3360-3368.

21. Futschik ME, Chaurasia G, Herzel H (2007) Comparison of human protein-protein interaction maps. Bioinformatics (Oxford, England) 23: 605-611.

22. Gavin AC, Aloy P, Grandi P, Krause R, Boesche M, Marzioch M, Rau C, Jensen LJ, Bastuck S, Dumpe1ffeld B, Edelmann A, Heurtier MA, Hoffman V, Hofert C, Klein K, Hudak M, Michon AM, Schelder M, Schirle M, Remor M, Rudi T, Hooper S, Bauer A, Bouwmeester T, Casari G, Drewes G, Neubauer G, Rick JM, Kuster B, Bork P, Russell RB, Superti-Furga G (2006) Proteome survey reveals modularity of the yeast cell machinery. Nature 440: 631-636.

23. Giaever G, Chu AM, Ni L, Connelly C, Riles L, Veronneau S, Dow S, Lucau-Danila A, Anderson K, Andre B, Arkin AP, Astromoff A, El-Bakkoury M, Bangham R, Benito R, Brachat S, Campanaro S, Curtiss M, Davis K, Deutschbauer A, Entian KD, Flaherty P, Fourn Y, Garfinkel DJ, Gerstein M, Gotte D, Guldener U, Hegemann JH, Hempel S, Herman Z, Jaramillo DF, Kelly DE, Kelly SL, Kotter P, LaBonte D, Lamb DC, Lan N, Liang H, Liao H, Liu L, Luo C, Lussier M, Mao R, Menard P, Ooi SL, Revuelta JL, Roberts CJ, Rose M, Ross-Macdonald P, Scherens B, Schimmack G, Shafer B, Shoemaker DD, Sookhiah-Mahadeo S, Storms RK, Strathern JN, Valle G, Voet M, Volckaert G, Wang CY, Ward TR, Wilhelmy J, Winzeler EA, Yang Y, Yen G, Youngman E, Yu K, Bussey H, Boeke JD, Snyder M, Philipsen P, Davis RW, Johnston M (2002) Functional profiling of the Saccharomyces cerevisiae genome. Nature 418: 387-391.

24. Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, Li Y, Hao YL, Ooi CE, Godwin B, Vitols E, Vijayadamodar G, Pochart P, Machineni H, Welsh M, Kong Y, Zerhusen B, Malcolm R, Varrone Z, Collis A, Minto M, Burgess S, McDaniel L, Stimpson E, Spriggs F, Williams J, Neurath K, Ioime N, Agee M, Voss E, Furtak K, Renzulli R, Aanensen N, Carrola S, Bickelhaupt E, Lazovatsky Y, DaSilva A, Zhong J, Stanyon CA, Finley RL Jr, White KP, Braverman M, Jarvie T, Gold S, Leach M, Knight J, Shimkets RA, McKenna MP, Chant J, Rothberg JM (2003) A protein interaction map of Drosophila melanogaster. Science 302: 1727-1736.

25. Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C, Richter J, Rubin GM, Blake JA, Bult C, Dolan M, Drabkin H, Eppig JT, Hill DP, Ni L, Ringwald M, Balakrishnan R, Cherry JM, Christie KR, Costanzo MC, Dwight SS, Engel S, Fisk DG, Hirschman JE, Hong EL, Nash RS, Sethuraman A, Theesfeld CL, Botstein D, Dolinski K, Feierbach B, Berardini T, Mundodi S, Rhee SY, Apweiler R, Barrell D, Camon E, Dimmer E, Lee V, Chisholm R, Gaudet P, Kibbe W, Kishore R, Schwarz EM, Sternberg P, Griss M, Hannick L, Wortman J, Berriman M, Wood V, de la Cruz N, Tonellato P, Jaiswal...
P. Seigfried T, White R (2004) The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res 32: D258-261.

26. Hart GT, Ramani AK, Marcotte EM (2006) How complete are current yeast and human protein-interaction networks? Genome Biol 7: 120.

27. Hermjakob H (2006). The HUPO Proteomics Standards Initiative - Overcoming the Fragmentation of Proteomics Data. Proteomics 6 Suppl 2: 34-38.

28. Hu Z, Mellor J, Wu J, DeLisi C (2004) VisANT: an online visualization and analysis tool for biological interaction data. BMC Bioinformatics 5: 17.

29. Iragne F, Nikolski M, Mathieu B, Aubé D, Sherman D (2005) ProViz: protein interaction visualization and exploration. Bioinformatics 21: 272-274.

30. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc Natl Acad Sci USA 98: 4569-4574.

31. Jayapandian M, Chapman A, Tarcea VG, Yu C, Elkiss A, Ianni A, Liu B, Nandi A, Santos C, Andrews P, Atthey B, States D, Jagadish HV (2007) Michigan Molecular Interactions (MiMI): putting the jigsaw puzzle together. Nucleic acids research 35: D566-571.

32. Kerrien S, Alam FY, Aranda B, Bancarz I, Bridge A, Derow C, Dimmer E, Feuermann M, Friedrichsen A, Huntley R, Kohler C, Khadake J, Leroy C, Liban A, Lieftink C, Montecchi-Palazzi L, Orchard S, Rost B, Robb K, Roechert B, Thorneycroft D, Zhang Y, Apweiler R, Hermjakob H (2007) IntAct—open source resource for molecular interaction data. Nucleic Acids Res 35: D561-565.

33. Kersey PJ, Duarte J, Williams A, Karavidopoulou Y, Birney E, Apweiler R (2004) The International Protein Index: an integrated database for proteomics experiments. Proteomics 4: 1985-1988.

34. Kim PM, Lu LJ, Xia Y, Gerstein MB (2006) Relating three-dimensional structures to protein networks provides evolutionary insights. Science 314: 1938-1941.

35. Krokan NJ, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, Li J, Pu S, Datta N, Tikuisis AP, Punna T, Peregrin-Alvarez JM, Shales M, Zhang X, Davey M, Robinson MD, Paccanaro A, Bray JE, Sheung A, Beattie B, Richards DP, Canadien V, Laliev A, Mena F, Wong P, Starostine A, Canete MM, Vlasblom J, Wu S, Orsi C, Collins SR, Chandran S, Haw R, Rilstone JJ, Gandi K, Thompson NJ, Musso G, St Onge P, Ghanny S, Lam MH, Butland G, Altaf-Ul AM, Kanaya S, Shilatifard A, O Shea E, Weissman JS, Ingles CJ, Hughes TR, Parkinson J, Gerstein M, Wodak SJ, Emili A, Greenblatt JF (2006) Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature 440: 637-643.

36. Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, Zeitlinger J, Jennings EG, Murray HL, Gordon DB, Ben W, Wyrick JJ, Tagne JB, Volkert TL, Fraenkel E, Gifford DK, Young RA (2002) Transcriptional regulatory networks in Saccharomyces cerevisiae. Science New York NY 298: 799-804.

37. Li S, Armstrong CM, Bertin N, Ge H, Milstein S, Boxem M, Vidalain PO, Han JD, Chesneau A, Hao T, Goldberg DS, Li N, Martinez M, Rual JF, Lamesch P, Xu L, Tewari M, Wong SL, Zhang LV, Berzig GF, Jacobot L, Voglio P, Reboul J, Hirozane-Kishikawa T, Li Q, Gabel HW, Elewa A, Baumgartner B, Rose DJ, Yu H, Bosak S, Sequerra R, Fraser A, Mango SE, Saxton WM, Strome S, Van Den Heuvel S, Piano F, Vandenhaute J, Sardet C, Gerstein M, Doucette-Stamm L, Gunsalus KC, Harper JW, Cusick ME, Roth FP, Hill DE, Vidal M (2004) A map of the interactome network of the metazoan C. elegans. Science 303: 540-543.

38. Maglott D, Ostell J, Pruitt KD, Tatusova T (2007) Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res 35: D26-31.

39. Mathivanan S, Periaswamy B, Gandhi T, Kandasamy K, Suresh S, Mohoodoor H, Ramachandra Y, Pandey A (2006) An evaluation of human protein-protein interaction data in the public domain. BMC Bioinformatics 7 Suppl 5: S19.

40. Orchard S, Kerrien S, Jones P, Ceol A, Chatr-Aryamontri A, Salwinski L, Nerothin J, Hermjakob H (2007) Submit your interaction data the IMEx way: a step by step guide to trouble-free deposition. Proteomics 7 Suppl 1: 28-34.

41. Pagel P, Kovac S, Oesterheld M, Brauner B, Dunger-Kaltenbach I, Frishman G, Montrone C, Mark P, Stumpflen V, Mewes HW, Ruepp A, Frishman D (2005) The MIPS mammalian protein-protein interaction database. Bioinformatics 21: 832-834.

42. Parrish JR, Gulyas KD, Finley RL Jr (2006) Yeast two-hybrid contributions to interactome mapping. Curr Opin Biotechnol 17: 387-393.
43. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proceedings of the National Academy of Sciences of the United States of America 96: 4285-4288.

44. Peri S, Navarro JD, Amanchy R, Kristiansen TZ, Jonnalagadda CK, Surendranath V, Niranján V, Muthusamy B, Gandhi TK, Gronborg M, Ibarrola N, Deshpande N, Shanker K, Shivasankar HN, Rashmi BP, Ramya MA, Zhao Z, Chandrika KN, Padma N, Harsha HC, Yatish AJ, Kavitha MP, Menezes M, Choudhury DR, Suresh S, Ghosh N, Saravana R, Chandran S, Krishna S, Joy M, Anand SK, Madavan V, Joseph A, Wong GW, Schiemann WP, Constantinescu SN, Huang L, Kohsrawi-Fari R, Steen H, Tewari M, Ghaffari S, Blobe GC, Garcia JG, Pevsner J, Jensen ON, Roepstorff P, Deshpande KS, Chinnaiyan AM, Hamosh A, Chakravarti A, Pandey A (2003) Development of human protein reference database as an initial platform for approaching systems biology in humans. Genome Res 13: 2363-2371.

50. Schijvenaars BJ, Mons B, Weeber M, Schuemie MJ, van Mulligen EM, Wain HM, Kors JA (2005) Thesaurus-based disambiguation of gene symbols. BMC bioinformatics 6: 149.

54. Stelzl U, Worm U, Lelkner K, Brembeck FH, Goehler H, Stroedicke M, Zetzschke S, Korn B, Birnie M, Lehrbach H, Wanker EE (2005) A human protein-protein interaction network: a resource for annotating the proteome. Cell 122: 957-968.

56. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, Qureshi-Emili A, Li Y, Godwin B, Conover D, Kalbfleisch T, Vijayadamanaga G, Yang M, Johnston M, Fields S, Rothberg JM (2000) A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature 403: 623-627.

58. Xu H, Fan JW, Hripcsak G, Mendonca EA, Markatou M, Friedman C (2007) Gene symbol disambiguation using knowledge-based profiles. Bioinformatics (Oxford, England) 23: 1015-1022.
The tYNA platform for comparative interactomics: a web tool for managing, comparing and mining multiple networks. Bioinformatics 22: 2968-2970.

61. Yu H, Luscombe NM, Lu HX, Zhu X, Xia Y, Han JD, Bertin N, Chung S, Vidal M, Gerstein M (2004) Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs. Genome Res 14: 1107-1118.

Additional Files

Additional file 1. The relational database design of the PIANA database.

Information is kept in four types of tables: 1) biological entity tables (protein and interaction); 2) biological entity identifiers (UniProt entry, gene names, ...); 3) biological entity information (protein features tables and interaction properties). All information is linked to the internal identifier of PIANA: proteinID (an index to all pairs [sequence, tax_id]).
### Database Name

| Database Name | Version of the database | File used in the parsing |
|---------------|-------------------------|--------------------------|
| NCBI taxonomy | Apr 03 2007             | names.dmp                |

### Sequence and cross-reference databases

| Database Name          | Version of the database | File used in the parsing               |
|------------------------|-------------------------|----------------------------------------|
| Uniprot Swissprot      | Apr 03 2007             | uniprot_sprot_rg_rx_OH_deleted.dat     |
| Uniprot Trembl         | Apr 03 2007             | uniprot_trembl_rg_rx_OH_deleted.dat    |
| NCBI GenBank           | Release 158 - April 2007| rel158.fsa_aa                         |
| NCBI Blast nr          | April 2007              | nr                                     |

### Cross-reference databases

| Database Name          | Version of the database | File used in the parsing               |
|------------------------|-------------------------|----------------------------------------|
| ncbi2pdb_pdbaa         | Apr 09 2007             | pdbaa                                  |
| ncbi2sprot             | Apr 09 2007             | swissprot                              |
| pdbsprotec             | 15-Jan-2007             | mapping.txt                            |
| gene2accession         | April 19 2007           | gene2accession                         |
| gene_info              | Apr 19 2007             | gene_info                              |
| refseq                 | Release 22              | RefSeq-release22.catalog               |

### Interaction databases

| Database Name | Version of the database | File used in the parsing                           |
|---------------|-------------------------|-----------------------------------------------------|
| string        | 6.3                     | MySQL database stringDB_v6_3_limited               |
| mint          | 2007-04-05              | 2007-04-05-ful.mitab25.txt                         |
| intact        | 23 april 2007           | intact.txt                                          |
| biogrid       | 2.0.26                  | BIOGRID-ALL-SINGLEFILE-2.0.26.tab.txt              |
| bind          | April 2007              | temporal_mysql_database                           |
| HPRD          | Release_6_01012007      | HPRD_Release_6_01012007.txt                       |
| mips          | March 2007              | temporal_mysql_database                           |
| dip           | 19.2.2007               | mif                                                 |

**Additional file 2.** Repositories used for generating the PIANA database used in this work.

For each external repository used to populate the PIANA database, the version and the file used are shown.
**Additional file 3.** Overlaps between repositories of protein sequences.

The overlap between UniProt Swiss-Prot, UniProt TrEMBL, and NCBI genpept is shown. The percentage over the total number of sequences in PIANA is shown in parenthesis. Two sequences must be identical in order to be considered a positive overlap. A table with overlaps between the NCBI non-redundant database (nr) and the three other databases is also provided.
Additional file 4. Function prediction based on common interaction partners in the interaction network from the Database of Interacting Proteins.

The percentage of shared GO terms is shown for each number of common interaction partners range (Figure 4A). We observed that the accuracy when using a partial subset of interactions is similar to that obtained by using the integrated network of interactions. However, the coverage provided by the partial set of interactions is much lower (Figure 4B).