Available online at www.sciencedirect.com

SciVerse ScienceDirect

Current Opinion in Structural Biology

Towards a detailed atlas of protein–protein interactions

Roberto Mosca[1,5], Tirso Pons[2,5], Arnaud Céol[1,3,5], Alfonso Valencia[2] and Patrick Aloy[1,4]

Protein interaction maps are the key to understand the complex world of biological processes inside the cell. Public protein databases have already catalogued hundreds of thousands of experimentally discovered interactions, and struggle to curate all the existing information dispersed through the literature. However, to be most efficient, standard protocols need to be implemented for direct submission of new interaction sets directly into databases. At the same time, great efforts are invested to expand the coverage of the interaction space and unveil the molecular details of such interactions up to the atomistic level. The net result will be the definition of a detailed atlas spanning the universe of protein interactions to guide the everyday work of the biologist.

Addresses
[1] Joint IRB-BSC Program in Computational Biology, Institute for Research in Biomedicine (IRB Barcelona), c/Baldiri Reixac 10-12, Barcelona 08028, Spain
[2] Structural Biology and Biocomputing Programme, Spanish National Cancer Research Center (CNIO), C/Melchor Fdz Almagro 3, Madrid E-28029, Spain
[3] Center for Genomic Science of IIT@SEMM, Istituto Italiano di Tecnologia (IIT), Via Adamello 16, Milan 20139, Italy
[4] Institució Catalana de Recerca i Estudis Avançats (ICREA), Passeig Lluís Companys 23, Barcelona 08010, Spain

Corresponding authors: Valencia, Alfonso (valencia@cnio.es)
and Aloy, Patrick (patrick.aloy@irbbarcelona.org)

[5] These authors contributed equally to this work.

Current Opinion in Structural Biology 2013, **23**:xx–yy

This review comes from a themed issue on **Protein–protein interactions**

Edited by **Joel Janin** and **Alexandre Bonvin**

http://dx.doi.org/10.1016/j.sbi.2013.07.005

Introduction

Genome-sequencing projects are almost routinely delivering nearly complete lists of the genes and gene products present in many organisms. However, taken individually, these components reveal relatively little about the functioning of most living entities. Complex biological systems are often regulated through the coordination of an intricate network of protein–protein interactions and thus, it is the relationship between molecules what will ultimately determine the behavior of the system. With the advent of high-throughput interactomics, biologists aim at drawing a complete map of the complex universe of interactions that populate the cell. Hundreds of thousands of interactions have been identified and stored in public databases during the last decade and the rate is growing exponentially [1]. Indeed, there are already more than 130 repositories containing protein–protein interaction (PPI) data [2], even if the vast majority of interactions are stored in only a few of them. These resources, each one mapping complementary regions of the interaction space [3], are becoming reference points for collecting and disseminating the current knowledge on protein interactions for several organisms. In a field where the data is produced at a breath-taking pace, and experimental techniques are constantly changing with the aim of providing richer and more detailed information, the effort of keeping up-to-date with the curation is daunting. The task is overwhelming, particularly considering the first estimations of the interactome size for several organisms, including human [4,5], that point out the fact that the interaction space is currently largely unknown. Furthermore, the interactome has an inherently dynamic nature, and varies among tissues, cellular processes and environmental conditions [6]. Thus, in this continuously changing scenario, it is paramount to define and enforce standards on the data formats to facilitate the compilation of new information. However, many thousands of PPIs have already been detected in several organisms and reported and, to be most useful, these interactions need to be processed and catalogued. To embrace this challenge the interaction databases have agreed on a *divide-et-impera* strategy, organizing and distributing the curation of the publications from different journals in order to avoid overlapping, and have adopted curation standards that allow both for accurate tracking of the original experimental information and for seamless integration of different datasets. Nevertheless there is a huge amount of data dispersed through the scientific literature, whose systematic and large-scale retrieval is only possible, at the moment, through the application of algorithms for automatic text-mining.

At the same time the relative low coverage of the tested interaction space has fostered the development of computational methods for the prediction of protein–protein interactions, which in the last few years have produced thousands of novel predictions of a quality comparable to the one of large scale experiments [7,8•].

While the knowledge of the interaction space has extended horizontally, it has been increasingly evident

Please cite this article in press as: Mosca R, et al.: Towards a detailed atlas of protein–protein interactions, Curr Opin Struct Biol (2013), http://dx.doi.org/10.1016/j.sbi.2013.07.005

the need of zooming into the molecular details of these interactions. By complementing the binary information provided by the networks with detailed experimental data on the interacting regions, influencing mutations and post-translational modifications and high-resolution three-dimensional (3D) structures, the information will be most valuable to many different areas of biological research. Mapping of structural data on interaction networks, in particular, has allowed, during the last years, systematic studies on the role of mutations in human genetic diseases and will be essential in next generation drug development strategies [9].

Here we review the current resources that organize the experimental knowledge of the interaction space, we comment on the latest efforts in extending this knowledge both through the prediction of previously unknown interactions and the annotation of their molecular details. We finally discuss possible applications of such a global and high-resolution atlas of protein–protein interactions.

Towards a complete and structured catalogue of PPIs

While sequencing technologies allow the decoding of entire genomes, our knowledge of protein interaction networks (or interactomes) relies on the application of experimental methods that differ in scale and scope, allowing the detection from few to thousands interactions of different types. Fortunately, these heterogeneous data are collected and catalogued in several public databases (Table 1). Overall, these databases contribute to the general completeness of the interaction space [3,10], each one of them covering a different corner, from the generic databases (e.g. MINT [11], DIP [12], BioGRID [13], BIND [14] and Intact [15]), which contain interactions without any restriction of scope, to the specialized ones that focus on particular species or subjects, such as MPIDB [16] that stores only microbial interactions or MatrixDB [17] containing only interactions occurring in the extracellular matrix. Integration of the different databases is therefore paramount to have the most complete vision of the interaction space. In order to facilitate this, and to avoid the huge discrepancies observed in the curation of the data [18], MINT, DIP and Intact, soon followed by BioGRID, MPIDB and MatrixDb, and more recently by I2D [19] and Molecular Connections (http://www.molecularconnections.com/home/en/home/products/netPro/), have created the IMEx consortium [20], agreeing on common curation rules and distributing the different scientific journals among the databases, thus insuring that every publication will be assigned to a single database [21,22]. HPRD [23], although not following community standards, remains one of the largest repositories for human interactions and BIND, which is no longer active, contains a valuable piece of data that has not yet been incorporated into the other repositories (Figure 1).

In addition, the Protein Standard Initiative-Molecular Interaction (PSI-MI) consortium [21,22] has defined a controlled vocabulary that consents a rich annotation of the experimental details allowing, for example, to distinguish between 'association' which 'indicates that the interaction is from an experimental method that identifies a loose co-complex' and 'direct interaction' which 'indicates that the two molecules are known to be in actual physical contact with each other' [20]. A recurrent mistake in the usage of PPI databases is the wrong assumption that they only report binary interactions. Other types of interactions are possible, like the physical association in groups of co-purified proteins in Affinity Purification followed by Mass Spectrometry (AP-MS) experiments. Often these groups comprise three or more proteins, and only for simplicity they are expanded to tabular form by exploding the groups into binary interactions between the bait and all its preys (SPOKE model). The strict rules of curation imposed by the IMEx standard allow the reconstruction of these complexes and avoid the binary ambiguity. The definition of the PSI-MI standard vocabulary has allowed to incorporate detailed descriptions of physical interactions inside pathways, which has been compiled in the BioPAX standard. BioPAX (Biological Pathway Exchange) is a standard language that aims to enable integration, exchange, visualization and analysis of biological pathway data. Its last release (Level 3) in 2010 [24] includes metabolism, signaling, gene regulation, genetic interactions and molecular interactions and covers most data available from molecular interaction and pathway databases. BioPAX Level 4 will incorporate cell type information, kinetics using SBML concepts and pathway graphical layout using SBGN concepts, increasing the interaction with data communities (G Bader, personal communication).

IMEx standards represent an important step forward in the correct cataloguing of PPIs within the different repositories. However, it is worth noting that all the entries curated before joining the consortium, as well as the vast amount of data in databases not complying with the standard, are not necessarily granted to have the same level of quality. This makes it difficult, for example, to distinguish binary interactions from other types of interactions in databases that automatically expand complexes according to a SPOKE or MATRIX (i.e. all against all) model, requiring to resort to the application of additional filters based on the annotated detection method, like in [25]. However, the general applicability of such strategy is questionable since, for instance, the methodological information is missing for up to 75% of the interactions contained in HPRD, one of the richest repositories of human interactions.

Adherence to the standards would also allow the integration of the data as inferred directly by the authors of the high-throughput experiments, using computational and

Please cite this article in press as: Mosca R, et al.: Towards a detailed atlas of protein–protein interactions, Curr Opin Struct Biol (2013), http://dx.doi.org/10.1016/j.sbi.2013.07.005