Chapter VI
Network Cleansing:
Reliable Interaction Networks

Paolo Marcatili
Sapienza University, Italy

Anna Tramontano
Sapienza University, Istituto Pasteur Fondazione Cenci Bolognetti, Italy

ABSTRACT

This chapter provides an overview of the current computational methods for PPI network cleansing. The authors first present the issue of identifying reliable PPIs from noisy and incomplete experimental data. Next, they address the questions of which are the expected results of the different experimental studies, of what can be defined as true interactions, of which kind of data are to be integrated in assigning reliability levels to PPIs and which gold standard should the authors use in training and testing PPI filtering methods. Finally, Marcatili and Tramontano describe the state of the art in the field, presenting the different classes of algorithms and comparing their results. The aim of the chapter is to guide the reader in the choice of the most convenient methods, experiments and integrative data and to underline the most common biases and errors to obtain a portrait of PINs which is not only reliable but as well able to correctly retrieve the biological information contained in such data.

INTRODUCTION

Interactions play a key role in every cellular activity and function. They are often related to diseases and therefore attract much interest in drug development projects. In the last few years the development of new experimental techniques for protein-protein interaction (PPI) detection together with the wide interest in systemic descriptions of biological processes gave birth to the field of *interactomics* aimed at studying the whole set of molecular interactions in a cell.

As in other -omics sciences, one, and perhaps the most important, issue to be faced concerns the great amount of noisy and unreliable information present in the data. Generally speaking, the data, mostly gathered with high-throughput experiments (Gavin et al., 2006; Gavin et al., 2002; Giot et al., 2003; Ho et al., 2002; Ito et al., 2001; Krogan et al., 2006; Li et al., 2004; Rual et al., 2005; Stelzl et al., 2005; Uetz et al., 2000), contain a sizeable number of false-positives and their reproducibility is not very satisfactory. Several authors have compared the results of

the experimental data collected in high-throughput experiments (G. D. Bader & Hogue, 2002; J. S. Bader, Chaudhuri, Rothberg, & Chant, 2004; Gentleman & Huber, 2007; Hart, Ramani, & Marcotte, 2006; Mrowka, Patzak, & Herzel, 2001; Sprinzak, Sattath, & Margalit, 2003; von Mering et al., 2002) and used the overlap between the detected interactions to infer the accuracy and coverage of the complete interactome of various organisms. The results of such comparisons, even if with some differences, seem to depict an extremely problematic situation, with accuracy values ranging from 10% to 40% of true positive interactions and a limited coverage (J. S. Bader, Chaudhuri, Rothberg, & Chant, 2004; Gentleman & Huber, 2007; Sprinzak, Sattath, & Margalit, 2003). In the last few years, thanks to the introduction of more reliable experimental techniques, of new filtering methods and of a somehow deeper understanding of the interactome such pessimistic scenario has been partially revised, but many of the original problems still remain. Despite the many efforts devoted to the analysis of the possible sources of noise in PPI maps and the continuous development of methods aimed at overcoming the problem, a satisfactory solution is still missing. As it has been shown in other -omics contexts, a coordinated effort at both the computational and experimental level is essential for allowing the data to be properly exploited.

Central to the understanding and the analysis of interaction maps is the definition of the interaction itself. Given the continue and fast development of paradigms and methods in interactomics, such definition is not straightforward and, in many cases, strongly dependent on the current scientific expectations and experimental limits.

The binding energy of a physical interaction can vary by orders of magnitude and involve large surface patches of the partners or just small regions; the complex itself can be transient or stable and/or obligatory; the interaction can be essential for the correct three-dimensional folding of both interacting partners and it may occur at specific times and locations in the life of a cell (Nooren & Thornton, 2003). Moreover, the term *interaction* is often used with rather different meaning: for example, we talk of functional or genetic interactions when the behavior of a protein or a gene is dependent upon the effect of another (Boone, Bussey, & Andrews, 2007).

When analyzing PPIs, all these different aspects should in principle be taken into account. This might be very difficult and, in some cases, impossible.

It follows that, in order to understand which information we can reliably extract from a PPI map, we need to define what we mean by interaction. Far from being only a semantic or ontological problem, this is a central issue that affects several of the fundamental questions that we might want to ask, some of which are:

- what do we expect to find as a result of different *experimental studies*;
- what do we define as a *true interaction*;
- which kind of data we need to *integrate* when we want to assign reliability levels to single PPIs or to whole interaction maps;
- how do we assess the results, i.e. which *gold standard* can we use.

This chapter will try and give an overview of the present situation with respect to these issues by first giving a brief definition of protein-protein interaction and next addressing each of the other questions in turn. We will focus on the possible biases, errors and critical choices in PPIs cleansing. Finally, we will describe the state of the art in the field, presenting the different classes of algorithms and comparing their results.

EXPERIMENTAL METHODS FOR PPI DETECTION

Most common experimental methods for PPI detection can be divided on the basis of two main characteristics: low vs high throughput studies and affinity purification-mass spectrometry (AP-MS) (Aebersold & Mann, 2003) versus yeast-two-hybrid (Y2H) technology (Shoemaker & Panchenko, 2007). The pros and cons of each of these methods are well known. For our purposes, it is useful to recall just a few concepts.

Y2H is an *in vivo* technique (even if it makes use of chimeric proteins different from the *wild type* ones) and it does not heavily rely on protein abundance. It is able to detect weak or transient PPIs, but it often fails to detect large complexes or obligate interactions. This method can introduce several artifacts in the data (auto-activators, membrane proteins) (Huang, Jedynak, & Bader, 2007; Vidalain, Boxem, Ge, Li, & Vidal, 2004).