Chapter X
Evolutionary Analyses of Protein Interaction Networks

Takashi Makino
University of Dublin, Trinity College, Ireland

Aoife McLysaght
University of Dublin, Trinity College, Ireland

ABSTRACT

This chapter introduces evolutionary analyses of protein interaction networks and of proteins as components of the networks. The authors show relationships between proteins in the networks and their evolutionary rates. For understanding protein-protein interaction (PPI) divergence, duplicated genes are often compared because they are derived from a common ancestral gene. In order to reveal evolutionary mechanisms acting on the interactome it is necessary to compare PPIs across species. Investigation of co-localization of interacting genes in a genome shows that PPIs have an important role in the maintenance of a physical link between neighboring genes. The purpose of this chapter is to introduce methodologies for analyzing PPI data and to describe molecular evolution and comparative genomics insights gained from such studies.

INTRODUCTION

Protein-protein interactions (PPIs) are one of the most important components of biological networks. An understanding of the evolution of PPIs is crucial to elucidating how the evolution of biological networks has contributed to diversification of extant organisms. The amount of information about PPIs has grown rapidly due to the development of a high-throughput two-hybrid system, mass spectrometry of co-immunoprecipitated protein complexes, and bioinformatics approaches such as text mining from the many individual studies reported in the literature. The extensive data allow us to analyze the protein interaction networks from an evolutionary aspect.

Evolutionary studies of protein interaction networks can be classified into at least five topics based on their focus (Fig. 1). The most studied topic is the relationship between the number of PPIs of a protein (**connectivity**) and protein evolution (Fig. 1A). It has been shown that protein **connectivity** correlates with evolutionary rates (Fraser *et al.*, 2002; Fraser *et al.*, 2003). Protein **connectivity** can be directly calculated from the protein interac-

tion network. To estimate evolutionary rates of proteins, we must first make multiple sequence alignments by using an application such as CLUSTAL W (Thompson *et al.*, 1994). The number of amino acid substitutions is estimated from the number of differences between the aligned sequences with a biologically realistic statistical model, e.g. Kimura's method. Many of the most widely used methods are implemented in the PHYLIP software package (http://evolution.genetics.washington.edu/phylip.html). The evolutionary distances can be translated into relative evolutionary rates by comparison with an older homologous sequence (an outgroup). When evolutionary rates of proteins are slow, the proteins are conserved even in distantly related species. In general, **orthologous** proteins between two species are identified as a pair of sequences that show reciprocal best hits in the sequence similarity search using all protein sequences from both species. In other words, a reciprocal best hit is evidence that the protein pair is related through a speciation event. BLAST (http://www.ncbi.nlm.nih.gov/blast/Blast.cgi) is the most commonly used sequence similarity search software. The very fact of easily detectable **orthologs** between distantly related species implies strong functional constraint and slow evolutionary rates without explicit rate estimation. The ability to detect **orthologs** in distantly-related genomes is one measure of conservation; the more distantly related the species where an **ortholog** is detectable the greater the conservation of the protein.

Not only the **connectivity** but also the structure of the protein interaction networks influence functional constraint on proteins (Fig. 1B). There are several indicators to describe **network structure** such as motif constituents, clustering coefficients, betweenness, centrality and so on. Wuchty and his colleagues perform the first survey of the PPI patterns between groups of two, three, four and five proteins in the yeast protein interaction network (defined as motif constituents) and examined their evolution (Wuchty *et al.*, 2003). When we focus on two proteins in a protein interaction network there is only one possible interaction. When there are three proteins, there are two possible interaction patterns. The most cohesive motif in the definition consists of five proteins and 10 PPIs (i.e. all proteins interacting with all others). Alternatively, clustering of proteins in the interaction network may be simplified into a classification of dense parts and sparse parts of the networks (Makino & Gojobori, 2006). Sparse and dense parts are defined using the clustering coefficients (Watts & Strogatz, 1998). Note that the clustering coefficients differ from motif constituents. The more PPI partners a protein has, the greater the chance the protein has a cohesive motif. On the other hand, the more PPI partners a protein has, the smaller the chance the protein has a high clustering coefficient because the number of possible interactions between those PPI partners increases.

Gene pairs with highly similar sequences in an organism's genome must be derived from a gene duplication event. Duplicated pairs are defined as pairs of sequences that 'hit' each other in a similarity search (e.g. BLAST as before) using protein sequences from a single organism. To understand the evolution of protein interaction networks it is useful to study the PPIs of **duplicated genes**. This is because immediately after duplication these genes will produce identical proteins, which, by definition, will behave identically in the protein interaction network (Fig. 1C). In fact, **duplicated gene** pairs often share PPI partners because they have a common ancestor (Deane *et al.*, 2002; Wagner, 2001). We can classify duplicated pairs into two groups depending on the absence or presence of common PPI partners as diverged or non-diverged pairs, respectively.

Most reports on the evolution of protein interaction networks deal with yeast protein interaction networks because yeast PPI data have been most abundant (Gavin *et al.*, 2006; Gavin *et al.*, 2002; Ho *et al.*, 2002; Ito *et al.*, 2000; Krogan *et al.*, 2006; Tarassov *et al.*, 2008; Uetz *et al.*, 2000). However recent global studies on PPIs have been performed in the prokaryotes *Helicobacter pylori* (Rain *et al.*, 2001) and *Escherichia coli* (Butland *et al.*, 2005), and also eukaryotes *Plasmodium falciparum* (LaCount *et al.*, 2005), *Caenorhabditis elegans* (Li *et al.*, 2004), *Drosophila melanogaster* (Formstecher *et al.*, 2005; Giot *et al.*, 2003) and Human (Rual *et al.*, 2005; Stelzl *et al.*, 2005). These comprehensive PPI data of multi species allow us to compare PPIs between species; evolutionary conserved interactions are known as **interologs** (Walhout *et al.*, 2000; Fig. 1D).

A functional **gene cluster** is defined as a gene pair having both functional and physical links on a genome (Fig. 1E). The type of functional link that is considered differs between studies and includes PPIs, interactions between subunits in a protein complex, and gene expression profiles (Hurst *et al.*, 2004; Makino & McLysaght, 2008; Poyatos & Hurst, 2006; Teichmann & Veitia, 2004). When functional links are mapped onto a genome one can observe that some genes have functional links to neighboring (i.e., genetically linked) genes in the genome.

Before identifying the **gene clusters**, we have to exclude tandem **duplicated genes** from the analysis, because **duplicated gene** pairs are often arrayed tandemly on a genome due to the mechanism of duplication (replication