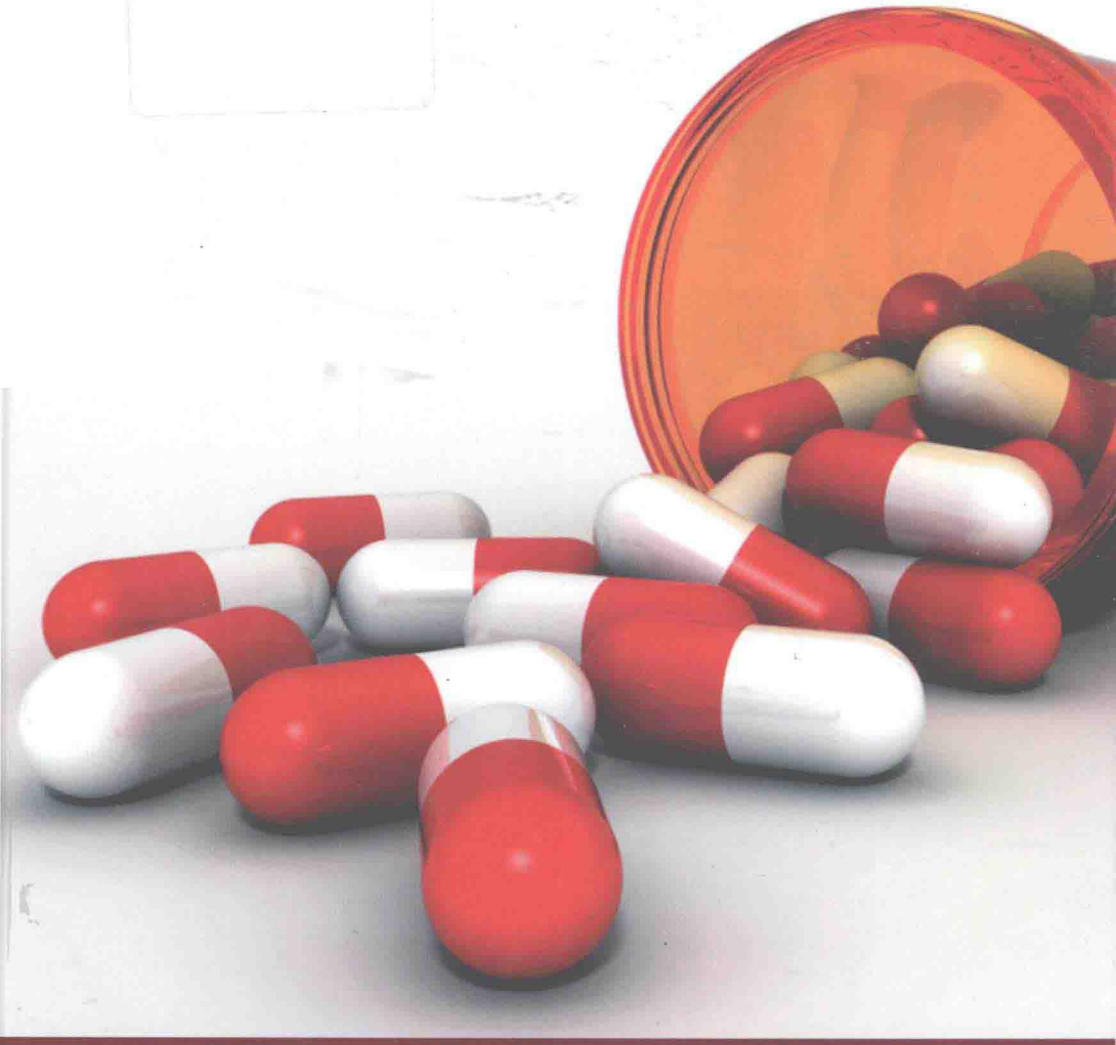

# Drug Design and Medicinal Chemistry

Erica Helmer

# Drug Design and Medicinal Chemistry

Edited by **Erica Helmer**



CALLISTO REFERENCE

New York

Published by Callisto Reference,
106 Park Avenue, Suite 200,
New York, NY 10016, USA
www.callistoreference.com

**Drug Design and Medicinal Chemistry**
Edited by Erica Helmer



International Standard Book Number: 978-1-63239-154-4 (Hardback)



Printed in China.

# Drug Design and Medicinal Chemistry

# Preface

It is often said that books are a boon to mankind. They document every progress and pass on the knowledge from one generation to the other. They play a crucial role in our lives. Thus I was both excited and nervous while editing this book. I was pleased by the thought of being able to make a mark but I was also nervous to do it right because the future of students depends upon it. Hence, I took a few months to research further into the discipline, revise my knowledge and also explore some more aspects. Post this process, I begun with the editing of this book.

This book covers the different aspects of drug design and medicinal chemistry. Recently, medicinal chemistry has become accountable for clarifying interactions of chemical molecules procedures, such that many experts in life sciences, from agronomy to medication, are occupied in medicinal study. This book comprises of researches centering on molecular features of drug metabolism, pro-drug production, in silico and chemical compounds used in applicable methods. It even deals with fundamental issues and developments in medicinal chemistry and drug design. Particular significance is given to both conjectural and investigational features of contemporary drug design. This book intends to provide some useful knowledge to students and even experts working on the above stated topic. This book is a compilation of data provided by some of the renowned experts working in this field of science for years.

I thank my publisher with all my heart for considering me worthy of this unparalleled opportunity and for showing unwavering faith in my skills. I would also like to thank the editorial team who worked closely with me at every step and contributed immensely towards the successful completion of this book. Last but not the least, I wish to thank my friends and colleagues for their support.

**Editor**

# Contents

1

# Analysis of Protein Interaction Networks to Prioritize Drug Targets of Neglected-Diseases Pathogens

Aldo Segura-Cabrera[1,5], Carlos A. García-Pérez[1], Mario A. Rodríguez-Pérez[2], Xianwu Guo[2], Gildardo Rivera[3] and Virgilio Bocanegra-García[4]
[1]*Laboratorio de Bioinformática*
[2]*Laboratorio de Biomedicina Molecular*
[3]*Laboratorio de Biotecnología Ambiental*
[4] *Laboratorio de Medicina de Conservación*
*Centro de Biotecnología Genómica, Instituto Politécnico Nacional*
[5]*U.A.M. Reynosa Aztlán, Universidad Autónoma de Tamaulipas, Reynosa*
*México*

## 1. Introduction

Many technological, social and biological systems have been modeled in terms of large networks providing invaluable insight in the understanding of such systems. Systems biology is an emerging and multi-disciplinary discipline that studies the interactions of cellular components by treating them as part of an integrated system. Thus, systems biology has shown that functional molecules are involved in complex networks of inter-relationships, and that most of the cellular processes depend on functional modules rather than isolated components. Large amounts of biological network data of different types are available, e.g., protein-protein interaction, transcriptional regulatory, signal transduction, and metabolic networks. Since proteins carry out most biological processes, the protein interaction networks (PINs) are of particular importance. The advancement of the functional genomics and systems biology of model organisms such as *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, and *Drosophila melanogaster* has contributed to the development of experimental and computational methods, and also to the understanding of human complex diseases. The availability of these methods has facilitated systematic efforts at creating large-scale data sets of protein interactions, which are modeled as PINs.

Usually, a PIN is represented as a graph where the proteins are the nodes and the interactions are the edges. According to the complex network theory, PINs are scale-free networks characterized by a power-law degree distribution. In scale-free networks, most nodes have a small number of links between them; whereas, a small percentage of nodes interact with a disproportionately large number of others. The nodes with a large number of links in PINs are called hub proteins. Functional genomics studies showed that in PINs, the deletion of a hub protein is lethal to the organism, a phenomenon known as the centrality-

lethality rule. This rule is widely believed to reflect the special importance of hubs in organizing the network, which in turn suggests the biological significance of network topology. Several well-known studied proteins that are implicated in human diseases are hub proteins. Examples include p53, p21, p27, BRCA1, ubiquitin, calmodulin, and others which play central roles in various cellular mechanisms.

Despite recent advances in systems biology of model organisms, the systems biology of human pathogenic organisms such as those that cause the so-called "neglected-diseases" has not received much attention. Neglected-diseases are chronic or related disabling infections affecting more than 1 billion people worldwide, mainly in Africa. Pathogens of neglected-diseases include: Protozoan parasites (e.g., *Leishmania* spp., *Plasmodium* spp., and *Trypanosoma* spp.), vector-borne helminthes (e.g., *Schistosoma* spp., *Brugia malayi*, and *Onchocerca volvulus*), soil-transmitted helminthes (e.g., *Ascaris lumbricoides* and *Trichuris trichura*), bacteria (e.g., *Mycobacterium tuberculosis* and *M. leprae*), and viruses (e.g., dengue and yellow fever virus). A number of factors limit the utility of existing drugs in neglected-diseases such as high cost, poor compliance, drug resistance, low efficacy, and poor safety. Since the evolution of drug resistance is likely to compromise every drug over time, the demand for new drugs and targets is continuous. The drug target identification is the first step in the drug discovery flow-through process. This step is complicated because a drug target must satisfy a variety of criteria. The important factors in this context are mainly related to the toxicity to host, and the essentiality of the target to the pathogen's physiology for growth and survival. Thus, the topological and functional analysis of neglected-disease pathogen PINs offers a potentially effective strategy for identifying and prioritizing new drug targets.

This chapter will introduce the reader to the basic concepts of network analyses and outline why it is important in terms of predicting protein function and essentiality. Work involving PINs of neglected-disease pathogens will be explained so that the reader will understand the current state in terms of its application to prioritize drug targets. The experimental and computational methods most likely to be used to identify and predict PINs, and the strategies for identifying multiple potential drug targets in neglected-disease pathogens will be also outlined using several biological databases in an integrated way.

To achieve this goal, the chapter includes three sections. Firstly, we present an outline of the conceptual development of network biology. The applied functional genomics involving the analysis of PINs of model organisms has led to developing methods and principles for elucidating protein function. We will also explain how these concepts are connected with protein essentiality to identify their "weak" points on the PINs of neglected-disease pathogens and its use for prioritizing drug targets. In the second section, we outline the experimental and computational methods that are most extensively to be used to identify and predict PINs. Some new approaches for predicting PINs are also introduced. These include the probabilistic integrated network methods which have shown the capability to increase the accuracy and coverage of the PINs. These primary research articles will be reviewed and the potential applications for the future be explained. This section mainly focused on analyzing the PINs of most prevalent neglected-disease pathogens in which the use of drugs is often limited by factors including high cost, low efficacy, toxicity, and the emergence of drug resistance. The potential use as an integrated strategy aimed at prioritizing and identifying drug targets of neglected-disease pathogens will be put forward, and the argument for future research involving the application of many tools and strategies will be discussed. In the final section,

we describe, amenably, the basic criteria to select pathogen drug targets, and the PINs of neglected-disease pathogens will be described in such a manner that the chapter will work as a source of key literature references for students and researchers. Papers will be reviewed to describe these basic principles, using key publications containing data and quantitative analyses (models, figures, tables) for PINs of some neglected-disease pathogens. We will describe novel lines of research; pros and cons of the use of PINs for prioritizing and identifying drug targets of neglected-disease pathogens.

## 2. Systems and network biology: Basic concepts

Systems biology is a holistic approach that involves the study of the inter-relationships of all the different elements in a biological system in order to understand non-deterministic behaviors that emerge from interaction between the cellular components and their environment and not by studying them in an isolated manner, one at a time (Hood and Perlmutter 2004, Weston and Hood 2004, Kohl and Noble 2009). Thus, the cell's behavior can be understood as a consequence of the complex interactions between its numerous constituents such as DNA, RNA, proteins, and metabolites. These interactions are also responsible for performing processes critical to cellular survival. For example, during transcription process the regulatory proteins can activate or inhibit the expression of genes or regulate each other as part of gene regulatory networks. Likewise, the cellular metabolism can be integrated into a metabolic network whose fluxes are regulated by enzymes. Similarly, the PINs represent how the proteins work together through interactions that lead to the modification of protein functions or new roles in protein complexes.

The biological systems consisting of interacting cellular components have led to the use of graph theory and mathematical tools based on graphs where the individual components are represented by nodes and the interactions by links (Fig. 1). Albert and Barabási (2002) have shown the general properties found among several networks ranging from the Internet to social and biological networks (Albert and Barabási 2002). The analysis of topology of those networks showed that they deviate substantially from randomly built networks as studied by Erdös and Rényi (Fig. 1a) (Erdös and Rényi 1960). Also, these networks did not show a well-shaped frequency distribution of the number of links per node as expected from randomly formed networks; instead, they showed a power-law distribution, which is characteristic of scale-free networks (Fig. 1b and 1c) (Amaral *et al.*, 2000, Albert 2005).

In scale-free network, the majority of nodes have only a few links, whereas very few nodes have a large number of links. Those nodes are called hubs and they represent the most vulnerable points of a network (Barabasi and Albert 1999, Albert *et al.*, 2000, Jeong *et al.*, 2001, Yu *et al.*, 2004a, Tew *et al.*, 2007). The topological features of networks can be quantified by measuring topological parameters whose information content provides a description from local (e.g., single nodes or links) to network-wide level (e.g., connections and relationships between nodes). For example, the nodes of a graph can be characterized by means of the number of links they have (the number of other nodes to which they are connected). This parameter is called "node degree". In directed networks, it is possible to distinguish the number of directed links that points toward the node (in-degree), and the number of directed edges that points outward the node (out-degree). The node degree characterizes individual nodes; however, in order to relate this parameter to whole network, a network degree distribution can be defined. The degree distribution P(k) represents the

fraction of nodes that have degree k and it is obtained by counting the number of nodes N(k) that have k = 1, 2... links and dividing it by the total number of nodes *N*. The degree distributions of numerous networks such as the Internet, social, and biological networks, follow a power law (Fig. 1b and 1c) which is defined by the functional equation $P(k) \sim k^{\gamma}$, where γ represents the degree exponent, taking usually values in the range between $2<\gamma<3$ (Barabasi and Oltvai 2004). This function is intimately linked to the growth of the network in which new nodes are preferentially attached to already established nodes, a property that is also thought to characterize the evolution of biological systems (Jeong *et al.*, 2000).

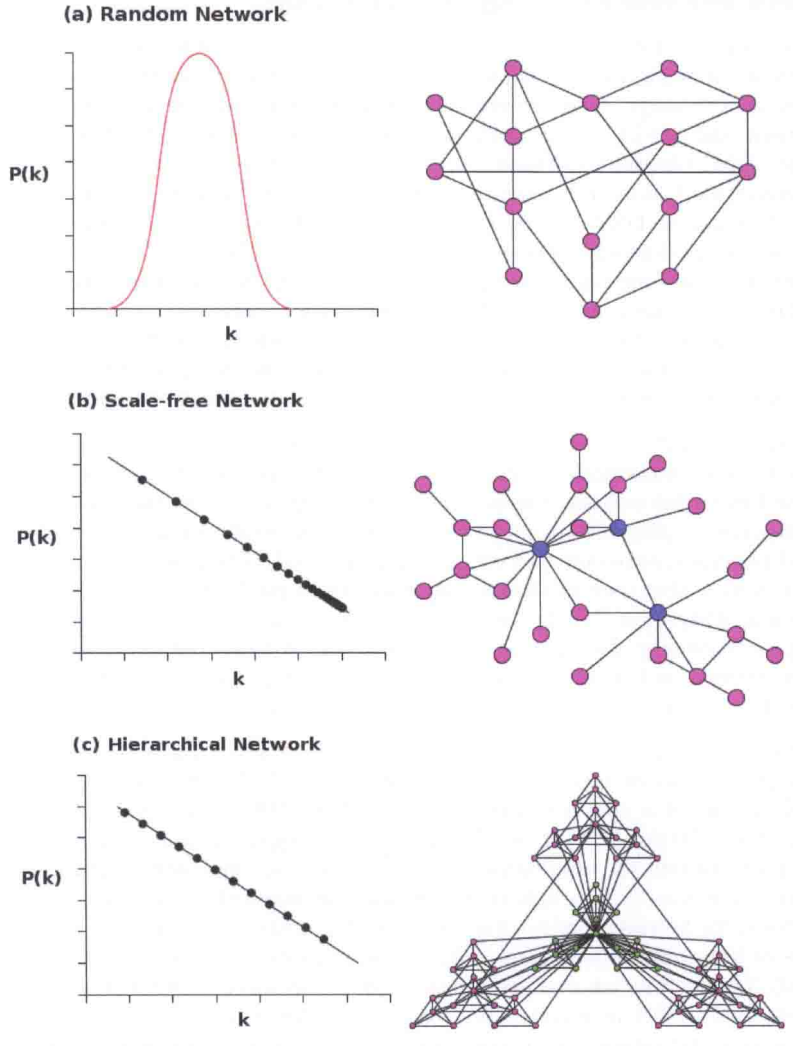


Fig. 1. Three types of network models and their associated distributions: (a) random network, (b) scale-free network, and (c) hierarchical network.

The distance between any two nodes in a network could be defined by the path length. In other words, it represents how many links we need to pass between two nodes. Nevertheless, it could have many alternative paths between two nodes in a network. The path with the smallest number of links between the selected nodes (shortest path) is of special interest. A common characteristic of several biological networks, including metabolic networks (Jeong *et al.*, 2000, Wagner and Fell 2001) and PINs (Giot *et al.*, 2003, Yook *et al.*, 2004) is that any two nodes can be connected with a path of a few links only. The main biological implications of this characteristic are related to: i) how the biological networks are capable of rapid responses to perturbations; ii) its capacity to employ alternative roads for the same input and output; and iii) the ability to efficiently compensate the perturbations in essential pathways.

Another important issue derived from network analysis is the concept of modularity, which can be used to describe how a group of physically or functionally linked nodes work together to achieve a particular function. The topological parameter used to quantify the modularity in a network is the clustering coefficient *Ci*, which represents the ratio between the number of links connecting nodes adjacent to node *i* and the total possible number of links among them (Watts and Strogatz 1998). It is worth noting that in first instance, the modularity concept might be in contradiction of the scale-free nature of the networks because the presence of modules implies that there are clusters of nodes that are relatively isolated from the rest of the network. However, it has been demonstrated that modularity and scale-free properties naturally co-occur in biological networks indicating that modules are not independent, instead, they are combined to form a hierarchical network (Fig. 1c) (Ravasz *et al.*, 2002).

Biological networks, including PINs and metabolic networks are good examples of network modularity because they exhibit high average *Ci*, which are associated to a high level of network robustness (Alon *et al.*, 1999, Ravasz *et al.*, 2002, Barabasi and Oltvai 2004). The most common representation of a module or cluster in a network is as a highly interconnected group of nodes. The biological implication of the modularity concept is that the nodes that integrate a module tend to participate in related biological processes and pathways; for example, protein and nucleic-acid synthesis, protein degradation, signal transduction, and metabolic pathways (Ma'ayan *et al.*, 2005). The analysis of experimental PINs have shown to have a remarkably modularity character (Giot *et al.*, 2003, Yook *et al.*, 2004). These findings in experimental PIN maps have been used to improve the understanding of the pleiotropic effects, and how perturbations on genes or proteins can propagate through the network and produce, in appearance, unrelated or extensive effects.

In addition to the modules, within a network, small and recurring sub-graphs, known as interaction motifs, with well-defined topologies can be identified (Fig. 2). The frequency analysis of these interaction motifs in networks revealed that they are over-represented when compared to a randomized version of the same network, suggesting that not all sub-graphs are equally significant in networks and that interaction motifs form functionally separable building blocks of cellular networks (Mangan and Alon 2003, Wuchty *et al.*, 2003, Alon 2007). For example, triangle motifs, also called feed-forward loops in directed networks, appear in both transcription-regulatory and neural networks. Likewise, there is evidence suggesting that specific motif type aggregates to form large motif clusters and that also appear to be commonly involved with certain functional roles (Milo *et al.*, 2002, Shen-

Orr *et al.*, 2002, Wuchty *et al.*, 2003). For example, in the *E. coli* transcription regulatory network, most motifs overlap, in which the specific motifs are no longer clearly separable (Shen-Orr *et al.*, 2002).

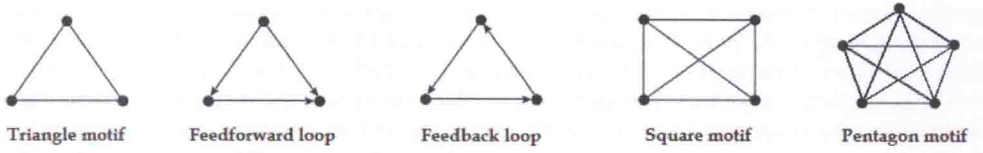


Fig. 2. Some types of interaction motifs found in biological networks.

The relevance of any node in mediating the communications flow among other nodes in the network is quantified by its betweenness centrality, which is defined as the total number of non-redundant shortest paths going through a certain node or edge (Freeman 1977). Girvan and Newman (2002), have proposed that the edges with high betweenness are the ones that are "between" network clusters; therefore, the information flow within a network could be altered by removing these edges (Girvan and Newman 2002). Dunn *et al.*, (2005) using an edge betweenness based-method have shown that clusters in PINs tend to share similar functions (Dunn *et al.*, 2005). Moreover, Yu *et al.*, (2007) have reconsidered the classical meaning of betweenness as a measure of the centrality of the nodes in a PIN. They have defined those nodes as "bottlenecks" with the highest betweenness centrality and find that bottlenecks nodes have a higher probability to be essential (Yu *et al.*, 2007).

It is worth noting that the topological parameters might be combined between them or with additional information of functional annotations regarding the network nodes (genes or proteins). Thus, a network provides testable predictions ranging from single interactions to essential genes and functional modules (del Rio *et al.*, 2009). Likewise, the functions of un-annotated genes or proteins can be also predicted on the basis of the annotation of their interacting partners. This approach to predict the protein/gene function is known as "guilty by association". Additionally, the integration of information related to diseases or specific phenotypes with network approaches also enhances the understanding of human diseases, pharmacology response, and phenotype prediction (Ideker and Sharan 2008, Lee *et al.*, 2008a, Lee *et al.*, 2010, Wang and Marcotte 2010, Lee *et al.*, 2011).

## 3. Methods to identify protein interactions networks (PINs)

### 3.1 Experimental methods

In the postgenomic era, the accumulation of protein-protein interaction data has enabled the biology systems studies at PINs levels (von Mering *et al.*, 2002). However, PIN analysis requires methods amenable to high throughput (HT) screening, such as large-scale versions of techniques like yeast two hybrid (Y2H) and tandem affinity purification coupled to mass spectrometry (TAP-MS) for performing systematic screens (Ito *et al.*, 2001a, Cusick *et al.*, 2005). In addition, there are a wide variety of methods to detect, analyze, and quantify protein interactions, including surface plasmon resonance spectroscopy, nuclear magnetic resonance (NMR), x-ray crystallography, and fluorescence-based technologies. These techniques provide detailed information on physical properties of protein interactions.

These methods are of paramount usefulness; however, herein, the techniques that can be applied to determine protein-protein interactions, at large-scale level, will be highlighted. In particular, the outcomes of Y2H system and TAP-MS are used further to perform *in silico* global network analysis. Both techniques were intensively applied to map the PIN of yeast, the first model organism with available PINs (Uetz *et al.*, 2000, Ito *et al.*, 2001b, Gavin *et al.*, 2002, Ho *et al.*, 2002, Ito *et al.*, 2002, Tong *et al.*, 2004, Yu *et al.*, 2008). Afterwards, large-scale efforts have been made to determine PINs for other model minor eukaryotic organisms: *D. melanogaster* (Giot *et al.*, 2003), and *C. elegans* (Li *et al.*, 2004); pathogenic microorganisms: *Helicobacter pylori, Campylobacter jejuni, Treponema pallidum, M. tuberculosis* (Wang et al., 2010), herpes simplex virus 1 (Lee *et al.*, 2008b), and Kaposi's sarcoma-associated herpesvirus (Uetz *et al.*, 2006, Rozen *et al.*, 2008), and major eukaryotic organisms: *Arabidopsis thaliana* (de Folter *et al.*, 2005) and humans (Rual *et al.*, 2005, Stelzl *et al.*, 2005, Gandhi *et al.*, 2006). Even though the PINs are not completed, the available PINs provide insight into how particular properties of proteins are integrated at systems level, and also, as a useful resource to predict the functional role of genes or proteins.

### 3.1.2 Yeast two-hybrid (Y2H) system

The Y2H system has considerably accelerated the *in vivo* large-scale screening of protein interactions enabling the detection of physically interacting proteins by using the modular organization of eukaryotic transcriptional activators. The eukaryotic transcription activators are formed by at least two distinct domains, one responsible of binding to a DNA region (BD) promoter and the other of activating the transcriptional processes (AD). It is well-known that splitting BD and AD domains will inactivate the transcriptional processes, but the transcription can be restored if a BD domain is re-associated with an AD domain (Fields and Song 1989). Thus, the standard Y2H system includes a DB domain fused to the "bait" protein-coding region and an AD domain fused to the "prey" protein-coding region. When DB-bait and AD-prey domains are co-expressed in the nucleus of yeast cells, "bait"-"prey" domain interaction reconstitutes a functional transcription factor that activates the transcription of one reporter gene (Fig. 3). The most used Y2H system is based on GAL4/LexA, where the GAL4 protein controls the expression of the LacZ gene encoding beta-galactosidase.

The main advantages of Y2H system are: i) the DNA ( not the protein) is manipulated to study both bait and prey proteins (Walhout and Vidal 2001a); ii) it allows to identify protein interactions *in vivo*; iii) to identify transitory protein interactions, and iv) it is amenable to high-throughput screening methods (Buckholz *et al.*, 1999, Uetz and Hughes 2000, Walhout and Vidal 2001b, Ito *et al.*, 2002, Rual *et al.*, 2005).

The drawbacks include: i) a high proportion of false-positives and negatives (Vidal and Legrain 1999, Ito *et al.*, 2002); ii) it forces sub-cellular localization of bait and prey in the yeast nucleus which might preclude certain interactions from taking place (Cusick *et al.*, 2005). For example, membrane protein interactions cannot be identified by standard Y2H system because the AD-prey fusion will be retained at the membrane, thus, avoiding the reconstitution of a functional transcription factor (Xia *et al.*, 2006); iii) the over-expression of tested proteins, thus modifying the relative concentrations of potential interaction partners in comparison to the *in vivo* state; iv) the presence of auto-activators, i.e. proteins initiating

transcription by themselves (Cusick *et al.*, 2005), and v) the differences in post-translational modifications and protein folding processes between yeasts and other organisms (Shoemaker and Panchenko 2007). Given these cons, several modifications have been made to improve the quality of the Y2H system results, including the development of membrane Y2H, the inclusion of different promoters of reporter genes, the use of low copy vectors, and the reduction of auto-activators. Once that these drawbacks are reduced, the quality of the Y2H system is significantly improved (Lehner *et al.*, 2004, Li *et al.*, 2004, Rual *et al.*, 2005, Yu *et al.*, 2008).

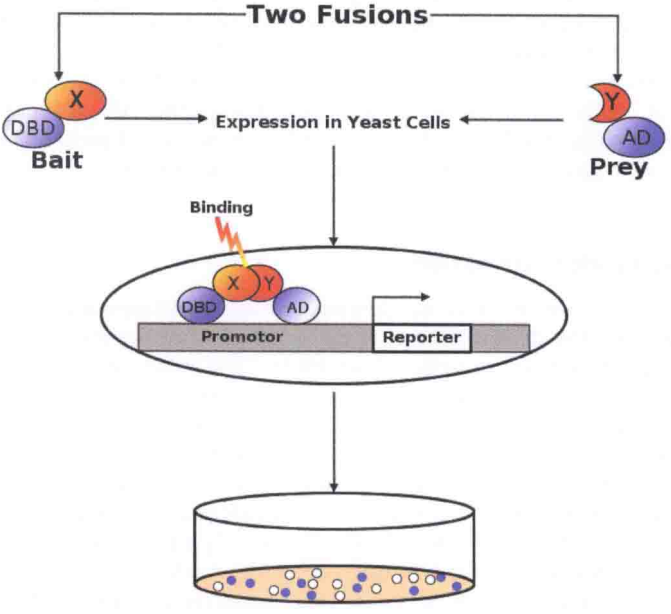


Fig. 3. The Y2H system. Y2H detects interactions between proteins X and Y, where X is linked to BD domain which binds to DNA region promoter.

### 3.1.3 Tandem affinity purification-tag coupled to mass spectrometry (TAP-MS)

TAP-MS method is a powerful approach to determine the composition of relevant protein complexes. In this method, a target protein-coding region is fused with a DNA sequence encoding an affinity tag which will be expressed with other cellular proteins, followed by two-step affinity purification (AP) and elucidation of the complex components by mass spectrometry (MS). A typical TAP tag is formed by an immunoglobulin interacting domain of protein A (protA) and a calmodulin-binding peptide (CBP) (Fig. 4). The protA/CBP binding domains are separated by a short recognition sequence for the site-specific tobacco-etch virus protease (TEV protease). The TEV site allows proteolytic elution of the protein complex from IgG-sepharose after the first affinity-purification step, which is based on the protA/IgG-sepharose interaction. The eluted protein complex is further purified by binding to a calmodulin affinity resin, eluted with EGTA and processed for identification with MS analyses.