Chapter 1

Introduction

Most cellular processes are carried out by multi-protein complexes. The identification and analysis of their components provide insight into how the ensemble of expressed proteins (proteome) is organized into functional units. A fundamental problem in modern biology is the identification of protein complexes and protein interactions in a cell.

Protein-protein interactions have been studied extensively. Different experimental methods are available to identify such interactions. They can be roughly divided into two main categories: small-scale (low-throughput) and large-scale (high-throughput) techniques. Given a set of proteins, small-scale techniques such as co-immunoprecipitation (co-IP) determine the interaction between one pair of proteins at a time [2]. More recently, large-scale techniques have been developed, such as yeast two-hybrid and tandem affinity purifications (TAP), that allow the simultaneous identification of large number of interacting pairs and protein complexes [14, 15, 20, 22, 26, 45].

The need to perform large-scale protein identification experiments has increased with the advent of genome-wide analysis. These experiments allow us to observe interactions between a great number of proteins, possibly even all proteins in a genome. When the number of proteins is in the thousands, the number of possibly interacting pairs is in the millions. To discover all these interactions using small-scale experiments is very labor-intensive and time-consuming, and in this situation large-scale experiments are preferred.

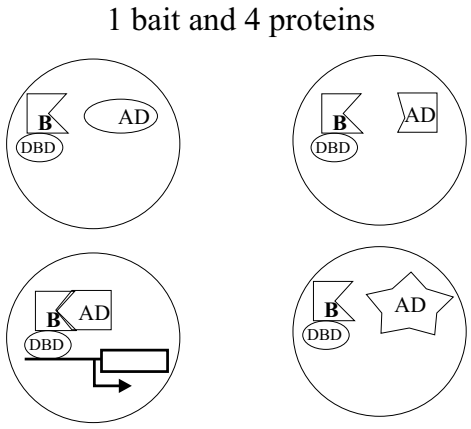

Figure 1.1: Principle of two-hybrid test. 'B' indicates the bait protein. In this example, a library of four proteins is transformed into the bait strain in order to express two interacting proteins in the same yeast cell. If two proteins interact, they reconstitute a transcription factor, which activates a reporter gene. Two-hybrid tests can be automated when large sets of proteins are under study. AD: transcriptional activation domain. DBD: DNA-binding domain.

1.1 Large-scale experimental methods

There are two relevant large-scale experimental techniques, namely yeast two-hybrid and tandem affinity purification.

1.1.1 Yeast two-hybrid arrays

The yeast two-hybrid system is a genetic method for detecting protein-protein interactions. Interactions between two proteins are detected through protein-protein interaction-dependent reporter gene activation *in vitro* [44]. This procedure is typically carried out by screening a protein against a random library of potential protein partners (Figure 1.1). Two-hybrid screening can be done in large-scale using a colony array format, in which each colony expresses a defined pair of proteins. Because the particular protein pair expressed in each colony is defined by its position in the array, positive signals can be used to identify interacting pairs of proteins.

The main advantage of two-hybrid arrays is their systematic nature, which may cover all proteins expressed by a genome. Its disadvantages are a significant number of false positives and the inability to identify interactions involving more than two proteins [44].

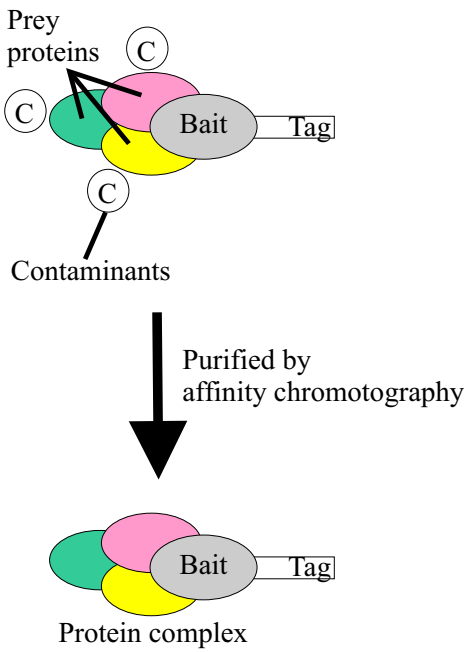

Figure 1.2: Schematic of Tandem Affinity Purification.

1.1.2 Tandem Affinity Purifications-Mass Spectrometry (TAP-MS)

Tandem affinity purifications followed by mass spectrometry identification (TAP-MS) overcomes the restriction to pairwise interaction. TAP-MS involves biochemical isolation of protein complexes and subsequent identification of their constituting proteins using mass spectrometry [3, 23, 47]. Sizable data sets using this method have been produced, in particular for yeast *Saccharomyces cerevisiae* [15, 20, 27], human *Homo sapiens* [7] and *Escherichia coli* [9]. A recent screening delivered interaction data for all proteins accessible by TAP-MS in yeast [14, 26].

In a single TAP-MS experiment, a focal "bait" protein is modified by integrating a polypeptide tag into the protein via standard recombinant DNA techniques (Figure 1.2). The bait protein is then expressed inside a cell where it may carry out its function as part of one or more protein complexes. To retrieve other proteins in a complex, the complex is purified from a cell lysate via affinity chromatography using the tag of the bait protein. A single experiment is referred to as a *purification*. A single purification is supposed to identify "prey" proteins forming protein complexes with the bait protein. The purified proteins are identified by standard mass spectrometry identification. Ideally, these purified proteins would constitute the entire complex encompassing all proteins interacting with the bait protein. Thus, in contrast to yeast two-hybrid experiments, which only reveal pair-

wise interactions, TAP-MS experiments are able to identify more complex protein interaction. In reality, false positive and false negative errors are common and handling them properly is a prerequisite for obtaining meaningful results. In this thesis, we develop techniques to analyze TAP-MS experimental results, based on different error models for tandem affinity purification.

1.1.3 Available data sets

All data sets are from the yeast *Saccharomyces cerevisiae* proteome. We focus on the experimental data from Gavin et al. [14, 15] because they provide raw purifications including several experiments with the same baits. Although the data set from Krogan et al. [26] is more recent and contains more proteins, information on repeated experiments is not available which makes it impossible to analyze the random experimental error.

- **Uetz00** (Uetz et al. [45]). The first publication that used a whole-genome two-hybrid array. Also one of the first reports in which all yeast proteins were studied systematically on the protein level.

- **Gav02** (Gavin et al. [15]). Experimental data produced by large-scale tandem-affinity purification and mass spectrometry (TAP-MS) in a large-scale approach to characterize multiprotein complexes. This study processed $1,739$ genes and obtained 589 purifications. Manual analysis of these purifications defined 232 overlapping protein complexes.

- **Gav06** (Gavin et al. [14]). A data set obtained under the same experimental condition as in Gavin et al. [15], but containing more proteins ($2,760$ proteins) in roughly $2,000$ purifications.

- **Krogan06** (Krogan et al. [26]). High-quality experimental data produced by large-scale tandem-affinity purification and mass spectrometry in a large-scale approach to characterize protein interactions. It contains the most number of proteins ($4,562$ different tagged proteins) and obtained $2,357$ successful purifications. However, it cannot be used to study experimental random error because it lacks information on repeated experiments.

1.1.4 Clustering solutions

In order to judge the quality of our results, we compare them to other clustering solutions and two benchmark data sets curated by experts. The clustering solutions we use for comparison are

- **Cellzome** : A manually curated solution published along with the experimental data by Gavin et al. [15]. Some purifications in this set were assigned to more than one complex, increasing the overlap between different complexes.
- **Krause** : A heuristically created solution by Krause et al. [24].
- **Gavin06 (core)** : A solution published along with the experimental data Gav06 [14] using a heuristic-based hierarchical clustering technique to obtain a subset of core complexes.
- **Gavin06 (all)** : An overlapping solution published along with the experimental data Gav06 [14] using the same technique as above. This solution consists of the core solution and many shared components.
- **MCL** : a clustering solution from applying the Markov Clustering algorithm [46] on an interaction graph.
- **MCODE** : a clustering solution from applying the MCODE algorithm [4] on an interaction graph.

We will introduce MCL and MCODE in the next chapter.

1.1.5 Protein complex annotation

We obtain protein complex annotation from two different sources. Both are curated by experts and derived from other experiments different from the above.

- **MIPS**: The MIPS data set [31] is a standard data set for benchmarking protein complexes. Note that it was largely created before high throughput data was available.
- **Reguly**: A manually curated dataset of protein interactions published recently, which includes protein complexes [36]. It is less selective than MIPS.

1.2 Overview of the thesis

In Chapter 2, we introduce the protein-protein interaction graph as a basic concept, and summarize several known methods that predict protein complexes from such a graph. The subsequent chapters are devoted to presenting our own methods.

We have studied two computational approaches in predicting protein complexes from protein complex purifications: (1) a combinatorial approach based on frequent itemsets, and (2) a partitioning approach based on Markov Random Fields. In Chapter 3, we present our approach for finding overlapping complexes by computing exact frequent itemsets; we model proteins as items and purifications as transactions in a database. In this framework, recurrent frequent itemsets define possible shared modules of protein complexes. In Chapter 4, we extend the exact problem to handle combinatorial error and probabilistic error. In Chapter 5, we present our approach based on Markov Random Fields and compare it with exact frequent itemsets. Finally, in Chapter 6, we present a visualization framework for overlapping protein complexes.