Related work

2.1 Protein complex prediction from networks

At a first glance, the prediction of protein complexes is a well-established problem. Data for physical interactions between proteins in the form of networks is quite abundant and so are clustering approaches that find dense areas in networks [1].

Clustering is a classical problem in computer science and mathematics and has the goal to group objects from a given universe into sets where objects within a set should be similar to each other and objects from different sets should be dissimilar according to an arbitrary similarity or distance measure [61, 62]. The exact definition depends on the specific goal and the context [62].
For a universe with a defined similarity measure one can easily transform the data to a graph structure by defining objects as the vertices and adding edges weighted by the measure for all pairwise comparisons. Additionally, edges could be discretized by removing the ones below a certain threshold and discarding the weights of the remaining ones. This yields an unweighted network in which only the strongly related nodes are connected [61]. The notion of a binary relatedness and its connection to distances enables to model even more problems within the framework of clustering, like social interactions or probable protein interactions. Finding the tighter related groups in a network, regardless of whether its edges are weighted or not, is then called graph clustering [61, 63, 64].

Protein-protein interaction networks (PPIN) offer a global view on the interactions between proteins of an organism by representing proteins as the nodes in the network and physical interactions, the protein-protein interactions (PPI), as an edge between the corresponding nodes. Plenty of methods exist to detect such physical interactions experimentally and modern high-throughput methods like yeast two-hybrid systems [65, 66] and (tandem) affinity purification accompanied with mass spectrometry [67, 68] made the annotation of whole organisms possible [69–72]. Unfortunately,

T. Will, *Predicting Transcription Factor Complexes,* BestMasters,
DOI 10.1007/978-3-658-08269-7_2, © Springer Fachmedien Wiesbaden 2015

the number of false positive and false negative interactions is very high. Indirect binding partners are, for example, often falsely reported as directly interacting ones. Furthermore, the overlap of datasets that are compiled using distinct experimental methods is surprisingly low due to individual strengths and weaknesses of each approach [70, 73–75]. The huge amount of uncertainty in the data is recently tackled quantitatively by the integration of different PPIN datasets and also additional heterogeneous data (like localization, coexpression or genomic evidence) into sophisticated statistical models. These models can then be used to obtain reliable weighted PPINs with interaction probabilities for each edge [69, 76, 77]. Section 3.2.2 will briefly address the Bayesian model that is used to construct PrePPI [71, 77], the PPIN used in this thesis.
Given that a set of proteins forms a complex to accomplish a common biological task, one expects this set to be a cohesive subset in such a PPIN [78–80].

According to [69] the complete characterization of a protein complex is a task that requires to solve several consecutive problems: (1) one needs to determine all member proteins of the complex, (2) the candidates must permit a connected topology of pairwise direct interactions, (3) these interactions have to be related to interactions between domains and distinct binding interfaces, and, with all this information one can (4) predict a 3D structure of the complex.
Section 2.2 will specifically address later steps, but most previous research has focused on the very first step which is basically a graph clustering problem. However, standard clustering is not ideal to detect complexes from PPIN. Many protein complexes are not only organized in a modular but also in a combinatorial fashion. They take part in several functional complexes and their associated nodes in the network therefore should belong to more than one cluster [81–86]. This has led to many sophisticated methods with very different underlying approaches. The first one, MCODE [80], works by iterative vertex reweighting, LCMA [87] searches for cliques and merges them, RNSC [88] optimizes a cost-function based on inter- and intra-complex edges, MCL [89] computes a flow within the network based on properties of its adjacency matrix, RRW [90] uses random walks in an iterative way and CFA [91] grows dense regions from k-connected subnetworks instead of cliques since protein complexes are not necessarily fully connected in a PPIN. A very successful recent method is ClusterONE [81] which will be explained in detail in Section 2.1.1.

Prediction methods always entail benchmarks to test their performance. Section 2.1.2 features common assessments in protein complex prediction that will be used in the thesis.
However, all the previously mentioned complex prediction approaches only apply graph-theoretic algorithms to the topology of the PPIN but neglect biologically important factors like structural limitations or combinatorial assembly which eventually leads to a high number of false positive predictions [69, 83, 86, 92–94]. Section 2.2 will elaborate the problems and introduce current solutions.

2.1.1 ClusterONE and cohesiveness

ClusterONE (short for clustering with overlapping neighborhood expansion) [81] is not only the opening track of Pink Floyd's 1994 album "The Division Bell"[1], but also one of the best performing complex prediction approaches available at the moment and one of the few that can handle weighted edges in PPINs as well as overlap of complexes.
It optimizes a very plausible metric for measuring the cluster quality called cohesiveness. The cohesiveness f for a set of proteins V in a network is defined as

$$f(V) = \frac{w^{in}(V)}{w^{in}(V) + w^{bound}(V) + p|V|}$$

where $w^{in}(V)$ denotes the total weight of edges between members of V (internal edges) and $w^{bound}(V)$ denotes the total weight of edges that connect V with the rest of the network (boundary edges). $p|V|$ serves as a penalty term with the purpose to model yet undiscovered interactions in the data. For $p > 0$ one assumes an additional boundary weight of p per protein in V. Figure 2.1 illustrates these definitions with an example and additionally introduces the notion of incident and boundary proteins.

Cohesiveness exactly assesses the structural properties of subnetworks we want to obtain: they should be densely connected but at the same time well separated from the outside. For example, $f(V) > 1/3$ implies that the proteins of the chosen subset have more internal than external weight on average. This satisfies the condition of a weak community [63].

Given some initial seed protein a growth algorithm iteratively increases

[1] `http://en.wikipedia.org/wiki/Cluster_One`

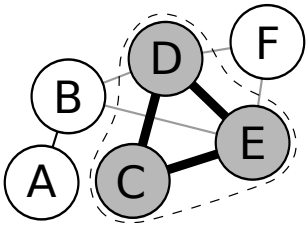

Figure 2.1: For convenience all edges in the network have unit weight and the corresponding weight annotation is omitted. The current members of the cohesive subset $V = \{C, D, E\}$ are shown darker and their internal edges are shown thicker, boundary edges are marked gray. Boundary edges can be thought to span a boundary (shown dashed) that separates the current dense subset V from the remaining network. This border defines the set of incident proteins $V_{inc} = \{B, F\}$, external vertices adjacent to the boundary, and boundary proteins $V_{bound} = \{D, E\}$, internal vertices at the boundary.
For the given example $w^{in}(V) = 3$, $w^{bound}(V) = 4$ and $f(V) = \frac{3}{7}$ if we neglect the penalty parameter p. Figure adapted from [1].

the cohesiveness using a greedy procedure. Based on the set of currently chosen proteins V, first the set of incident proteins V_{inc} and the set of boundary proteins V_{bound} are determined. All proteins in V_{inc} are adjacent to some protein in V and could be added to V in the step, all members in V_{bound} are on the boundary of V and could be removed from V. Figure 2.1 clarifies this on an example. For each of these possibilities to expand V or to shrink V, it is tested how the change would affect the cohesiveness $f(V')$ of the thus modified set V'. The algorithm then chooses the single addition/removal with the highest increase in cohesiveness as long as it can be further increased. If no further increase is possible it returns a locally optimal solution (see Algorithm 2.1).
ClusterONE can be run in a seeded mode, where the user explicitly supplies a list of seed proteins for its growth procedure algorithm, or one can leave this selection to the program. In this mode the growth is initially started from the protein with the largest number of connections (highest degree). After the completion of a single growth process, from all the proteins that are not yet included in one of the complex candidates so far, again the one with the highest degree is chosen as the next seed protein. This is done until there are no more proteins to consider.
In the next step of the algorithm highly overlapping complex candidates are merged. At first an overlap graph G is constructed where each previously

Algorithm 2.1 Iterative cohesiveness optimization starting in v_0

```
startprotein: V_0 ← {v_0}
t ← 0 {t: step number}
repeat
  V_{t+1} ← V_t
  determine current V_inc and V_bound

  check if addition/removal is valuable:
  for ∀v ∈ V_inc do
    V' ← V_t ∪ {v}
    if f(V') > f(V_{t+1}) then
      V_{t+1} ← V'
    end if
  end for
  for ∀v ∈ V_bound do
    V' ← V_t \ {v}
    if f(V') > f(V_{t+1}) then
      V_{t+1} ← V'
    end if
  end for

until V_t ← V_{t+1} {as long as further increase is possible}
return V {output locally optimal cohesive subset}
```

determined complex candidate is a vertex in the graph and two vertices A and B are connected by an edge if the overlap score $\omega(A, B)$ between the two sets is larger than 0.8 (for definition see Section 2.1.2). All candidates within the same connected component in G are then merged into single protein complex candidates; single vertices in G are carried over to the predicted output set without any merging step.
In the final step complex candidates with less than three members or below a certain density are discarded and the remaining ones returned.

2.1.2 Quality measures for complex prediction

Since complex prediction is already a well-established problem, plenty of benchmarks exist to assess the quality of such predictions. They can be clearly separated into two distinct but complementary categories: measurements based on the mutual agreement with reference complexes and measurements that account for plausible biological relationships within the predicted clusters.

Comparison with reference protein complexes

Given reliable and as complete as possible reference data of protein complexes for a certain organism, one can compare predictions to this compilation of known complex assemblies.
Unlike many other prediction problems, perfect matches to known complexes are rarely seen in protein complex prediction which raises the need for specifically adapted quality measures. A compilation of common ones is covered in the following paragraphs.

Overlap score and related scores

The overlap score suggested by Bader et al. [80] is a measure of overlap for pairs of sets and the foundation of many metrics to assess the quality of complex prediction approaches. Benchmarks based on the overlap score have been used in plenty of publications in slightly different variants and under different names [80, 81, 83, 93–95].
The overlap score ω between two sets of proteins A and B is defined as

$$\omega(A, B) = \frac{|A \cap B|^2}{|A||B|}.$$

Given some threshold t and $\omega(A, B) > t$ or $\omega(A, B) \geq t$ (depending on the exact definition) we call A and B *matched*.
The actual threshold can in principle be set as desired. In their original publication [80] evaluated this question for different thresholds and suggested t should be between 0.2 and 0.3 to get rid of biologically insignificant overlaps. In the remaining thesis we will use $\omega(A, B) > 0.25$ as used by [81] and [93]. The reasoning behind 0.25 is the fact that, given the compared complexes A and B are equally large, they share at least half of their proteins.

With the notion of a match one can compute a precision P as the fraction of predicted complexes that can be matched to a reference and a recall R as the fraction of reference complexes matched by predicted complexes. Additionally, one can define a combined F-score (or F-measure) using the harmonic mean

$$F = 2\frac{PR}{P+R}$$

which is very common in information retrieval and was also used frequently in complex prediction [83, 95, 96]. All scores range from 0 to 1. $P = 1$ means that all predicted complexes can be related to true ones and $R = 1$ means that all known complexes are predicted.
The naming of the individual scores is inconsistent across publications but the quintessence is covered by this paragraph.

Geometric accuracy

The geometric accuracy goes back to Brohee and van Helden [92] and is based on a combination of the clustering-wise sensitivity Sn and positive predictive value PPV. Both metrics work on a contingency table $T = [t_{ij}]$, which is an $n \times m$ matrix where row i corresponds to the i^{th} among the n reference complexes, column j to the j^{th} predicted complex and t_{ij} denotes the number of shared proteins between reference i and prediction j. The cardinality of reference complex i is given by n_i. Then Sn and PPV are defined as:

$$Sn = \frac{\sum_{i=1}^{n} \max_{j=1}^{m} t_{ij}}{\sum_{i=1}^{n} n_i}$$

$$PPV = \frac{\sum_{j=1}^{m} \max_{i=1}^{n} t_{ij}}{\sum_{j=1}^{m} \sum_{i=1}^{n} t_{ij}}.$$

These two scores, each one optimizing a different direction, are joined using the geometric mean to obtain the final score:

$$Acc = \sqrt{Sn \times PPV}.$$

It is necessary to balance the two different metrics in the above manner since both can be artificially boosted by rather extreme predictions. The clustering-wise sensitivity Sn can be cheated by putting every protein in one huge cluster, while the positive predictive value PPV could be maximized

by putting every protein in its own cluster.
The usage of the geometric mean instead of the arithmetic mean is beneficial in this case because then the combined score *Acc* is penalized stronger if one of the scores is comparably low. The arithmetic mean would give a false idea of the prediction quality in the extreme cases mentioned above: since one can easily get $Sn = 1$ or $PPV = 1$, an arithmetic accuracy would achieve $Acc_{arith} > 0.5$ for a hardcoded prediction [81, 95].

Maximum matching ratio

The Maximum matching ratio (MMR) is a rather recent quality measure proposed by [81] and is based on a maximal weight one-to-one mapping between reference and predicted complexes.
To establish this mapping a bipartite graph is constructed where reference complexes are related to predicted complexes by an edge if their overlap score is larger than zero. Now one can find one-to-one mappings using the definition of a matching from graph theory. A matching of a graph is a set of edges that do not share common endpoints, which means no prediction is assigned to several references and vice versa. A matching of a graph is maximal if there is no matching for this graph with a larger cardinality and maximum if there is no matching with a bigger sum of edge weights [97].
The MMR is defined as the sum of edge weights of the maximum matching divided by the number of reference complexes to be matched. The MMR is set in relation to the number of reference complexes and not offset against the number of predicted complexes. There are good reasons for this. Reference complexes are inherently incomplete, therefore unmatched predicted complexes should not be penalized [81].

Biological relevance

The second category of benchmarks is concerned with the assessment of the biological relevance of predicted protein complexes. It can be thought of as a biological plausibility check conceived to complement the measurements that rely on incomplete reference complex data.

Colocalization
The test for colocalization is based on the assumption that proteins within a complex must be locatable in a common compartment [76]. Given location data for all involved proteins, the colocalization score is defined as the average fraction of proteins encountered in the most common compartment

within the complex weighted by the size of the complex [81, 98]. In the particular case of transcription factor complexes only the nucleus is an appropriate locality. Nucleus colocalization is thus defined as:

$$Coloc_{nucl} = \frac{\sum_{\forall \text{ complexes } i} \frac{\text{members of complex } i \text{ located in the nucleus}}{|\text{complex } i|} |\text{complex } i|}{\sum_{\forall \text{ complexes } i} |\text{complex } i|} = \frac{\sum_{\forall \text{ complexes } i} \text{members of complex } i \text{ located in the nucleus}}{\sum_{\forall \text{ complexes } i} |\text{complex } i|} .$$

Gene Ontology enrichment

The last test has become a standard analysis in all areas of computational biology and assesses functional homogeneity within the complexes based on an extensive genome-wide annotation.

The Gene Ontology (GO) annotation [99] is a standardized representation for attributes of genes and gene products across species and databases. The GO defines a hierarchical relationship between annotation terms and is structured as a directed acyclic graph. Individual terms are represented as nodes and directed edges connect them to more specific terms to establish a parent/child-relationship where each term can be child of several parents. Consequently, for every term a gene is annotated with, it is furthermore associated with all the less specific parents of that term. The three distinct ontology domains, the roots of these trees, are:

molecular function: biochemical activity of the protein

biological process: biological objective to which the protein contributes

cellular component: localization where the protein is active

GO terms can be used to conduct an enrichment analysis. If a set of proteins is related functionally or members contribute to a common pathway one expects to find evident GO terms in the set more often than by mere chance. Thus, annotations that are found in an examined set of proteins are tested for overrepresentation against a suitable background such as all genes of an organism or all genes covered by the microarray used in the study. This type of analysis has become a quasi-standard in the investigation of biological data and was also used in the context of protein complexes before [81, 95, 100, 101].

The probability P to observe k or more proteins annotated with term X in a set of m proteins by chance is given by a hypergeometric distribution:

$$P = \sum_{i=k}^{m} \frac{\binom{M-K}{m-i}\binom{K}{i}}{\binom{M}{m}}$$

where M denotes the number of proteins in the background, m the number of proteins in the studied set, K all proteins in the background annotated with term X, and k the number of proteins in the studied set annotated with X. Term X is then said to be enriched in the studied set at significance level p if $P < p$. Since multiple hypothesis testing is performed to examine if the set of interest contains an enriched GO term, significance levels have to be adjusted. A simple and conservative method to correct them is the Bonferroni correction whereby the significance levels are simply divided by the number of tests [102].
To utilize the idea of GO enrichment for the assessment of complex prediction quality the overrepresentation score is defined as the fraction of complex candidates with at least one enriched annotation at significance level $p = 0.05$ [81, 100].

2.2 Protein complex prediction beyond protein interaction networks

PPINs provide a holistic view on protein connectivity which, with respect to applicability in the prediction of protein complexes, misses certain important layers of information [69]. Interaction networks offer a compilation of assumed pairwise interactions that are thrown together to a static entity but in reality the network is highly dynamic and intrinsically controlled by protein expression, affecting the current state of the network in time and space, and spatial constraints. To enable complex formation all involved proteins must be expressed at the same time, in spatial proximity and capable of forming a stable binding topology devoid of any binding site competition [69, 83–86, 103, 104]. Hence, predicted clusters in PPINs are not necessarily valid biological complexes and their interpretation can be quite ambiguous and even lead to false positive complex predictions as Figure 2.2 illustrates. Nonetheless, dense connectivity and slight seclusion still suggest that clusters in PPINs form at least functional modules, what are groups of proteins transiently interacting with each other on cellular

http://www.springer.com/978-3-658-08268-0

Predicting Transcription Factor Complexes
A Novel Approach to Data Integration in Systems Biology
Will, T.
2015, XIX, 142 p. 29 illus., Softcover
ISBN: 978-3-658-08268-0