BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 20 2012, pages 2624–2631
doi:10.1093/bioinformatics/bts469

Systems biology

Advance Access publication August 20, 2012

FACETS: multi-faceted functional decomposition of protein interaction networks

Boon-Siew Seah[1,3,*], Sourav S. Bhowmick[1,3,*] and C. Forbes Dewey, Jr[2,3]

[1]School of Computer Engineering, Nanyang Technological University, Singapore, [2]Biological Engineering Department, Massachusetts Institute of Technology, Cambridge, MA 02139, USA and [3]Singapore-MIT Alliance, Nanyang Technological University, Singapore

Associate Editor: Burkhard Rost

ABSTRACT

Motivation: The availability of large-scale curated protein interaction datasets has given rise to the opportunity to investigate higher level organization and modularity within the protein–protein interaction (PPI) network using graph theoretic analysis. Despite the recent progress, systems level analysis of high-throughput PPIs remains a daunting task because of the amount of data they present. In this article, we propose a novel PPI network decomposition algorithm called FACETS in order to make sense of the deluge of interaction data using Gene Ontology (GO) annotations. FACETS finds not just a single functional decomposition of the PPI network, but a *multi-faceted atlas* of functional decompositions that portray alternative perspectives of the functional landscape of the underlying PPI network. Each *facet* in the atlas represents a distinct interpretation of how the network can be functionally decomposed and organized. Our algorithm maximizes interpretative value of the atlas by optimizing *inter-facet orthogonality* and *intra-facet cluster modularity*.

Results: We tested our algorithm on the global networks from *IntAct*, and compared it with gold standard datasets from MIPS and KEGG. We demonstrated the performance of FACETS. We also performed a case study that illustrates the utility of our approach.

Contact: seah0097@ntu.edu.sg or assourav@ntu.edu.sg

Supplementary information: Supplementary data are available at the *Bioinformatics* online.

Availability: Our software is available freely for non-commercial purposes from: http://www.cais.ntu.edu.sg/~assourav/Facets/

Received on March 2, 2012; revised on June 27, 2012; accepted on July 22, 2012

1 INTRODUCTION

The massive amount of biological interaction datasets presents the opportunity to study higher order organization and modularity of interaction networks. High-throughput interaction experiments, however, introduce new challenges to visualization and analysis of biological interaction data. A common thread that runs through high throughput generated data is information overload, i.e. the explosion of data that makes intuitive and meaningful functional analysis difficult, even overwhelming. In case of protein–protein interaction (PPI) data, decomposing the network into functional modules is often the key step to understanding the overall picture of the functional relationships that underlie the data. Consequently, graph clustering methods that decompose PPI networks into their functional constituents are increasingly pertinent (Lavallée-Adam *et al*., 2009).

In general, graph clustering algorithms discover regions of dense connectivity that represent protein complexes or functionally coherent processes (Bader and Hogue, 2003; Krogan *et al*., 2006; Seah *et al*., 2012). Unfortunately, these methods output *only a single optimal functional decomposition* of the PPI network. Consequently, a PPI network can only be decomposed and viewed from a single perspective, whereas in reality there are often multiple different perspectives (decompositions) associated with the functional organization of the underlying network, all of which are distinct and equally valid. We refer to each of these decompositions as a *facet* because they visualize the organization of a PPI network from a unique view, providing a distinct interpretation of the organization of the underlying network. For example, consider the toy transcriptional regulatory network depicted in Figure 1. A typical decomposition, based on an existing graph clustering technique (e.g. mcode in Bader and Hogue, 2003), identifies dense regions of the network, which correspond to the decomposition of protein complexes as shown in *Facet 1*. However, this network can also be viewed from other different perspectives. For instance, it can be organized by the types of signaling pathways involved in it (*Facet 2*). Notice that the decomposition from this perspective is markedly different from the complex-based decomposition. Furthermore, different proteins in the network may undergo various modifications such as acetylation, phosphorylation and ubiquitination. Hence, yet another way to decompose the network is by their modification effects as depicted in *Facet 3*. Clearly, in larger real-world networks the possibility of uncovering multiple, distinct functional decompositions are real.

At first glance, it may seem that we can tune the clustering parameters of existing graph clustering techniques in order to generate multiple facets or decompositions. Unfortunately, such tuning only generates an exponential number of *slightly perturbed* decompositions with incremental differences (see Supplementary Material). In other words, such strategy does not generate functionally unique decompositions. In contrast, it is imperative to ensure that the decompositions or facets are *distinctive*, i.e. they are maximally different from each other. This is because every facet should provide a fresh and informative perspective to the organization of the network, rather than providing just incremental differences with respect to other facets.

*To whom correspondence should be addressed.

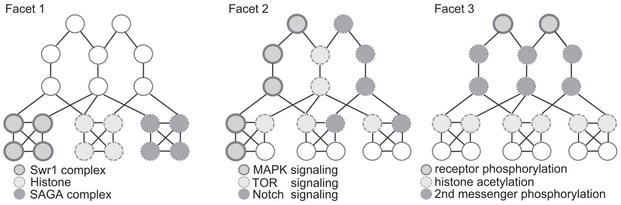

Fig. 1. Illustration of multi-faceted PPI network decomposition

Our contribution. We propose a novel algorithm called FACETS that discovers an *atlas* of functionally unique decompositions from a PPI network, portraying alternative views of the functional landscape of the network (detailed in Section 2). Each decomposition or facet represents a distinct interpretation of how the network can be functionally decomposed and organized. Since a key objective is to obtain n unique facets that are informative and orthogonal (We use the term orthogonal to describe the idea of distinctive clusters, rather than its precise mathematical meaning.), our algorithm maximizes interpretative value of the atlas by optimizing *intra-facet cluster modularity* and *inter-facet orthogonality*. *Intra-facet cluster modularity* captures the aim of decomposing a PPI network G based on a particular functional and/or structural view. For instance, based on complexes and localized structures, G can be decomposed into protein complexes. If we consider regulatory processes as a functional concept, then G can be decomposed into signaling and regulation pathways, an entirely different decomposition. *Inter-facet orthogonality*, on the other hand, demands that each of the n facets are structurally distinctive and functionally apart from each other. We propose a novel *objective function* that models these intuitions and FACETS exploit it to discover a set of distinct facets. Specifically, we exploit both the PPI graph structure and the rich functional information provided by Gene Ontology (GO) annotations to guide facets construction.

In Section 3 we evaluate the performance of FACETS on real world PPI datasets. We also compare FACETS-generated decompositions against several gold standard datasets. We demonstrate its superiority over tested graph clustering methods. We illustrate the robustness of FACETS against noise. Finally, we conduct a case study to illustrate how multi-faceted decompositions identified multiple organization maps of the human autophagy system (Behrends *et al*., 2010).

Related work. Multi-view clustering is a poorly studied problem in the data mining community (Qi and Davidson, 2009). Still, there are several works that have focused on multi-view clusterings in image and text mining domain (Niu and Dy, 2010). One approach projects data into an alternative subspace (Cui *et al*., 2007). Another approach generates alternative clustering through the use of *must-link* and *cannot-link* constraints (Wagstaff and Cardie, 2000). In meta-clustering (Caruana *et al*., 2006), a large number of clusters are generated and clusters which are truly different are selected. All of the aforementioned approaches, however, assume data points in the *vector space* that allow the notion of metric distances in a Euclidean geometry. On the other hand, our problem demands a multi-view clustering methodology on *attributed graphs*, which requires a graph clustering paradigm on both structure and annotation. To the best of our knowledge, multi-view clustering paradigm has not been applied in clustering biological networks to identify pertinent functional modules from multiple perspectives.

Ensemble clustering methods generate an ensemble of near-optimal decompositions (Agarwal and Kempe, 2008; Massen and Doye, 2005; Navlakha and Kingsford, 2010). These methods have been used to increase the quality and confidence of the decomposition and understand network dynamics. The near-optimal decompositions generated, however, have no notion of the orthogonality that this work is seeking. Instead, ensemble clusterings create a large number of perturbed solutions, making them unsuitable as an atlas of functionally distinct decompositions. For instance, in (Navlakha and Kingsford, 2010), a small network of 32 nodes generated at least 82 permutations of clusterings.

2 MATERIALS AND METHODS

In this section, we formally introduce the multi-faceted functional decomposition problem. We begin by defining some terminology that we shall be using in the sequel. We use the network in Figure 1 as running example in this article.

2.1 Terminology

An undirected network $G = (V, E)$ contains a set of vertices V, representing biological entities like proteins or genes, a set of edges E, representing interactions between the entities. A *functional module* $C_i = (V_c^i, E_c^i)$ is a subnetwork of G such that $V_c^i \subset V$ and E_c^i is the set of edges induced by V_c^i from G. A *facet* (*decomposition* or *view*) of G, denoted by F, is a set of functional modules $\{C_1, \ldots, C_m\}$ representing a specific functional concept. Functional modules within a facet F are allowed to overlap. In the sequel, we use the terms facet, view and decomposition interchangeably. A *functional atlas* (or atlas for brevity) of G, denoted by A, is a set of facets $\{F_1, F_2, \ldots, F_n\}$ that represents distinctive functional landscapes of G. Figure 1 depicts an atlas of three facets, with each facet decomposing the network into three functional modules.

In order to support the idea of functionally orthogonal views, we utilize GO annotations associated with proteins. Given a GO directed acyclic graph (DAG) $D = (V_{\mathrm{GO}}, E_{\mathrm{GO}})$, the ordered set $\Delta = \langle \Delta_1, \Delta_2, \ldots, \Delta_d \rangle$ is a topological sort of D, where Δ_i represents a single GO term. Each vertex $v \in V$ is associated with a d-dimensional *function association vector* $\Delta_v \in \{0, 1\}^d$, such that $\Delta_v = \langle \Delta_1^v, \Delta_2^v, \ldots, \Delta_d^v \rangle, \Delta_i^v \in \{0, 1\}$, where $\Delta_i^v = 1$ if and only if the term $\Delta_i \in D$ or its descendants are associated with protein v, and $\Delta_i^v = 0$ if otherwise. Note that Δ_v is an indicator vector that indicates GO terms that are associated with v.

A *facet candidate bundle* $B_i = \{G_1, G_2, \ldots, G_m\}$ is a set of connected subnetworks of G such that for every $G_k \in B_i$, there is a shared GO term Δ_i within every $v \in V_k$. Δ_i represents the common function of the candidate subnetwork. A facet candidate bundle B_i represents the superset of facet F_i and it contains a large permutation of subnetworks that satisfy a particular functional concept. Typically, $|F_i| \ll |B_i|$. A *function bundle* $\omega_i = \{\Delta_1, \Delta_2, \ldots \Delta_m\}$ is the set of shared GO annotations of B_i, i.e. $\omega_i = \bigcup_{G_k \in B_i} \Delta_{G_k}$. To illustrate these concepts, consider the PPI network in Figure 1. Suppose that B_1 is a facet candidate bundle with $\omega_1 = \{\Delta_1, \Delta_2\}$, where Δ_1 represents the `Swr1 complex` GO term and Δ_2 the Histone term. In the subgraph with `Swr1 complex` label in Facet 1, every node in that subgraph is annotated with `Swr1 complex` term. Thus, the subgraph is a valid member of B_1. Any subgraph made up of `Histone`-labeled nodes is also a valid member of B_1. If B_2 represents the facet candidate bundle with $\omega_2 = \{\Delta_3\}$, where Δ_3 represents `cellular component`, then the `Swr1 complex`-labeled subgraph is also a valid member of B_2 (`Swr1 complex` is a `cellular component`).

Furthermore, every subgraph in Facet 1 whose nodes are labeled is a valid member of B_2, but not neccessarily a valid member of B_1. One can see that B_i contains a set of subgraphs that shares specific functional concepts depending on the functional terms in ω_i. We define the function $f\colon P(V_{go}) \rightarrow$ A given by $f(\omega_i) = F_i$ to make explicit the association between a functional bundle and its corresponding facet.

A *function bundle partition* $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$ is the set of function bundles that form a partition of all GO terms V_{GO}, i.e. $\bigcup_{\omega_i \in \Omega} = V_{\text{GO}}$. In the next section, we shall impose further constraints on facet candidate bundles and function bundles such that the shared GO terms of the subnetworks within each facet candidate bundle shares high functional commonality and the terms shares in one facet are distinct from the terms in another facet.

2.2 Problem formulation

The goal of multi-faceted functional decomposition problem is to identify an atlas of n distinct facets of G that maximizes *inter-facet orthogonality* and *intra-facet cluster modularity*. Each facet depicts a higher-order organization of modules of G. Recall that inter-facet functional orthogonality demands that each of the n facets is based on an orthogonal functional concept—facets that are distinctive and functionally apart from each other. Hence, we propose two criteria that model the intra-facet functional modularity and inter-facet orthogonality of an atlas solution. Next, we propose an *objective function* that models and scores an atlas of G.

2.2.1 Intra-facet cluster modularity Intra-facet cluster modularity enables us to seek clusters that are both structurally and functionally modular. Given ω_i, Ω and G, *ω-restricted* decomposition procedure (denoted by g_ω) computes a decomposition of G into F_i such that F_i satisfies the following criteria:

Criterion 1. Every module $C_j \in F_i$ should be *functionally bounded* by ω_i. Let $D_{C_j} = \{\Delta_1, \Delta_2, \ldots, \Delta_m\}$ be the set of shared terms in C_j, i.e. for every $v \in V_c^j$, v must be annotated with every $\Delta_i \in D_{C_j}$. Then, the *functional boundedness* of module C_j by ω_i is given by $r(C_j, \omega_i) = D_{C_j} \cap \omega_i$. A cluster C_j is bounded by ω_i if $r(C_j, \omega_i) \neq \emptyset$. An ω_i-restricted decomposition of a facet draws from a restricted search space of subnetworks in G whose vertice shares at least a term within ω_i. Intuitively, this means that for any subnetwork to be considered as a module, it must first be sharing a term in ω_i. Even if a subnetwork is dense, it must yield to sparser subnetwork candidates if it is not enriched with terms within ω_i. In the example of Figure 1, if ω_1 is terms of protein complexes, then any subgraphs enriched with complex terms is in the search space for *Facet 1*. In contrast, the modules of *Facet 2*, enriched with signaling terms, would be invalid candidates for *Facet 1* decomposition. This restricted search space is modeled by facet bundle B_i, where any valid candidate facet cluster C_j of facet F_i must belong to B_i.

Criterion 2. A facet F_i decomposes G by maximizing a clustering objective function $o(F_i)$ while satisfying the above criterion. These criteria are determined by the specific graph clustering algorithm that is adapted for creating a facet; for generality we let this be the objective function $o(F_i)$ that has to be maximized by the graph clustering algorithm. For instance, every module $C_j \in F_i$ has to be structurally dense and/or functionally coherent (i.e. every node in module shares a common function), the coverage of F_i has to be high, and the amount of overlap between modules should be low. For example, modules of *Facet 2* maximize $o(F_2)$ while satisfying the ω_2 bound, despite not forming dense modules. This is because all dense modules formed are enriched with complex terms, violating the ω_2 bound.

2.2.2 Inter-facet orthogonality Since we want every facet in the atlas to be functionally and structurally distinct, modules within a facet, as whole, should be structurally and functionally distinct from modules within another facet. We discuss two independent distance measure between facets: *functional orthogonality* and *structural orthogonality*.

Functional orthogonality is indirectly controllable by the function bundles attached to facets, which determines the types of allowable modules through the aforementioned restriction. By increasing inter-bundle functional orthogonality, we increase the functional distinctiveness of each facet. To impose functional orthogonality, we introduce the following constraint: for every $\omega_i, \omega_j \in \Omega, \omega_i \cap \omega_j = \emptyset$ and $\bigcup_{\omega_i \in \Omega} = V_{\text{GO}}$. This requires that Ω actually partitions the terms of the GO DAG. The *functional distance measure* between Δ_i and Δ_j, denoted by $d(\Delta_i, \Delta_j)$, measures the functional dissimilarity between the terms. In this article, $d(\Delta_i, \Delta_j)$ is simply computed as the length of the shortest path between the terms: $d_f(\Delta_i, \Delta_j) = \min_{\Delta_r \in R} |p(\Delta_r, \Delta_i)| + |p(\Delta_r, \Delta_j)|$, where R is the set of common ancestors of term Δ_i and Δ_j and $|p(i,j)|$ is the length of the shortest path from node Δ_i to Δ_j in D. The *candidate function specificity* $s(\Delta_i, C_u)$ is defined as $s(\Delta_i, C_u) = \frac{|\{\Delta_i \in \Delta_v | v \in V_c^u\}|}{|\{\Delta_i \in \Delta_v | v \in V\}|}$. $s(\Delta_i, C_u)$ measures the specificity of a shared GO term, which we will later use to weigh the contribution of the term. For instance, a cluster C_j of 5 nodes that share the `biological process` GO term in a network of 1000 `biological process` annotated nodes has a low specificity value of 0.005 with respect to the term.

Likewise, we define structural orthogonality. The *structural distance measure* between two clusters C_u and C_v is defined as $d_s(C_u, C_v) = 1 - |E_C^u \cap E_C^v| \ / |\{(v_i, v_j) | v_i \in V_C^u \cap V_C^v, v_j \in V_C^u \cup V_C^v, (v_i, v_j) \in E_C^u \cup E_C^v\}|$. The distance is 0 if C_u and C_v shares all edges and 1 if C_u and C_v shares no edges.

Following that, we define $t(\Omega, A)$ as the linear combination of inter-facet functional and structural orthogonality, as follows:

$$t(\Omega, A) = \gamma \sum_{\substack{\omega_i, \omega_j \in \Omega, \\ i \neq j}} \left\{ \sum_{\substack{\Delta_j \in D_{C_j}, \\ C_j \in f(\omega_j)}} \sum_{\substack{\Delta_t \in D_{C_i}, \\ C_i \in f(\omega_i)}} s(\Delta_i, C_i) s(\Delta_j, C_j) \frac{d_f(\Delta_j, \Delta_i)}{|V_p^j||V_p^i|} \right\}$$

$$+(1-\gamma) \sum_{\substack{\Delta_u \in D_{C_u}, C_u \in F_i \\ F_i \in A}} \sum_{\substack{\Delta_v \in D_{C_v}, C_v \in F_j \\ F_j \in A, i \neq j}} s(\Delta_u, C_u) s(\Delta_v, C_v)\, d_s(C_u, C_v)$$

The parameter γ weighs the contribution of d_s against d_f, and is set to attain balanced contribution from both terms. Note that $t(\Omega, A)$ quantifies the pairwise orthogonality between two function bundles. The higher the score, the greater the orthogonality.

2.3 Problem definition

The multi-faceted functional decomposition of G is defined as the problem of simultaneously constructing the atlas of decompositions $A = \{F_1, \ldots, F_n\}$, and the function partition $\Omega = \{\omega_1, \ldots, \omega_n\}$, such that the following objective function is maximized:

$$\begin{array}{ll} \max\limits_{A, \Omega} & \lambda t(\Omega, A) + (1-\lambda)|A|^{-1} \sum\limits_{F_i \in A} o(F_i) \\ \text{subject to} & C_s \in B_i \forall C_s \in F_i, 1 \leq i \leq n \end{array}$$

The right half of the terms captures the cost function of decomposing G into A; the left half, decomposing D into Ω. The parameter λ controls the weightage between the two terms. Observe that one has to optimize these criteria simultaneously over the space of A and Ω. Otherwise, one may end up with a poor objective score. For instance, if $t(\Omega, A)$ is high (meaning highly orthogonal partitioning), but Ω is improperly partitioned such that one ends up with ω_i that allow only poor decompositions, then the $o(F_i)$ score would be very low. Due to the interdependence of the criteria, optimizing the aforementioned function is computationally expensive.

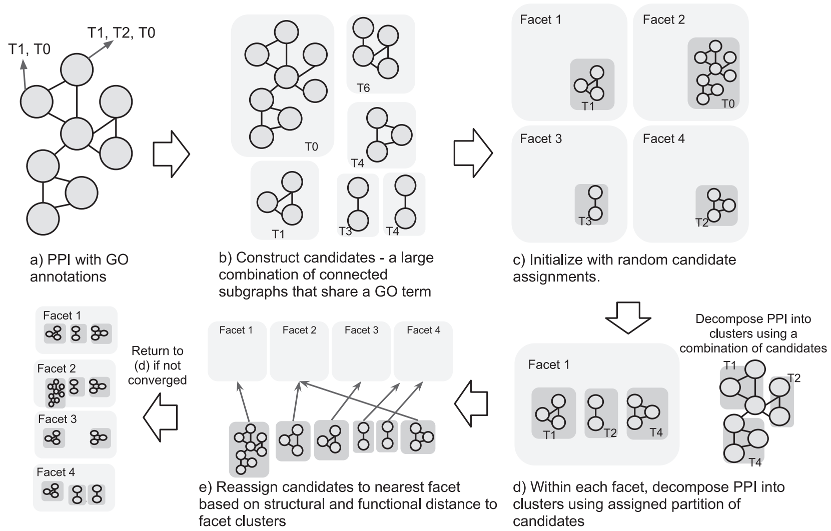
T1, T0
T1, T2, T0
a) PPI with GO annotations
T0
T6
T4
T1
T3
T4
b) Construct candidates - a large combination of connected subgraphs that share a GO term
Facet 1
Facet 2
T1
T0
Facet 3
Facet 4
T3
T2
c) Initialize with random candidate assignments.
Decompose PPI into clusters using a combination of candidates
T1
T2
T4
Facet 1
T1
T2
T4
d) Within each facet, decompose PPI into clusters using assigned partition of candidates
Facet 1
Facet 2
Facet 3
Facet 4
e) Reassign candidates to nearest facet based on structural and functional distance to facet clusters
Return to (d) if not converged
Facet 1
Facet 2
Facet 3
Facet 4

Fig. 2. Illustration of the FACETS algorithm. (**a**) GO annotated PPI network is used as input. (**b**) The set of candidate subnetworks are computed. (**c**) An initial set of modules are randomly assigned to a facet. Candidate subnetworks are then assigned to their nearest facet based on function and structure distance. (**d**) For each facet, decomposition is performed to identify modules that are functionally contained by the facet candidate bundle. (**e**) The candidate subnetworks are reassigned based on their distance to the new set of modules identified. Convergence is achieved when the number of terms reassigned to a different facet drops below the threshold parameter θ. Otherwise, Steps (d–e) are repeated

2.4 FACETS algorithm

Generally, the problem of finding clusters that maximizes typical clustering objective functions that relate to graph density is known to be NP-hard (Jagota, 1995). Hence, the FACETS algorithm is a heuristic implementation that attempts to find a local maximum of the objective function. Our heuristics is a step-wise iterative approach that incrementally optimizes Ω and A, one at a time (Fig. 2). Intuitively, given an attributed PPI network (e.g. Fig. 2a), Ω is incrementally updated by using each facet in A as functional centroids, and then using the centroids to partition D. A is updated through ω-restricted decomposition using the updated Ω. The FACETS algorithm consists of two phases: the *initialization* phase (Fig. 2b), and the *iteration* phase (Fig. 2c–d) (see Supplementary Material). We describe each of them in turn.

Initialization. This phase creates an initial set of decompositions for the second phase. We perform graph clustering on G to obtain an initial set of modules. To this end, the FUSE (Seah *et al.*, 2012) algorithm, summarized in Supplementary Material, is utilized. Each module of this set is then randomly associated with a facet, randomly distributing the modules over an initial set of facets. Following that, we construct *candidates subnetworks*, which use subnetworks of G that satisfy ω_i-restricted decomposition constraint. To generate candidates exhaustively is prohibitively expensive. Instead, candidates for a facet F_i are generated as follows: for every GO term $\Delta \in \omega_i$, we obtain the induced subnetwork in G whose nodes are annotated with Δ or its descendants. The subnetwork is then decomposed into connected components, each forming a candidate subnetwork G_j. Let $\Delta_j^C = \Delta$ be the term associated with this candidate. Candidates formed this way can vary greatly in resolution of the annotation that its nodes share (for example, $\Delta_j^C =$ `biological process`), and can be highly overlapping.

Iteration. This phase—the actual optimization phase—is performed in rounds. Let i denote the i-th iteration of the algorithm. At each round, the algorithm updates A and Ω in two sequential steps. To evaluate algorithm convergence, we introduce *functional reassignment*—the number of terms in Δ that is reassigned to a different function bundle after Step 1 of i-th iteration. This score measures the rate of change of Ω, indicating how close the algorithm is to convergence. Observe that when Ω is fixed, the algorithm reaches a steady state. The algorithm reaches convergence and terminates when the functional reassignment at i-th iteration drops below *convergence threshold* θ, a user-defined parameter.

Step 1. **Update** Ω. In this step, we assume that A is a constant and update Ω to increase $t(\Omega, A)$. For each $F_i \in A$, the enriched functional terms of the modules in F_i serve as centroids for partitioning D into orthogonal concepts; these enriched terms as whole form the centroid of ω_i, which is associated with F_i. We then reassign every candidate subnetwork to its nearest centroid to form a partition Ω. The convergence properties of such centroid-based partitioning approaches (e.g. K-means) has been well studied (Bottou and Bengio, 1994). For every $G_j \in B_i, 1 \le i \le n$, we determine its *closest* centroid by considering G_j's average functional and structural distance to functional modules within a facet. The facet that is closest to G_j is indicated by:

$$d_c(G_j, F_k) = \begin{cases} 1 & \text{if} \frac{1}{Z(F_k)} \sum\limits_{C_i \in F_k} \; s(\Delta_i^C, C_i)\,\phi(C_i, G_j) \\ & \le \frac{1}{Z(F_{k'})} \sum\limits_{C_i \in F_k{'}} \; s(\Delta_i^C, C_i)\,\phi(C_i, G_j) \\ & k' \neq k, \text{where} \\ & \phi(C_i, C_j) = \gamma \dfrac{d_f(\Delta_i^C, \Delta_j^C)}{|V_p^j||V_p^i|} + (1-\gamma) d_s(C_i, C_j) \\ & z(F) = \sum\limits_{C_i \in F} s(\Delta_i^C, C_i) \\ 0 & \text{otherwise} \end{cases}$$

Following that, G_j is reassigned to nearest facet candidate bundle B_k (superset of F_k) and Ω is updated based on where every $\Delta_j^C \in V_{\mathrm{GO}}$ is assigned to. Each function bundle $\omega_i \in \Omega$ represents functional terms that are most closely associated with F_i, and the decomposition of F_i in the following step will be bounded by the updated ω_i. Function partitioning depends on the atlas of decompositions because not every partition of the GO DAG is capable of forming a modular decomposition of functional modules.

Step 2. **Update** A. In this step, we update A to maximize the objective function while fixing Ω. To support ω_i-restricted decomposition of F_i, we propose an algorithm that employs profit maximization model (discussed below) and runs in iterations. At each iteration, we score candidate subnetworks based on a profit maximization model and greedily selects the best scoring candidate as member in F_i. An iteration runs for every $F_i \in A$ before moving to the next iteration. Every candidate considered for F_i must satisfy the ω_i-restricted decomposition constraint, i.e. the candidate subnetwork must be enriched with terms in ω_i. In other words, $G_j \in B_i$.

We now describe the profit maximization model for scoring a candidate $G_j \in B_i$. Every $v \in V$ is assigned an information budget. A candidate G_j extracts, from each $v \in V_j^G$, some information revenue from the budget pool. The revenue extracted is correlated to the edge density of the subnetwork, with modular candidates giving high revenue. Each time a candidate is selected, revenue is removed from the budget pool and a cost is incurred. A penalty cost is incurred for a candidate that is structurally similar to selected clusters in other facets $F_{i'} \neq F_i$. This penalty is modeled by $\text{cost}(G_j) = \sum_{C' \in F_{i'}, i' \neq i} d_s(G_j, C')$, which utilizes the structural distance measure d_s described earlier. At each iteration, the candidate that contributes the highest information profit (revenue minus cost) is selected. To summarize, a clustering in F_i that yields high overall revenue have subgraphs with high facet modularity $o(F_i)$, whereas a clustering with low overall cost yields high inter-facet orthogonality $t(\Omega, A)$. Given a fixed Ω, the set of facets A with maximum overall profit maximizes the objective function. The algorithm above approximates this through greedy heuristic.

3 RESULTS

3.1 Experiment settings

The FACETS algorithm is implemented in Scala (Odersky *et al.*, 2004). We now present the experiments conducted to study the performance of FACETS and report some of the results here. All experiments were executed on a 1.66 GHz Intel Core 2 Duo T5450 machine with 3GB memory. We primarily used the global human PPI network from *IntAct* (Kerrien *et al.*, 2007), as well as the *yeast*, *fruit fly*, and *human autophagy* networks from *IntAct* (Table 1). In all experiments, we set the convergence threshold $\theta = 5$. The weight γ is set to 0.091 to balance the contribution of structure and function (equal order of magnitude). We utilize only the `cellular process` sub-domain of the GO so that the facets are created not merely based on different GO domains, but on more subtle functional differences.

Table 1. Datasets used

Dataset	No of nodes	No of edges	Source
H. sapiens	9131	34 362	IntAct (Kerrien *et al.*, 2007)
S. cerevisiae	4768	40 457	IntAct
D. melanogaster	3114	6472	IntAct
Human autophagy	1241	3555	IntAct

3.1.1 Evaluation criteria To measure the similarity/dissimilarity between facets or decompositions, we employed the *Jaccard index* (JI) (Ben-Hur *et al.*, 2002) evaluation measure, which is widely used to compare clusterings based on counting the agreement or disagreement of co-clustered pairs of proteins. The reader may refer to Supplementary Material for definitions of the measure.

3.2 Experiment results

3.2.1 Quantitative assessment Table 2 shows the quantitative comparison between facets. We measure the inter-facet decomposition similarity using the JI score. The low clustering similarity scores between facets show that they are decomposed distinctively. This reflects significant organizational differences between modules of signaling pathways and modules of protein complexes. We measured the *coverage* of a facet and the *extent* of coverage overlap between the facets. Let the coverage of a facet F_k be $Cvg(F_k) = |\bigcup_{V_c \in F_k} V_c|$. Also, let the extent of coverage overlap between the F_i and F_j be $Ext(F_i, F_j) = \frac{|V_i \cap V_j|}{|V_i|}$, where $V_i = \bigcup_{V_c \in F_i} V_c$ and $V_j = \bigcup_{V_c \in F_j} V_c$. The extent of overlap between facets reaches up to 0.316. Consequently, the overlap is not insignificant, implying that the facets are not partitions of G.

3.2.2 Validation on real data In this experiment, we compare the FACETS atlases of the global human network to gold standard functional modules. The gold standard datasets were constructed as follows: (i) `MIPS`—We use the set of 571 human complexes (of more than three proteins) from MIPS (Mewes *et al.*, 2002) to represent the decomposition of the human interactome into complexes. (ii) `KEGG-metabolic`—To represent decomposition into metabolic modules, we use 67 human

Table 2. Comparison between facets of the *H. sapiens* PPI network ($n = 6$)

			JI score						Coverage overlap					
Facets	No of modules	Coverage	Facet 1	Facet2	Facet 3	Facet 4	Facet 5	Facet 6	Facet 1	Facet 2	Facet 3	Facet 4	Facet 5	Facet 6
1	89	294	1.0	0.014	0.065	0.0050	0.0070	0.079	1.0	0.316	0.142	0.081	0.044	0.112
2	280	1079	0.014	1.0	0.0040	0.119	0.0050	0.0070	0.086	1.0	0.077	0.09	0.082	0.079
3	106	372	0.065	0.0040	1.0	0.0010	0.0	0.013	0.112	0.225	1.0	0.029	0.059	0.086
4	94	419	0.0050	0.119	0.0010	1.0	0.0	0.0080	0.057	0.233	0.026	1.0	0.028	0.052
5	114	390	0.0070	0.0050	0.0	0.0	1.0	0.0010	0.033	0.228	0.056	0.03	1.0	0.038
6	72	306	0.079	0.0070	0.013	0.0080	0.0010	1.0	0.107	0.281	0.104	0.071	0.049	1.0

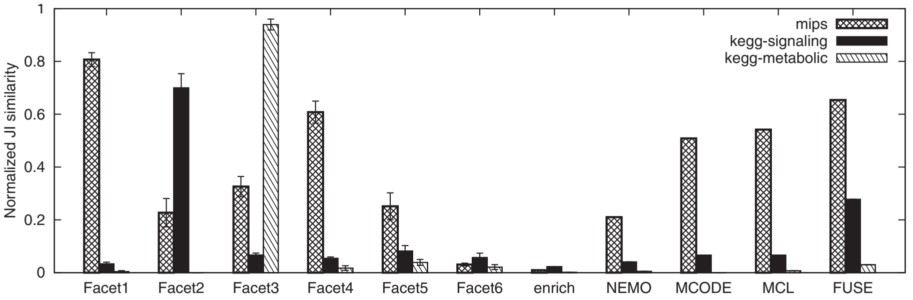

Fig. 3. Comparison between the decomposition similarities of FACETS, other methods and gold standard decompositions

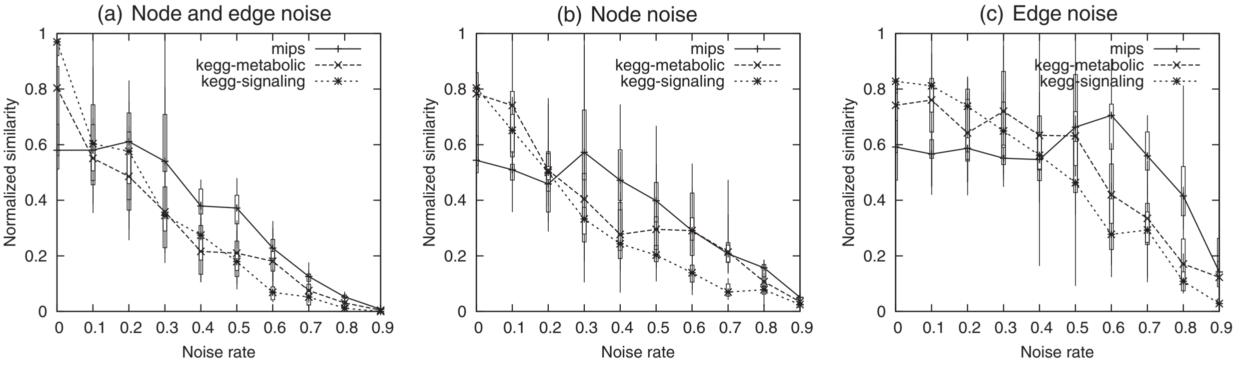

Fig. 4. Effect of noise on FACETS algorithm

metabolic networks from KEGG, each representing a single functional module. (iii) KEGG-signaling—We use 23 human signal transduction pathways from KEGG to represent decomposition into signaling pathways. The gold standard decompositions were chosen such that each represents a distinct functional organization of the human network. As such, we consider each gold standard dataset as a facet of the human network, and the set of these three datasets as the gold standard atlas of the human network. We then compared these datasets against the atlas of facets obtained through our algorithm and determine if there is a distinctive one-to-one mapping between our facet and a gold standard facet. We set $n = 6$ and repeated the tests fifteen times under different starting conditions to account for variability in facets output. We also compare the similarity scores against graph clustering methods, namely Markov clustering (MCL) (Krogan *et al.*, 2006), mcode (Bader and Hogue, 2003), nemo (Rivera *et al.*, 2010) and fuse (Seah *et al.*, 2012). These methods create a single decomposition of the human network. We removed clusters with fewer than three proteins. We also compare against GO term enrichment (enrich) (Boyle *et al.*, 2004), which does not utilize structural information. Following that, we measure the clustering similarities between the gold standard datasets and the decompositions obtained. Figure 3 shows the clustering similarities between modules in gold standard datasets and modules in facets as well as tested graph clustering methods. The JI was used to measure the agreement between pairs of decompositions. We normalize the scores so that the highest JI score obtained, within each gold standard dataset, is adjusted to 1.

We consider the facet best associated with a gold standard decomposition by comparing their relative scores. The gold standard datasets were uniquely mapped to a distinct facet: KEGG-metabolic is most similar to *Facet 3*, KEGG-signaling is most similar to *Facet 2* and MIPS is most similar to *Facet 1*. This unique mapping demonstrates that from a clustering perspective, the facets have significant functional orthogonality such that they are uniquely associated with different functional organization maps. *Facet 6* has poor similarity to the gold datasets, indicating a set of clusters that could be functionally distinct from these datasets.

In contrast, the tested graph clustering methods share common similarity patterns. Clusters are largely from a single dominant perspective—those of protein complexes (MIPS). We argue that objective functions based on dense connectivity tend to favor protein complex structures over other decompositions like metabolic pathways. GO term enrichment, on the other hand, generates output with little similarity to all gold standard datasets, indicating that annotations alone are unable to specifically identify important functional modules within a large PPI network. This is supported by the fact that functional analysis of large networks often involve graph clustering prior to term enrichment (Krogan *et al.*, 2006).

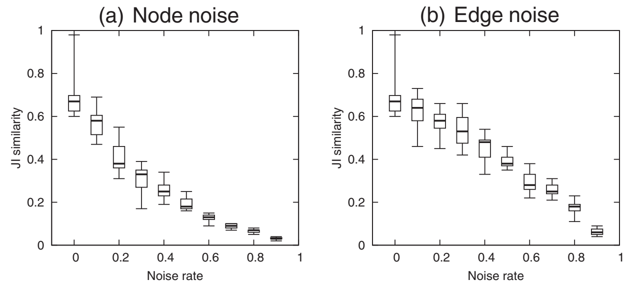

Fig. 5. Effect of initial starting point versus noise on FACETS algorithm

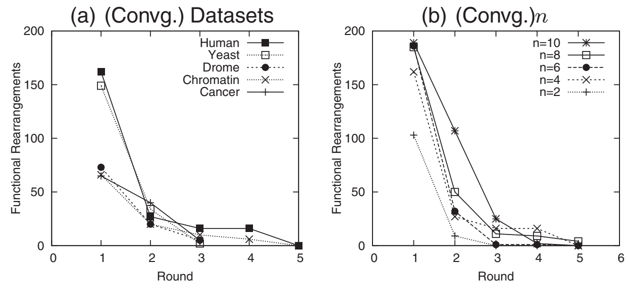

Fig. 6. Rate of convergence of FACETS algorithm

3.2.3 Robustness To study the robustness of FACETS, we tested the effect of annotation perturbations and edge deletions of the input network on FACETS output. Random edge deletion (*edge noise*) simulates the effect of removing false positive interactions in high-throughput interaction datasets, whereas annotation perturbation (*node noise*) simulates errors in curated annotations. Figure 4(a–c) shows the effect of edge and node noise on FACETS, varying from 0 to 100% noise. The figure shows clustering similarities (JI similarity) between the best scoring facets and gold standard datasets under increasing noise perturbations. We repeated each test fifteen times with different randomization seed. We observed that FACETS output quality drops gradually under increasing edge and node noise conditions. This demonstrates that the algorithm is robust to small noise perturbations. In case of edge noise, we noted that the quality of output only drops rapidly past the 0.5 noise ratio. This is desirable given that false-positive rates in yeast two-hybrid and TAP experiments range between 0.35 to 0.7 (Hart *et al*., 2006). MIPS clusters, which consist of densely interconnected clusters, are most robust to edge noise effects. The effect of node noise is comparatively greater, but quality degradation remains gradual.

3.2.4 Effect of initial starting point Given that FACETS belong to the class of hill-climbing methods, the algorithm output is dependent on the initial starting point. To this end, we study the effects of multiple random initial starting points. We compared the variability in clustering output due to starting point versus variability due to noise effects to give a sense of the magnitude of variability. We set a single facet output as the reference output, and compared its JI similarity with outputs from different starting points and increasing noise effects. The boxplot Figure 5(a and b) shows the effect of initial starting point versus noise on facets. At 0 noise rate, the variability in JI similarity is due to initial starting point. Given the fact that high-throughput datasets are inherently noisy (as mentioned above), the variability due to starting point is less significant. In addition, Figure 4(a–c) shows the effect of starting points with respect to gold standard datasets when one observes the similarity at 0 noise rate.

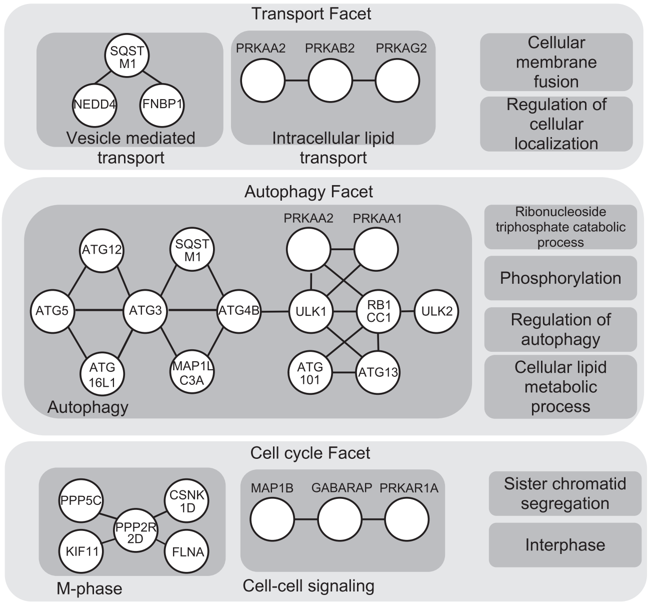

Fig. 7. Multiple facets (subset) illustrating the functional organization of the human autophagy network under different perspectives

3.2.5 Convergence Figure 6(a and b) shows the functional reassignments after the *i*-th iteration. We conducted the tests on varying types of datasets with $n = 6$. We also vary the number of facets per atlas ($n = 2$–6) on the global human network. All tests converge in <9 rounds, demonstrating FACETS' ability to converge quickly to a solution. Larger datasets such as the human network requires more iterations to complete. The number of iterations required also tends to increase with the number of facets n.

3.2.6 Case study: human autophagy system To illustrate the utility of multi-faceted decomposition, we analyze the functional organization human autophagy system. The functional map of this system was manually constructed in (Behrends *et al*., 2010). We generated the facets of the human autophagy network ($n = 6$), and a subset of the results is shown in Figure 7. The automatically generated facets show the pertinent roles of vesicle transport and lipid membrane metabolism in autophagy, which is consistent with the manually constructed map. Additionally, the network can also be clustered from the perspective of cell cycle and apoptosis regulation modules, which is not depicted in the manual map. This demonstrates the possibility of having multiple perspectives that organize a network.

4 CONCLUSION

In this article, we propose FACETS, a data-driven and generic algorithm for generating multi-faceted functional decompositions of a PPI network, providing multiple perspectives of the functional organization landscape of the network. Our experimental validation with real-world PPI networks demonstrates effectiveness of FACETS in generating functionally distinctive facets. As future work, we intend to extend FACETS to evaluate both annotated and unannotated regions of the PPI network.

Conflict of Interest: none declared.

REFERENCES

Agarwal,G. and Kempe,D. (2008) Modularity-maximizing graph communities via mathematical programming. *Eur. Phys. J. B.*, **66**, 409–418.

Bader,G. and Hogue,C. (2003) An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, **4**, 2.

Behrends,C. *et al.* (2010) Network organization of the human autophagy system. *Nature*, **466**, 68–76.

Ben-Hur,A. *et al.* (2002) A stability based method for discovering structure in clustered data. *Pac. Symp. Biocomput*, World Scientific, Lihue, Hawaii, USA, 6–17.

Boyle,E. *et al.* (2004) GO:TermFinder–open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics*, **20**, 3710–3715.

Bottou,L. and Bengio,Y. (1994) Convergence properties of the k-means algorithms. In *Advances in Neural Information Processing Systems 7 (NIPS'94)*, MIT Press, Denver, Colorado, USA.

Brohée,S. and van Helden,J. (2006) Evaluation of clustering algorithms for protein–protein interaction networks. *BMC Bioinformatics*, **7**, 488.

Caruana,R. *et al.* (2006) Meta clustering. In *Proceedings of the Sixth IEEE International Conference on Data Mining*, IEEE Computer Society, Hong Kong, China.

Cui,Y. *et al.* (2007) Non-redundant multi-view clustering via orthogonalization. In *Proceedings of the Seventh IEEE International Conference on Data Mining*, IEEE Computer Society, Omaha, Nebraska, USA.

Hart,G.T. *et al.* (2006) How complete are current yeast and human protein-interaction networks? *Genome Biol*, **7**, 120.

Jagota,A. (1995) Approximating maximum clique with a Hopfield network. *IEEE Trans. Neural Netw.*, **6** (3), 724–735.

Kerrien,S. *et al.* (2007) IntAct-open source resource for molecular interaction data. *Nucleic Acids Res.*, **35** (Database issue), D561–5.

Krogan,N.J. *et al.* (2006) Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature*, **440** (7084), 637–43.

Lavallée-Adam,M. *et al.* (2009) Detection of locally over-represented GO terms in protein–protein interaction networks. In *Proceedings of the Thirteenth Annual International Conference on Research in Computational Molecular Biology*, Springer, Tucson, AZ, USA.

Massen,C.P. and Doye,J.P. (2005) Identifying communities within energy landscapes. *Phys. Rev. E*, **71** (4), 046101.

Mewes,H.W. *et al.* (2002) MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.*, **30** (1), 31–4.

Navlakha,S. and Kingsford,C. (2010) Exploring biological network dynamics with ensembles of graph partitions. *Pac. Symp. Biocomput*, World Scientific, Fairmont Orchid, Hawaii, USA, 166–177.

Niu,D. and Dy,J.G. (2010) Multiple non-redundant spectral clustering views. *27th International Conference on Machine Learning (ICML 2010)*, Omnipress, Haifa, Israel, 831–838.

Odersky,M. *et al.* (2004) An overview of the Scala programming language. *EPFL Technical Report IC/2004/64*, N/A (technical report), Denver, Colorado, USA.

Qi,Z. and Davidson,I. (2009) A principled and flexible framework for finding alternative clusterings. In *Proceeding of the Fifteenth ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, ACM, Paris, France.

Rivera,C.G. *et al.* (2010) Network module identification in Cytoscape. *BMC Bioinformatics*, **11** (Suppl. 1), S61.

Seah,B.S. *et al.* (2012) FUSE: towards multi-level functional summarization of protein interaction networks. *BMC Bioinformatics*, **13** (Suppl. 3), S10.

Wagstaff,K. and Cardie,C. (2000) Clustering with instance-level constraints. *Proceedings of the Seventeenth International Conference on Machine Learning*, Morgan Kaufmann, Stanford, California, USA.