Connectivity Inference in Mass Spectrometry based Structure Determination

Deepesh Agarwal, Julio Araujo, Christelle Caillouet, Frédéric Cazals, David Coudert, Stéphane Pérennes

Connectivity Inference in Mass Spectrometry based Structure Determination

D. Agarwal, J. Araujo, C. Caillouet, F. Cazals, D. Coudert, S. Pérennes

RESEARCH REPORT
N° 8320
June 2013
Project-Team ABS and COATI

ISSN 0249-6399   ISRN INRIA/RR--8320--FR+ENG

Connectivity Inference in Mass Spectrometry based Structure Determination

D. Agarwal*, J. Araujo*†, C. Caillouet*†, F. Cazals*‡,
D. Coudert*†‡, S. Pérennes*†

Project-Team ABS and COATI

Research Report n° 8320 — June 2013 — 20 pages

Abstract: We consider the following Minimum Connectivity Inference problem (MCI), which arises in structural biology: given vertex sets $V_i \subseteq V, i \in I$, find the graph $G = (V, E)$ minimizing the size of the edge set E, such that the sub-graph of G induced by each V_i is connected. This problem arises in structural biology, when one aims at finding the pairwise contacts between the proteins of a protein assembly, given the lists of proteins involved in sub-complexes. We present four contributions.

First, using a reduction of set cover, we establish that MCI is APX-hard. Second, we show how to solve the problem to optimality using a mixed integer linear programming formulation (`MILP`). Third, we develop a greedy algorithm based on union-find data structures (`Greedy`), yielding a $2(\log_2 |V| + \log_2 \kappa)$-approximation, with κ the maximum number of subsets V_i a vertex belongs to. Fourth, application-wise, we use the MILP and the greedy heuristic to solve the aforementioned connectivity inference problem in structural biology. We show that the solutions of `MILP` and `Greedy` are more parsimonious than those reported by the algorithm initially developed in biophysics, which are not qualified in terms of optimality. Since MILP outputs a set of optimal solutions, we introduce the notion of *consensus solution*. Using assemblies whose pairwise contacts are known exhaustively, we show an almost perfect agreement between the contacts predicted by our algorithms and the experimentally determined ones, especially for consensus solutions.

Key-words: Connectivity Inference Connected induced sub-graphs, network design, APX-hard, Mixed integer linear program, Greedy algorithm, Mass spectrometry, Protein assembly, Structural biology, Biophysics, Molecular machines

* INRIA Sophia-Antipolis - Méditerranée

† Univ. Nice Sophia Antipolis, CNRS, I3S, UMR 7271, 06900 Sophia Antipolis, France

‡ Correspondence to Frederic.Cazals@inria.fr or to David.Coudert@inria.fr

RESEARCH CENTRE
SOPHIA ANTIPOLIS – MÉDITERRANÉE

2004 route des Lucioles - BP 93
06902 Sophia Antipolis Cedex

Inférence de la connectivité pour la détermination de structure en spectrométrie de masse

Résumé : Nous considérons le problème d'INFÉRENCE DE CONNECTIVITÉ MINIMALE (MINIMUM CONNECTIVITY INFERENCE ou MCI) qui se pose en biologie structurale: étant donnés des ensembles de sommets $V_i \subseteq V, i \in I$, trouver le graphe $G = (V, E)$ minimisant la taille de l'ensemble des arêtes E, de telle sorte que le sous-graphe de G induit par chaque ensemble V_i soit connexe. Ce problème se pose en biologie structurale pour la determination des contacts plausibles entre les protéines d'un assemblage à partir des listes de protéines présentes dans des sous-complexes. Nous présentons quatre contributions.

Premièrement, nous montrons que le problème MCI est APX-hard en utilisant une réduction de *set cover*. Deuxièmement, nous présentons une formulation en programme linéaire mixte (`MILP`) permettant de résoudre MCI de façon optimale. Troisièmement, nous proposons un algorithme glouton (`Greedy`) basé sur des structures de données *Union-Find*. Nous montrons que cet algorithme est une $2(\log_2 |V| + \log_2 \kappa)$-approximation de l'optimal, où κ est le nombre maximum d'ensembles V_i contenant un sommet donné. Quatrièmement, d'un point de vue appliqué, nous utilisons l'approche MILP et l'algorithme glouton pour résoudre le problème MCI en biologie structurale. Nous montrons que les solutions calculées par `MILP` et `Greedy` sont plus parcimonieuses que celles produites par l'algorithme utilisé à ce jour en bio-physique — lequel n'est pas qualifié en terme d'optimalité. Les algorithmes `MILP` et `Greedy` générant des ensembles de solutions, nous introduisons la notion de *solution consensus*. En utilisant le cas d'assemblages dont les contacts sont connus de façon exhaustive, nous montrons un accord presque parfait entre les contacts determinés par nos algorithmes et ceux determinés expérimentalement, en particulier pour les solutions consensus.

Mots-clés : Inférence de la connectivité, Sous-graphe induit connexe, APX-hard, programme linéaire mixte, algorithme glouton, spectrométrie de masse, assemblage protéique, biologie structurale, biophysique, machine moléculaire

Contents

1 Introduction

1.1 Connectivity Inference for Macro-molecular Assemblies

Macro-molecular assemblies. Building models of macro-molecular machines is a key endeavor of biophysics, as such models not only unravel fundamental mechanisms of life, but also offer the possibility to monitor and to fix defaulting systems. Example of such machines are the eukaryotic initiation factors which initiate protein synthesis by the ribosome, the ribosome which performs the synthesis of a polypeptide chain encoded in a messenger RNA derived from a gene, chaperonins which help proteins to adopt their 3D structure, the proteasome which carries out the elimination of damaged or misfolded proteins, etc. These macro-molecular assemblies involve from tens to hundreds of molecules, and range in size from a few tens of Angstroms (the size of one atom) up to 100 nanometers.

But if atomic resolution models of small assemblies are typically obtained with X-ray crystallography and/or nuclear magnetic resonance, large assemblies are not, in general, amenable to such studies. Instead, their reconstruction by *data integration* requires mixing a panel of complementary experimental data [4]. In particular, information on the hierarchical structure of an assembly, namely its decomposition into sub-complexes (complexes for short in the sequel) which themselves decompose into isolated molecules (proteins or nucleic acids) can be obtained from mass spectrometry.

Mass spectrometry. Mass spectrometry (MS) is an analytical technique allowing the measurement of the mass-to-charge (m/z) ratio of molecules [22], based on three devices, namely a source to produce ions from samples in solution, an analyzer separating them according to their m/z ratio, and a detector to count them. The process results in a m/z spectrum, whose deconvolution yields a mass spectrum, i.e. an histogram recording the abundance of the various complexes as a function of their mass. Considering this spectrum as raw data, two mathematical questions need to be solved. The first one, known as stoichiometry determination (SD), consists of inferring how many copies of the individual molecules are needed to account for the mass of a mode of the spectrum [6, 2]. The second one, known as connectivity inference, aims at finding the most plausible connectivity of the molecules involved in a solution of the SD problem.

Connectivity inference. Given a macro-molecular assembly whose individual molecules (proteins or nucleic acids) are known, we aim at inferring the connectivity between these molecules. In other words, we are given the vertices of a graph, and we wish to figure out the edges it should have. To constrain the problem, we assume that the composition, in terms of individual molecules, of selected complexes of the assembly is known. Mathematically, this means that the vertex sets of selected *connected subgraphs* of the graph sought are known. To see where this information comes from, recall that a given assembly can be chemically denatured i.e. split into complexes by manipulating the chemical conditions prior to ionization. In extreme conditions, complete denaturation occurs, so that the individual molecules can be identified using MS. In milder conditions, multiple overlapping complexes are generated: once the masses of the proteins are known, the list of proteins in each such complex is determined by solving the aforementioned SD problem [20]. As a final comment, it should be noticed that in inferring the connectivity, *smallest-size networks* (i.e. graphs with as few edges as possible) are sought [3, 25]. Indeed, due to volume exclusion constraints, a given protein cannot contact all the remaining ones, so that the minimal connectivity assumption avoids speculating on the exact (unknown) number of contacts.

Mathematical Model. Let $G = (V, E)$ be a graph, where V is the set of vertices and E the set of edges. We denote $G[V']$, respectively $G[E']$, the subgraph of G induced by $V' \subseteq V$, resp. by $E' \subseteq E$.

Consider an assembly together with the list of constituting proteins, as well as a list of associated complexes. Prosaically, we associate to each protein a vertex $v \in V$ and to each complex $i \in I \subseteq \mathbb{N}$ a subset $V_i \subseteq V$, such that if the protein v belongs to the complex i, then $v \in V_i$. Our goal is to infer the connectivity inside each complex of proteins. Therefore, we need to select a set of edges E_i between the vertices of V_i such that the graph $G_i = (V_i, E_i)$ is connected. The Minimum Connectivity Inference problem is to find a graph $G = (V, E)$ with minimum cardinality set of edges E such that the subgraph $G[V_i]$ induced by each V_i, $i \in I$, is connected. Formally, we state the problem as follows.

Definition 1 (Minimum Connectivity Inference problem, MCI).

Inputs: *A set V of n vertices (proteins) and a set of subsets (complexes) $C = \{V_i \mid V_i \subseteq V \text{ and } i \in I\}$.*

Constraint: *A set E of edges is* feasible *if $G[V_i] \subseteq G = (V, E)$ is connected, for every $i \in I$.*

Output: *A feasible set of edges E with minimum cardinality.*

Related work. The connectivity inference problem was first addressed in [25] using a two-stage algorithm, called *network inference* (`NI` in the sequel). First, random graphs meeting the connectivity constraint are generated, by incrementally adding edges at random. Second, a genetic algorithm is used to reduce the number of edges, and also boost the diversity of the connectivity. Once the average size of the graphs stabilizes, the pool of graphs is analyzed to spot highly conserved edges.

From the Computer Science point of view, MCI is a network design problem in which one wants to choose a set of edges with minimum cost to connect entities (e.g., routers, antennas, etc.) subject to particular connectivity constraints. Typical examples of such constraints are that the subgraph must be k-connected, possibly with minimum degree or maximum diameter requirements (see [19] for a survey). Such network design problems are generally hard to solve. To the best of our knowledge, the problem of ensuring the connectivity for different subsets of nodes has not been addressed before.

2 Preliminaries and Hardness

2.1 Simplifying an Instance of MCI: Reduction Rules

Let (V, C) be an instance of MCI. We denote by $(V, C) \setminus u$ the instance (V', E') of MCI obtained from (V, C) by removing u from V and all the subsets of C it belongs to. So we have $V' = V \setminus \{u\}$ and $C' = \{V_i \setminus \{u\} \mid V_i \in C \text{ and } i \in I\}$. Moreover, we denote by $\mathrm{OPT}((V, C))$ the cardinality of an optimal solution of MCI for the instance (V, C). Let us now denote $C(v) = \{i \mid V_i \ni v\} \subseteq I$, the set of complexes of the protein $v \in V$. We observe that we can apply the following reduction rules to any instance of MCI:

Lemma 1 (Reduction Rules). *Let (V, C) be an instance of MCI.*

1. *If $V_i \in C$ is such that $|V_i| = 1$, then any feasible solution for $(V, C \setminus V_i)$ is also feasible for (V, C), and we have $OPT((V, C \setminus V_i)) = OPT((V, C))$;*

2. If $C(u) \subseteq C(v)$, for some $u, v \in V$, then a feasible solution for (V, C) is obtained from a feasible solution for $(V, C) \setminus u$ by adding the edge uv, and we have $OPT((V, C)) = OPT((V, C) \setminus u) + 1$;

The proof is provided in the technical report [1].

By applying Lemma 1, we conclude that we can reduce the input instances of MCI to instances where every subset V_i has at least two vertices, every vertex appears in at least two subsets V_i and V_j with $i \neq j$, and the sets $C(u)$ and $C(v)$ are different, for any two vertices u and v.

2.2 Hardness

We establish that MCI is APX-hard, by showing a reduction of the Set Cover problem. The Set Cover problem is defined as follows:

Definition 2 (Set Cover problem)**.**

Inputs: *a ground set $\mathcal{X} = \{x_1, \ldots, x_m\}$, a collection $\mathcal{F} = \{X_i \subseteq \mathcal{X},\ i \in I\}$ and a positive integer k.*

Question: *does there exist $J \subseteq I$ such that $\bigcup_{i \in J} X_i = X$ and $|J| \leq k$?*

It is well-known that the Set Cover problem is NP-complete [13] and that this problem cannot be approximated in polynomial-time by a factor of $\ln n$, unless $P = NP$ [17, 5]. In order to prove our NP-completeness result, let us formally define the decision version of MCI as:

Definition 3 (Decision version of the Connectivity Inference problem, CI)**.**

Inputs: *A set of vertices V, a set of subsets $C = \{V_i \mid V_i \subset V$ and $i \in I\}$ and a positive integer k.*

Constraint: *A set E of edges is* feasible *if $G[V_i] \subseteq G = (V, E)$ is connected, for every $i \in I$.*

Question: *Does there exists a feasible set E of edges such that $|E| \leq k$?*

Theorem 1. *The decision version of the* Connectivity Inference *problem is NP-complete.*

The proof is provided in the technical report [1].

From the reduction used in the proof of Theorem 1 and the previous results on Set Cover problem [17, 5], we conclude that MCI is APX-hard:

Corollary 1. *There exists a constant $\mu > 0$ such that approximating MCI within $1 + \mu$ is NP-hard.*

3 Solving the Problem to Optimality using Mixed Integer Linear Programming

3.1 Flow Based Formulation

To solve an instance (V, C) of the MCI problem, we introduce one binary variable y_e for each edge $e = uv$ of the undirected complete graph on $|V|$ vertices $K_{|V|}$, to determine whether edge e is selected in the solution. Thus, the objective function consists of minimizing the sum of the y variables, as specified by Eq. (1). To solve this problem, we form the directed graph $D = (V, A)$

in which each edge $e = uv$ of the complete graph $K_{|V|}$ is replaced by two directed arcs (u, v) and (v, u). The solution using MILP satisfies the following constraints:

▷ *Connectivity constraints.* To enforce the connectivity of each complex, we select one vertex s_i per subset $V_i \in C$ as the source of a flow that must reach all other vertices in V_i using only arcs in $D[V_i]$. We introduce continuous variables $f_a^i \in \mathbb{R}^+$ to express the quantity of flow originating from s_i and circulating along the arc $a = (u, v)$ (i.e. from node u to v), with $u, v \in V_i$. Constraint (2), the flow conservation constraint of Eq. (2), expresses that $|V_i| - 1$ units of flow are sent from s_i, and each vertex u_i collects 1 unit of flow from s_i and forwards the excess it has received from s_i to its neighbors in $D[V_i]$.

▷ *Capacity constraints.* We also introduce a continuous variable $x_a \in [0, 1]$, with $a = (u, v) \in A$ and $u, v \in V$, that is strictly positive if arc a carries some flow and 0 otherwise. In other words, no flow can use arc a when $x_a = 0$ as ensured by Constraint (3).

▷ *Symmetry constraints.* If there is some flow on arc (u, v) or (v, u) in D, then variable x is strictly positive and so the corresponding edge uv must be selected in the solution, meaning that $y_e = 1$, as ensured by Constraints (4) and (5).

Denoting $\mathcal{E}$ the edges of the complete graph $K_{|V|}$, and $A_i^+(u)$ (resp. $A_i^-(u)$) the subset of arcs of $D[V_i]$ entering (resp. leaving) node u, the formulation reads as:

$$\min \quad \sum_{e \in \mathcal{E}} y_e \tag{1}$$

$$\text{s.t.} \sum_{a \in A_i^+(u)} f_a^i - \sum_{a \in A_i^-(u)} f_a^i = \begin{cases} |V_i| - 1 & \text{if } u = s_i \\ -1 & \text{if } u \neq s_i \end{cases} \qquad \forall u \in V_i,\ V_i \in C \tag{2}$$

$$f_a^i \leq |V_i| \cdot x_a, \qquad \forall\, V_i \in C, a \in A \tag{3}$$

$$x_{(u,v)} \leq y_{uv}, \qquad \forall\, uv \in \mathcal{E} \tag{4}$$

$$x_{(v,u)} \leq y_{uv}, \qquad \forall\, uv \in \mathcal{E} \tag{5}$$

Observe that this formulation can be turned into a decision formulation, by removing the objective and adding the constraint of Eq. (6). If the formulation becomes infeasible, the optimal solution as more than k edges.

$$\sum_{e \in \mathcal{E}} y_e \leq k \tag{6}$$

$$\sum_{e \in E_\ell} y_e < k \qquad \forall E_\ell \in \mathcal{S} \tag{7}$$

Moreover, we can use the decision formulation to enumerate all feasible solutions for an instance (V, C, k). To do so, we use Constraints (7), where $\mathcal{S}$ is the set of feasible solutions that have already been found. This constraint prevents finding twice a solution. We first set $\mathcal{S} = \emptyset$, then we add it to all newly found solutions and repeat until the problem becomes infeasible for a solution of size k.

3.2 Implementation

The formulation has been implemented using IBM CPLEX solver 12.1, the corresponding software being named MILP in the sequel. Starting from the complete graph of size $|V|$, MILP allows one to compute one optimal solution, or the set of all solutions involving at most $OPT + k$ edges. For $k = 0$, one gets the set of all optimal solutions, denoted $\mathcal{S}_{MILP}$ in the sequel.

4 Approximate Solution based on a Greedy Algorithm

4.1 Design and Properties

We now propose a greedy algorithm for MCI. Starting from the empty graph $G^0 = (V, E^0 = \emptyset)$, Algorithm 1 iteratively builds a graph $G^t = (V, E^t)$, with $E^t = E^{t-1} \cup \{e^t\}$. The edge $e^t = uv$ chosen at step t aims at reducing the number of connected components in the induced subgraphs $G^{t-1}[V_i]$ of G^{t-1}, for $i \in C(u) \cap C(v)$. More formally, at step t, we choose an edge e^t maximizing $m_t(e = uv)$ among all pairs $u, v \in V$, with $m_t(e = uv)$ the number of complexes containing u and v, and such that u and v belong to two different connected components of $G^{t-1}[V_i]$. The quantity $m_t(e = uv)$ is called the *priority* of the edge e.

Algorithm 1 Greedy algorithm for MCI

```
Require: $V = \{v_1, \ldots, v_n\}$ and $C = \{V_i \mid V_i \subseteq V$ and $i \in I\}$.
Ensure: A set $E$ of edges such that $G[V_i] \subseteq G = (V, E)$ is connected, for every $i \in I$.
$t := 1$, $E^0 := \emptyset$
while there exists a disconnected graph $G^{t-1}[V_i]$, for some $i \in I$ do
    Find edge $e^t$ maximizing the priority $m_t(e)$
    $E^t := E^{t-1} \cup \{e^t\}$ and $t := t + 1$
return $E_{t-1}$
```

Proposition 1. *Algorithm 1 is a* $2(\log_2 |V| + \log_2 \kappa)$*-approximation algorithm for MCI, with* κ *being the maximum number of subsets* V_i *a vertex belongs to.*

The proof is provided in the technical report [1].

Proposition 2. *When* $\max_{v \in V} |C(v)| = 2$*, Algorithm 1 always returns an optimal solution.*

The proof is provided in the technical report [1].

4.2 Implementation

In the following, we sketch an implementation of Algorithm 1, denoted `Greedy` in the sequel, which does not scan every candidate edge in E_t to find the (or a) best one, but instead maintains the priorities of all candidate edges.

Consider the following data structures:

- a priority queue Q associating to each candidate edge e its priority defined by $m_t(e)$. Note that the initial priority is given by $m_0(e = uv) = |C(u) \cap C(v)|$.
- a union-find data structure UF_i used to maintain the connected components of the induced graph $G^t[V_i]$. We assume in particular the existence of a function Find_vertices() such that UF_i.Find_vertices(u) returns the vertices of the connected component of the graph $G^t[V_i]$ containing the vertex u.

Upon popping the edge $e^t = (u, v)$ from Q, the following updates take place:

Update of the priority queue Q**.** For each complex V_i such that e^t triggers a merge between two connected components of $G^t[V_i]$, consider the two sets of vertices associated to these components, namely $K_{i,u} = UF_i$.Find_vertices(u) and $K_{i,v} = UF_i$.Find_vertices(v). The priority of all edges in the set $K_{i,u} \times K_{i,v} \setminus \{e^t\}$ is decreased by one unit.

Update of the union-find data structures. For each complex V_i such that e^t triggers a merge between two connected components of $G^{t-1}[V_i]$, the union operation $UF_i.\mathrm{Union}(UF_i.\mathrm{Find}(u), UF_i.\mathrm{Find}(v),)$ is performed.

It should be noticed that up to the logarithmic factor involved in the maintenance of Q, and up to the factor involving the inverse of Ackermann's function to run the union and find operations [24], the update complexity is output sensitive in the number of candidate edges affected in $K_{i,u} \times K_{i,v}$.

5 Experimental Results

5.1 Test Set: Assemblies of Interest

We selected three assemblies investigated by mass spectrometry, as explained in Section 1, for which we also found reference contacts between pairs of constituting proteins, against which to compare the output of our algorithms.

As explained in the supplemental section 8, we classified all collected contacts into three sets, namely crystal contacts (set $\mathrm{C}_{\mathrm{Xtal}}$) observed in high resolution crystal structures, cross-linking contacts (set C_{XL}), obtained by so-called cross-linking experiments, and miscellaneous dimers (set $\mathrm{C}_{\mathrm{Dim}}$) obtained by various biophysical experiments. In case the crystal contacts are not available, we define the reference set of contacts as $\mathrm{C}_{\mathrm{Exp}} = \mathrm{C}_{\mathrm{Dim}} \cup \mathrm{C}_{\mathrm{XL}}$. The three systems we selected are:

▷**Yeast Exosome.** The exosome is a 3'- 5' exonuclease assembly involved in RNA processing and degradation, composed of 10 different protein types with unit stoichiometry [10]. A total of 21 complexes were determined by mass spectrometry. (See also the supplemental Table 2 for the reference contacts.)

▷**Yeast 19S Proteasome lid.** Proteasomes are assemblies involved in the elimination of damaged or misfolded proteins, and the degradation of short-lived regulatory proteins. The most common form of proteasome is the 26S, which involves two filtering caps (the 19S), each cap involving a peripheral lid, composed of 9 distinct protein types each with unit stoichiometry [21]. A total of 14 complexes were determined by mass spectrometry. (See also the supplemental Table 3 for the reference contacts.)

▷**Eukaryotic Translation factor eIF3.** Eukaryotic initiation factors (eIF) are proteins involved in the initiation phase of the eukaryotic translation. They form a complex with the 40S ribosomal subunit, initiating the ribosomal scanning of mRNA. Among them, eIF3 consists of 13 different protein types each with unit stoichiometry [26]. A total of 27 complexes were determined by mass spectrometry. (See also the supplemental Table 4 for the reference contacts.)

5.2 Assessment Method

Let $\mathcal{S}_{MILP}$ be the set of optimal solutions returned by `MILP`, and let S_{NI} and S_{G} be the solutions computed by the algorithms `NI` [25] and `Greedy` respectively.

Consider an ensemble of solutions $\mathcal{S}$. The *size of a solution* $S \in \mathcal{S}$, denoted $\mid S \mid$, is its number of contacts. The *precision* of a solution S w.r.t. a reference set of contacts C is defined as the size of the intersection, i.e. $P_{\mathrm{MILP};C}(S) =\mid S \cap C \mid$. The precision is maximum if $S \subset C$, in which case no predicted contact is a false positive. The notion of precision makes sense if the reference contacts are exhaustive, which is the case for the exosome (since a crystal structure is known) and for the proteasome lid (exhaustive list of cross-links). We summarize the precision

Table 1 Size and precision of solutions. First section of the table: assembly, number of protein types, and size of the reference set C; second and third sections: size and precision for the solution returned by the algorithms `NI` [25] and `Greedy`; fourth and fifth sections: size and precision of algorithm `MILP`, for the whole set of optimal solutions $\mathcal{S}_{MILP}$, and for consensus solutions $\mathcal{S}^{cons.}_{MILP}$. NB: **The assignment of contacts was done manually [26]; NC^*: assembly not connected

Complex	#types	Ref. set C	$\lvert C\rvert$	$\lvert$ S$_{\text{NI}}$ $\rvert$	$P_{\text{MILP};C}(\text{S}_{\text{NI}})$	$\lvert$ S$_{\text{G}}$ $\rvert$	$P_{\text{MILP};C}(\text{S}_{\text{G}})$	$\lvert$ S$_{\text{MILP}}$ $\rvert$	$\lvert \mathcal{S}_{MILP} \rvert$	$P_{\text{MILP};C}(\mathcal{S}_{MILP})$	$\lvert \mathcal{S}^{cons.}_{MILP} \rvert$	$P_{\text{MILP};C}(\mathcal{S}^{cons.}_{MILP})$
Exosome	10	C_{Xtal}	26	12	12(100%)	10	10 (100%)	10	1644	(7, 9, 10)	12	(8, 9, 10)
19*S Lid*	9	C_{Exp}	16	9 $(NC)^*$	7(77.8%)	10	8 (80%)	10	324	(6, 7, 10)	18	(8, 8, 10)
eIF3	13	C_{Exp}	19	17**	14 (82.3%)	14	9 (64.2%)	14	2160	(8, 9, 11)	432	(8, 9, 10)

of the ensemble of solutions $\mathcal{S}$, denoted $P_{\text{MILP};C}(\mathcal{S})$, by the triple (min, median, max) of the precisions of the solutions $S \in \mathcal{S}$.

The *score of a contact* appearing in a solution is the number of solutions from $\mathcal{S}$ containing it, and its *signed score* is its score multiplied by ±1 depending on whether it is a true or false positive w.r.t C. The *score of a solution* $S \in \mathcal{S}$ is the sum of the scores of its contacts. Finally, a *consensus solution* is a solution achieving the maximum score over S, the set of all such solutions being denoted $\mathcal{S}^{cons.}$. Note that the score of a solution is meant to single out the consensus solutions from a solution set $\mathcal{S}$, while the signed score is meant to assess the solutions in $\mathcal{S}$ w.r.t a reference set.

5.3 Results

Except for the analysis of Table 1, due to the lack of space, we focus on the exosome (Fig. 1, and supplemental Table 2 for the reference contacts.).

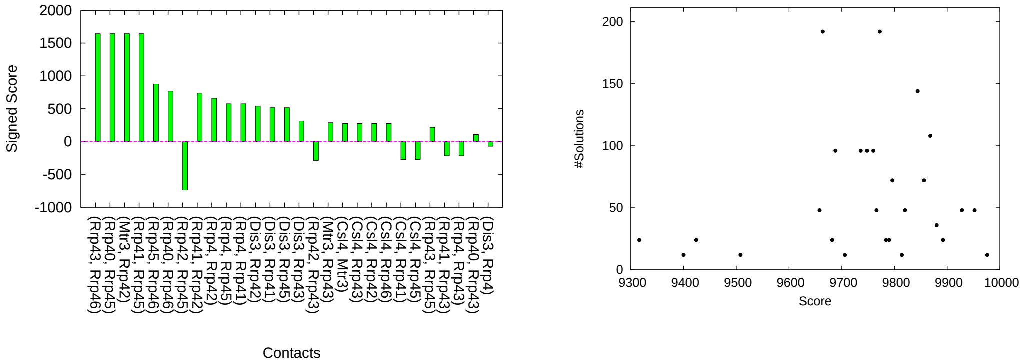
$\triangleright$**Parsimony and precision.** It is first observed that on the three systems, the algorithms `MILP` and `Greedy` are more parsimonious than `NI` (Table 1). For example, on the exosome, 10 edges are used instead of 12. The precision is excellent ($\geq$ 80%) for the three algorithms on the two systems where the reference set of contacts is exhaustive (exosome and lid).

$\triangleright$**Contact scores for $\mathcal{S}_{MILP}$ on the exosome.** Two facts emerge (Fig. 2(A)). First, four ubiquitous contacts are observed, while the remaining ones vary in the frequency. Second, there are few false positive overall. An interesting case is (Rrp42, Rrp45), which has the 7th highest count. The two polypeptide chains Rrp42 and Rrp45 are found in 16 out of 21 complexes used as input, accompanied in all cases by Rrp41. Interestingly, the point of closest approach between Rrp42 and Rrp45 in the crystal structure is circa 24Å, and this gap is precisely filled by Rrp41. That is, these three chains behave like a *rigid body.* Further inspection of the structure and of the behavior of `MILP` on such patterns is needed to explain why the edge (Rrp42, Rrp45) is reported.

$\triangleright$**Scores for consensus solutions on the exosome.** It is first observed that 12 consensus solutions amidst 1644 optimal ones are observed (Table 1 and Fig. 2(B)). In moving from $\mathcal{S}_{MILP}$ to $\mathcal{S}^{cons.}_{MILP}$, the precision increases from $(7, 9, 10)$ to $(8, 9, 10)$ — as also seen by a Pearson correlation coefficient of -0.51 between the mean false positive count per score, and the score (of a solution).

$\triangleright$**Overall assessment.** The consensus solutions from `MILP` are more parsimonious than those form `NI`, and compare favorably in terms of precision.

Figure 2 Exosome (A) Signed scores for contacts in $\mathcal{S}_{MILP}$, w.r.t $\mathrm{C}_{\mathrm{Xtal}}$ (B) Distribution of scores for solutions in $\mathcal{S}_{MILP}$



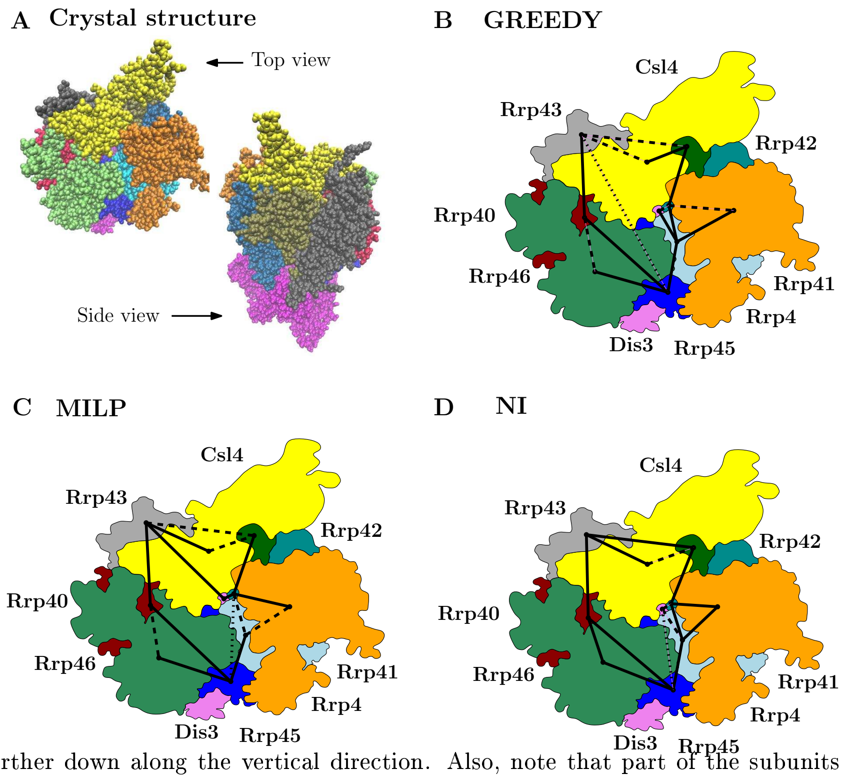


Figure. 1. Yeast Exosome: contacts computed by the algorithms. (A) Top and side view of the crystal structure [18, PDB 4IFD]. **(B,C,D)** Structure decorated with one edge per contact. the dash style reads as follows: *bold*: contacts in $S \cap \mathrm{C}_{\mathrm{Xtal}}$; *dotted*: contacts in S but not in $\mathrm{C}_{\mathrm{Xtal}}$; *dashed*: contact in $\mathrm{C}_{\mathrm{Xtal}}$ but not in S (Note that only most prominent contacts in $\mathrm{C}_{\mathrm{Xtal}}$ are shown to avoid cluttering). Note that a long edge i.e. an edge between two subunits that appear distant on the top view of the assembly corresponds to a contact of these subunits located further down along the vertical direction. Also, note that part of the subunits Dis3 and Rrp42 are visible in the middle of the assembly and are trapped in between Csl4, Rrp40, Rrp41. The contact node therefore is placed there for convenience.

6 Conclusion and Outlook

A key endeavor of biophysics, for macro-molecular systems involving up to hundreds of molecules, is the determination of the pairwise contacts between these constituting molecules. The corresponding problem, known as connectivity inference, is central in mass-spectrometry based studies, which over the past five years, has proved crucial to investigate large assemblies. In this context, this paper presents a thorough study of the problem, encompassing its hardness, a greedy strategy, and a mixed integer programming algorithm. Application-wise, the key advantage of our methods w.r.t. the algorithm *network inference* developed in biophysics, is that we fully master all optimal solutions instead of a random collection of solutions which are not qualified w.r.t. the optimum. As shown by careful experiments on three assemblies recently scrutinized by other bio-physical experiments (exosome, proteasome lid, eIF3), our predictions are in excellent agreement with the experimental contacts. We therefore believe that our algorithms should leverage the interpretation of protein complexes obtained by mass spectrometry, a research vein currently undergoing major developments.

From a theoretical standpoint, a number of challenging problems deserve further work. The first one is to understand the solution space as a function of the number of input vertex sets and the structure of the unknown underlying graph. This problem is also related to the (output-sensitive) enumeration of connected subgraphs of a given graph. The second challenge is concerned with the generalization where the stoichiometry (the number of instances) of the proteins involved is more than one. In that case, complications arise since the connectivity information associated to the vertex sets of the connected subgraphs is related to protein types, while the connectivity sought is between protein instances. This extension would allow processing cases such as the nuclear pore complex, the biggest assembly known to date in eukaryotic cells, as it involves circa 450 protein instances of 30 different protein types, some of them present in 16 copies. The third one is of geometric flavor, and is concerned with the 3D embedding of the graph(s) generated. Since the nodes represent proteins and since two proteins must form a bio-physically valid interface if they touch at all, information on the shape of the proteins could be used to find plausible embeddings that would constrain the combinatorially valid solutions. This would be especially helpful to recover the edges which are known from experiments, but do not appear in exact or approximate solution to the minimal connectivity problem. Finally, the Minimum Connectivity Inference problem also deserves investigation when the pool of candidate edges is a subset of the complete graph, which is especially relevant since pre-defined sets of edges may have been reported by a variety of experiments, some of them producing false positives.

References

[1] D. Agarwal, J. Araujo, C. Caillouet, F. Cazals, D. Coudert, and S. Pérennes. Connectivity inference in mass spectrometry based structure determination. Technical Report 8320, Inria, 2013.

[2] D. Agarwal, F. Cazals, and N. Malod-Dognin. Stoichiometry determination for mass-spectrometry data: the interval case. 2013. Submitted; available as INRIA tech report http://hal.inria.fr/hal-00741491.

[3] F. Alber, S. Dokudovskaya, L. M. Veenhoff, W. Zhang, J. Kipper, D. Devos, A. Suprapto, O. Karni-Schmidt, R. Williams, B.T. Chait, M.P. Rout, and A. Sali. Determining the Architectures of Macromolecular Assemblies. *Nature*, 450(7170):683–694, Nov 2007.

[4] F. Alber, F. Förster, D. Korkin, M. Topf, and A. Sali. Integrating Diverse Data for Structure Determination of Macromolecular Assemblies. *Ann. Rev. Biochem.*, 77:11.1–11.35, 2008.

[5] N. Alon, D. Moshkovitz, and S. Safra. Algorithmic construction of sets for k-restrictions. *ACM Trans. Algorithms*, 2:153–177, April 2006.

[6] S. Bocker and Z. Liptak. A fast and simple algorithm for the money changing problem. *Algorithmica*, 48(4):413–432, 2007.

[7] Q. Cai, A. Todorovic, A. Andaya, J. Gao, J. A. Leary, and J.H.D. Cate. Distinct regions of human eIF3 are sufficient for binding to the hcv ires and the 40s ribosomal subunit. *Journal of molecular biology*, 403(2):185–196, 2010.

[8] F. Cazals, F. Proust, R. Bahadur, and J. Janin. Revisiting the Voronoi description of protein-protein interfaces. *Protein Science*, 15(9):2082–2092, 2006.

[9] L. ElAntak, A. G. Tzakos, N. Locker, and P. J. Lukavsky. Structure of eIF3b rna recognition motif and its interaction with eIF3j. *Journal of Biological Chemistry*, 282(11):8165–8174, 2007.

[10] H. Hernández, A. Dziembowski, T. Taverner, B. Séraphin, and C.V. Robinson. Subunit architecture of multimeric complexes isolated directly from cells. *EMBO reports*, 7(6):605–610, 2006.

[11] J. Janin, R. P. Bahadur, and P. Chakrabarti. Protein-protein interaction and quaternary structure. *Quarterly reviews of biophysics*, 41(2):133–180, 2008.

[12] A. Kao, A. Randall, Y. Yang, V. R. Patel, W. Kandur, S. Guan, S. D. Rychnovsky, P. Baldi, and L. Huang. Mapping the structural topology of the yeast 19s proteasomal regulatory particle using chemical cross-linking and probabilistic modeling. *Molecular & Cellular Proteomics*, 2012.

[13] R. M. Karp. Reducibility among combinatorial problems. In *Complexity of Computer Computations*, The IBM Research Symposia Series, pages 85–103. Plenum Press, New York, March 1972.

[14] K. Lasker, F. Förster, S. Bohn, T. Walzthoeni, E. Villa, P. Unverdorben, F. Beck, R. Aebersold, A. Sali, and W. Baumeister. Molecular architecture of the 26s proteasome holocomplex determined by an integrative approach. *Proceedings of the National Academy of Sciences*, 109(5):1380–1387, 2012.

[15] E. Levy, E-B.Erba, C. Robinson, and S. Teichmann. Assembly reflects evolution of protein complexes. *Nature*, 453(7199):1262–1265, 2008.

[16] S. Loriot and F. Cazals. Modeling macro–molecular interfaces with intervor. *Bioinformatics*, 26(7):964–965, 2010.

[17] C. Lund and M. Yannakakis. On the hardness of approximating minimization problems. *J. ACM*, 41:960–981, September 1994.

[18] D. L. Makino, M. Baumgärtner, and E. Conti. Crystal structure of an rna-bound 11-subunit eukaryotic exosome complex. *Nature*, 495(7439):70–75, 2013.

[19] S. Raghavan. *Formulations and algorithms for network design problems with connectivity requirements.* PhD thesis, MIT, Cambridge, MA, USA, 1995.

[20] M. Sharon and C.V. Robinson. The role of mass spectrometry in structure elucidation of dynamic protein complexes. *Annu. Rev. Biochem.*, 76:167–193, 2007.

[21] M. Sharon, T. Taverner, X.I. Ambroggio, R.J. Deshaies, and C.V. Robinson. Structural organization of the 19s proteasome lid: insights from ms of intact complexes. *PLoS biology*, 4(8):e267, 2006.

[22] F. Stengel, R. Aebersold, and C. V. Robinson. Joining forces: integrating proteomics and cross-linking with the mass spectrometry of intact complexes. *Molecular & Cellular Proteomics*, 11(3), 2012.

[23] C. Sun, A. Todorovic, J. Querol-Audí, Y. Bai, N. Villa, M. Snyder, J. Ashchyan, C. S. Lewis, A. Hartland, S. Gradia, et al. Functional reconstitution of human eukaryotic translation initiation factor 3 (eIF3). *Proceedings of the National Academy of Sciences*, 108(51):20473–20478, 2011.

[24] R. E. Tarjan. *Data Structures and Network Algorithms*, volume 44 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 1983.

[25] T. Taverner, H. Hernández, M. Sharon, B.T. Ruotolo, D. Matak-Vinkovic, D. Devos, R.B. Russell, and C.V. Robinson. Subunit architecture of intact protein complexes from mass spectrometry and homology modeling. *Accounts of chemical research*, 41(5):617–627, 2008.

[26] M. Zhou, A. M. Sandercock, C. S. Fraser, G. Ridlova, E. Stephens, M. R. Schenauer, T. Yokoi-Fong, D. Barsky, J. A. Leary, J. W. Hershey, J. A. Doudna, and C. V. Robinson. Mass spectrometry reveals modularity and a complete subunit interaction map of the eukaryotic translation factor eIF3. *Proceedings of the National Academy of Sciences*, 105(47):18139–18144, 2008.

7 Appendix: Theory

7.1 Proofs

Proof for lemma 1:

Proof. To prove Statement 1, observe that V_i will always induce a connected subgraph, since it contains a single vertex. Consequently, it is trivially satisfied without the inclusion of any edge.

To prove Statement 2, observe first that given a feasible solution for $(V, C) \setminus u$, we can construct a feasible solution for (V, C) by adding the edge uv. Hence we have $\text{OPT}((V, C)) \leq \text{OPT}((V, C) \setminus u) + 1$.

Let us now construct a feasible solution for $(V, C) \setminus u$ from a feasible solution E for (V, C). Let $G = (V, E)$ be a graph and suppose first that $uv \notin E$. Let $P = uw_1 \dots w_p v$ be a uv-path in G, $p \geq 1$. Such a path exists since E is a feasible solution for (V, C) and $C(u) \subseteq C(v)$. So for every $i \in C(u)$, $G[V_i]$ is a connected subgraph containing both u and v. Observe that the edge uw_1 only appears in the subgraphs $G[V_i]$, for $i \in C(u) \cap C(w_1)$. Then we claim that the set E' obtained from E by removing edge uw_1 and adding edge uv is also a feasible solution for (V, C). In fact, the removal of uw_1 can only disconnect the subgraphs $G[V_i]$ for $i \in C(u)$. Moreover, the subgraphs $G[V_i]$ that become disconnected after the removal of uw_1 will have exactly two connected components, one containing u and the other containing v. Then, the addition of edge uv reconnects the disconnected subgraphs, and we have $|E| = |E'|$. Suppose now that $uv \in E$ and that there exists an edge $uw \in E$ with $w \neq v$. By using the previous argument, we construct the set E'' of edges from E by removing the edge uw and adding the edge vw. Obviously, E'' is a feasible solution for (V, C) and $|E| = |E''|$. Altogether, we can construct a feasible solution E^* for (V, C) such that $|E| = |E^*|$ and the only edge incident to u is uv. Furthermore, $E^* \setminus uv$ is a feasible solution for $(V, C) \setminus u$ since $C(u) \subseteq C(v)$, and so $\text{OPT}((V, C) \setminus u) \leq OPT((V, C)) - 1$. □ □

Proof for theorem 1:

Proof. Given a set E of edges, one can check in polynomial time whether $|E| \leq k$ and each induced subgraph $G[V_i]$ is connected, for every $i \in I$. Therefore the problem is in NP.

Let $I_{SC} = (\mathcal{X}, \mathcal{F}, k)$ be an input for the Set Cover problem. We construct an instance $I_{CI} = (V, C, k')$ to the decision version of the Connectivity Inference problem in such a way that I_{SC} is true if, and only if, I_{CI} is also true in the following way:

1. The vertex set V is partitioned into two sets A and B with $|A| = |B|$;
2. To each subset $X_i \in \mathcal{F}$, we associate a vertex $v_i \in A$ and a vertex $v'_i \in B$. So $|V| = 2|\mathcal{F}|$;
3. To each pair $v_i v_j$ (resp. $v'_i v'_j$) we associate a subset $V_{ij} = \{v_i, v_j\}$ (resp. $V_{i'j'} = \{v'_i, v'_j\}$) in C;
4. To each element $x_i \in \mathcal{X}$ we associate a subset $V_i \in C$, $i \in \{1, \dots, m\}$;
5. We have $x_i \in X_j$ if and only if we have $v_j, v'_j \in V_i$, for every $i \in \{1, \dots, m\}$ and $j \in \{1, \dots, |I|\}$;
6. We set $k' = k + 2\binom{|\mathcal{F}|}{2}$.

Observe that step 3 forces any feasible solution for CI to include all the edges $v_i v_j$ and $v'_i v'_j$, for every $i, j \in \{1, \dots, |\mathcal{F}|\}$. All the subsets V_{ij} and $V_{i'j'}$ of C are connected in any feasible solution of I_{CI}. Furthermore, each feasible solution has at least $2\binom{|\mathcal{F}|}{2}$ edges. In the next arguments, we refer to this set of edges as EC.

Let $J \subseteq I$, $|J| \leq k$, be a true solution for the Set Cover problem. We claim that $E = \{v_i v_i' \mid i \in J\} \cup EC$ is a true solution for CI in the corresponding instance I_{CI}. First, observe that $|E| \leq k + 2\binom{|\mathcal{F}|}{2}$, and recall that all the subsets V_{ij} and $V_{i'j'}$ of C are connected thanks to the edges of EC. Now let us check that the remaining subsets V_i, $1 \leq i \leq m$, are also connected. By contradiction, suppose that it is not the case and let $G = (V, E)$. Thus, there exists a subset $V_i \subseteq V$ such that $G[V_i]$ is not connected. Recall that all the edges EC are forced to be in the feasible solution E and thus the only possibility is that the vertices in $A \cap V_i$ are not connected in the solution to the vertices of $B \cap V_i$. However, x_i must have been covered by a subset X_j, for some $j \in J$, in the true solution to the Set Cover problem. Consequently, by construction of I_{CI}, the subset V_i contains both v_j and v_j', and so the subgraph $G[V_i]$ should be connected since edge $v_j v_j'$ belongs to E, a contradiction.

On the other hand, let E^* be a true solution for (V, C, k'), let $EM = E^* \setminus EC$ be the edges of the solution E^* that have one endpoint in A and the other in B. Since $|EC| = 2\binom{|\mathcal{F}|}{2}$, we know that $k \geq |EM|$. We claim that with $|EM|$ subsets of $\mathcal{F}$ we can satisfy I_{SC}. Observe that, if there is an edge $v_i v_j' \in EM$ with $i \neq j$, this edge can be replaced by the edge $v_i v_i'$ and the solution is still feasible. In fact, all the complexes that were connected by $v_i v_j'$ belong to the intersection of $C(v_i) \cap C(v_j')$ and we have $C(v_i) = C(v_i')$. So we obtain a feasible solution with k' edges whose end-points are both included in either A or B, or one end-point is v_i and the other is $v_{i'}$ for some $i \in J$ with $|J| = |EM| \leq k$. By construction of I_{CI}, the set J is a feasible solution to I_{SC}. □ □

Proof for proposition 1:

Proof. Let $M = \sum_{v \in V} |C(v)|$ and remark that $m_t(e^t) \geq m_{t+1}(e^{t+1})$. We divide the steps of the algorithm into $\log_2 M$ phases. During each phase x, the value of $m_t(e^t)$ remains in the interval $[a_x, 2a_x]$, where $a_0 = \max_{u,v \in V} |C(u) \cap C(v)|$ and $a_x = a_{x-1}/2 = a_0/2^x$. Since $a_{\log_2 M} = a_0/2^{\log_2 M} = a_0/M < 1$ and that an edge is selected only if it connects at least two components, we need only $\log_2 M$ phases. Let Λ_x be the number of selected edges during phase x. We observe that the output of Algorithm 1 is $SOL = \sum_{x=1}^{\log_2 M} \Lambda_x$. Let also δ_x be the number of components of the graphs $G^t[V_i]$ that have been connected during that phase. We have

$$a_x \Lambda_x \leq \delta_x \leq 2a_x \Lambda_x \tag{8a}$$

Since during phase x we have reduced the number of components by δ_x, we know that the remaining number of components to connect at the beginning of the phase was at least δ_x. Furthermore, the maximum value $m_t(e)$ of an edge during phase x was upper bounded by $2a_x$ and we have $m_t(e^t) \geq m_{t+1}(e^{t+1})$. So to connect these remaining components, we need at least $\delta_x/2a_x$ edges. Hence we have

$$OPT \geq \frac{\delta_x}{2a_x} \tag{8b}$$

Using Eq. 8a, we obtain that $2 \cdot OPT \geq \Lambda_x$. Now, summing over all phases, we obtain

$$\sum_{x=1}^{\log_2 M} 2 \cdot OPT = 2 \cdot OPT \cdot \log_2 M \quad \geq \quad \sum_{x=1}^{\log_2 M} \Lambda_x = SOL \tag{8c}$$

Finally, since $\kappa = \max_{v \in V} |C(v)|$, we have $M \leq |V| \cdot \kappa$ and the result follows. □ □

Proof for proposition 2:

Proof. By Lemma 1, we may assume that $m_t(e) \leq 1$, for every $e \in E$ and for every t. Consequently, when an edge e^t is chosen, it will be useful to connect only one complex. Thus, all the edges that are chosen by the algorithm are necessary (in the sense that one edge would be necessary to connect such complex) and the solution is optimal. □ □

8 Lists of Contacts for the Assemblies Studied

In section 8.1, we provide a classification of various contacts reported in the literature, classified as a function of the experimental technique they were observed with. These contact categories are used to define reference edge sets used for the assessment of the edges reported by our algorithms.

We proceed in section 8.2 with the corresponding lists for the assemblies studied.

8.1 Pairwise Contacts within Macro-molecular Complexes

Crystal contacts: [$\mathrm{C_{Xtal}}$] A high-resolution crystal structure of an assembly can be seen as the gold standard providing all pairwise contacts between its constituting molecules. Given such a crystal structure, all pairs of molecules are tested to check whether they define a contact. A pair defines a contact provided that in the solvent accessible (SAS) model of the assembly [1], two atoms from these partners define an edge in the α-complex of the assembly for $\alpha = 0$, as classically done to define macro-molecular interfaces [8, 11, 16].)

This protocol actually calls for one comment. For protein interfaces, it is generally accepted that any biologically specific contact has a surface area beyond 500Å^2, or equivalently, involves at least 50 atoms on each partners [11]. For assemblies, because of the promiscuity of molecules, this threshold does not apply directly. As an example, consider the number of atoms observed at interfaces for the yeast exosome complex (Fig. 1). While selected interfaces meet the usual criterion, others involve a handful of atoms. For this reason, in addition to $\mathrm{C_{Xtal}}$, we defined a set $\mathrm{C_{Xtal}^-}$ involving the most prominent contacts only (14 contacts out of 26). We note in passing that the existence of a hierarchy of interface size within a protein assembly has been reported in [15, 11].

Cross-linking (set $\mathrm{C_{XL}}$). Cross-linking is an analytical technique which consists in chemically linking surface residue of two proteins located nearby. This technique is used to identify protein-protein interactions, upon disrupting the cell and identifying the cross-linked proteins. The outcome allows identifying interacting proteins within an assembly, but also transient interactions which get stabilized by the cross-linker. The distance between the two amino-acids cross-linked is circa 25Å, including the length of the linker and the span of the side-chains of the two amino-acids involved.

Due to this distance, the two proteins cross-linked may not form an interface in the sense defined above. However, cross-linking contacts are considered as interfacial contacts in [12], defining a *low-resolution topology.*

Dimers obtained from various biophysical experiments (set $\mathrm{C_{Dim}}$). The following experiments deliver information on the existence of a dimer involving two proteins:

[1] Given a van der Waals models, the corresponding SAS model consists of expanding the atomic radii by 1.4Å, so as to account for an implicit layer of water molecules on the model. The SAS model also allows capturing intersections between atoms which are nearby in 3D space, but are not covalently bonded.

- Mass spectrometry (MS^1) or Tandem Mass spectrometry (MS^2): upon collecting a dimer, and since no re-arrangement occurs in gas phase, the two proteins form a dimer in the assembly analyzed.

- Tandem affinity purification (TAP): a bait put on one protein pulls down another protein, upon capturing the marked protein on a affinity purification column.

- Co-immuno-precipitation of two proteins: as above.

- Native Agarose Gel electrophoresis: two proteins are inferred to be interacting if instead of two sharp bands (assuming mol. wt. to be different) a broad band spread over a range of molecular weight is observed.

- NMR titrations: information of the interacting residues of one protein is inferred from the perturbation of the chemical shifts of the interfacial residues obtained when adding the partner.

Contacts observed in Yeast two-hybrid assays (set $\mathbf{C_{Y2H}}$). Because such contacts are prone to false positive, we reported such contacts within our tables for the sake of completeness, but did not use them to assess our algorithms.

Reference sets used for our assessments. Owing to the reliability of the three sets of contacts $\mathrm{C_{Xtal}}, \mathrm{C_{Dim}}$, and $\mathrm{C_{XL}}$, for a given assembly, we define the interface contacts

$$\mathrm{C_{Exp}} = \mathrm{C_{Xtal}} \cup \mathrm{C_{Dim}} \cup \mathrm{C_{XL}}. \tag{9}$$

8.2 Contacts for the Assemblies of Interest

NB: No cross-linking or cryo-EM or x-ray data is available for eIF3.

Table 2 List of contacts determined from experiments for Yeast Exosome

C_{Xtal} (26 contacts)			C_{Dim} (7 contacts)
X-Ray Crystallography, 2.8 Å[18]			TAP, MS^1, MS^2,
Chains	Subunits	#Interface atoms	*Partial Denaturation* [10] [25]
CG	(Rrp43, Rrp40)	2	(Rrp43, Csl4)
EI	(Rrp42, Csl4)	6	(Rrp45, Rrp40)
AF	(Rrp45, Mtr3)	19	(Rrp46, Rrp40)
FH	(Mtr3, Rrp4)	24	(Rrp45, Rrp46)
DF	(Rrp46, Mtr3)	54	(Rrp45, Rrp41)
AH	(Rrp45, Rrp4)	59	(Rrp43, Rrp46)
HI	(Rrp4, Csl4)	60	(Rrp42, Mtr3)
AC	(Rrp45, Rrp43)	72	
DI	(Rrp46, Csl4)	79	
AJ	(Rrp45, Dis3)	95	
GI	(Rrp40, Csl4)	117	
CJ	(Rrp43, Dis3)	148	
CI	(Rrp43, Csl4)†	211	
BE	(Rrp41, Rrp42)	223	
EJ	(Rrp42, Dis3)	231	
AG	(Rrp45, Rrp40)†	245	
EF	(Rrp42, Mtr3)†	313	
FI	(Mtr3, Csl4)	327	
AD	(Rrp45, Rrp46)†	349	
BH	(Rrp41, Rrp4)	352	
CD	(Rrp43, Rrp46)†	369	
BJ	(Rrp41, Dis3)	371	
DG	(Rrp46, Rrp40)†	411	
CF	(Rrp43, Mtr3)	446	
EH	(Rrp42, Rrp4)	458	
AB	(Rrp45, Rrp41)†	463	

† signifies those contacts which are also recovered by other biophysical experiments, TAP, MS^1, MS^2

Table 3 List of contacts determined from experiments for Yeast 19S Proteasome Lid

C_{Dim} (3 contacts)	C_{XL} (14 contacts)		C_{Y2H} (9 contacts)
$Nat. Agarose\,Gel$ [21]	$CX - DSSO, DSS, BS3$	$References$	*Ref* - [21]
(Rpn5, Rpn8)	(Rpn3, Rpn7)	[12][14]	(Rpn3, Rpn7)
(Rpn6, Rpn8)	(Rpn3, Rpn8)	[12]	(Rpn3, Rpn12)
(Rpn8, Rpn9)†	(Rpn3, Rpn12)	[12]	(Rpn5, Rpn6)
	(Rpn3, Sem1)	[12][21]	(Rpn5, Rpn8)
	(Rpn5, Rpn6)	[12]	(Rpn5, Rpn9)
	(Rpn5, Rpn9)	[12][14]	(Rpn5, Rpn11)
	(Rpn6, Rpn7)	[12]	(Rpn8, Rpn9)
	(Rpn6, Rpn11)	[12]	(Rpn8, Rpn11)
	(Rpn7, Rpn11)	[12]	(Rpn9, Rpn11)
	(Rpn7, Sem1)	[12][21]	
	(Rpn8, Rpn9)	[12]	
	(Rpn8, Rpn11)	[12]	
	(Rpn3, Rpn5)	[21]	
	(Rpn3, Rpn11)	[14]	

Table 4 List of contacts determined from experiments for eIF3

C_{Dim} (19 contacts)	C_{AUX} (1 contact)
$TAP, MS^1, MS^2, Partial\,Denaturation$ [26]	
(a, b)	(b, e)
(b, i)	
(b, g)	
(d, e)	
(e, l)	
(f, h)	
(f, m)	
(g, i)	
(h, m)	
(k, l)	
$Immuno - precipitation$ [26]	
(b, f)	
$NMR\,Titrations$ [9]	
(b, j)	
$Native\,Agarose\,Gel$, [23] [7]	
(a, c)	
(b, c)	
(b, h)	
(j, c)	
(j, f)	
(j, h)	
(j, k)	

RESEARCH CENTRE
SOPHIA ANTIPOLIS – MÉDITERRANÉE

2004 route des Lucioles - BP 93
06902 Sophia Antipolis Cedex

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr

ISSN 0249-6399