Unveiling Contacts within Macromolecular Assemblies by Solving Minimum Weight Connectivity Inference (MWC) Problems* 

Deepesh Agarwal, Christelle Caillouet, David Coudert, and Frederic Cazals

Consider a set of oligomers listing the subunits involved in subcomplexes of a macromolecular assembly, obtained e.g. using native mass spectrometry or affinity purification. Given these oligomers, connectivity inference (CI) consists of finding the most plausible contacts between these subunits, and minimum connectivity inference (MCI) is the variant consisting of finding a set of contacts of smallest cardinality. MCI problems avoid speculating on the total number of contacts but yield a subset of all contacts and do not allow exploiting a priori information on the likelihood of individual contacts. In this context, we present two novel algorithms, MILP-W and MILP-Wb. The former solves the minimum weight connectivity inference (MWC), an optimization problem whose criterion mixes the number of contacts and their likelihood. The latter uses the former in a bootstrap fashion to improve the sensitivity and the specificity of solution sets.

Experiments on three systems (yeast exosome, yeast proteasome lid, human eIF3), for which reference contacts are known (crystal structure, cryo electron microscopy, cross-linking), show that our algorithms predict contacts with high specificity and sensitivity, yielding a very significant improvement over previous work, typically a twofold increase in sensitivity.

The software accompanying this paper is made available and should prove of ubiquitous interest whenever connectivity inference from oligomers is faced. Molecular & Cellular Proteomics 14: 10.1074/mcp.M114.047779, 2274–2284, 2015.

1 CONNECTIVITY INFERENCE FROM SETS OF OLIGOMERS

Structural Inference from Oligomers and Contacts—Unraveling the function of macromolecules and macromolecular machines requires atomic level data, both in their static and dynamic dimensions, the latter coding for thermodynamic and kinetic properties (1). However, obtaining even static snapshots of large systems remains a tour de force, so that alternative methods are being developed, based in particular on reconstruction by data integration (RDI)1, a strategy aiming at producing models of assemblies using complementary experimental data (2). In its full generality, RDI accommodates both structural and purely combinatorial data (3). The former typically consists of crystallographic (high resolution) structures, electron microscopy maps, and NMR models. The latter comprise information on the composition and copy numbers of subunits, as well as pairwise contacts. Given a large assembly, information on oligomers (i.e. subcomplexes of the assembly) can be obtained by methods such as tandem affinity purification (4) or native mass spectrometry (5, 6), and such oligomers can be complemented by information on pairwise protein-protein interactions (7, 8). More specifically, oligomers of varying size can be obtained under various experimental conditions. While stringent conditions (e.g. low pH) result in complete dissociation of the assembly so that the individual molecules are identified, less stringent conditions result in the disruption of the assembly into multiple overlapping oligomers. Assembled together, such oligomers can be used to infer contacts within the assembly (5). In the context of RDI, the models stemming from such analysis do not, in general, achieve atomic resolution. They can, however, be used to bridge the gap to atomic level models of subsystems of the assembly under scrutiny (9, 10).

Unweighted and Weighted Connectivity Inference: MCI and MWCI—Consider a macromolecular assembly consisting of subunits (typically proteins or nucleic acids). Assume that these subunits are known but that the pairwise contacts between them are unknown or only partially known—in this latter case, the presence or the likelihood of selected contacts is known. Connectivity inference (CI) is the problem concerned with the elucidation of contacts between these subunits, as it ideally aims at producing one contact for each pair of subunits.

The abbreviations used are: RDI, reconstruction by data integration; CI, connectivity inference; MCI, minimum connectivity inference; MWCI, minimum weight connectivity inference; Cvg score, coverage score; MS, mass spectrometry; MS/MS, tandem mass spectrometry; SAS, solvent accessible surface; TAP, tandem affinity purification; MILP, mixed integer linear programming.
sharing an interface in the assembly (Fig. 1). Note that mathematically, the subunits may be seen as the nodes of a graph whose edges are defined by the contacts. Thus, in the sequel, we use contacts and edges interchangeably.

To address CI, let an oligomer formula be a list of subunits defining a connected component within the assembly. That is, an oligomer formula is the description of the composition of the oligomer, giving the number of copies of each molecule. We define a connectivity inference specification (specification for short) as a list of oligomers. The solution of a CI problem consists of a set of contacts, denoted \( S \) in the sequel. This set is called a valid edge set or a solution provided that for each oligomer and also for the whole complex: Restricting the edges from \( S \) to the vertices of an oligomer formula yields a connected graph (Fig. 1). In defining valid solutions, two critical questions arise: How many contacts should one seek, and should all edges be treated on an equal footing?

On the Number of Contacts—In the absence of a priori knowledge on the likelihood of individual contacts, a solution is naturally assessed by its number of contacts. Mastering this size is nontrivial since the number of interfaces between subunits in the assembly is unknown. On the one hand, the trivial solution involving all possible edges is uninteresting since it is likely to contain a large number of false positives. On the other hand, one may solve the minimum connectivity inference problem (MCI), namely the variant of CI minimizing the number of contacts used. To do so, observe that minimally connecting an oligomer merely requires a tree, that is a graph whose number of edges is the number of vertices (subunits) minus one. Thus, solving MCI consists of choosing for each oligomer the tree yielding the solution of minimum size. Yet, in doing so, one is likely to generate false negatives. Given these two extremes, one goal of this work is to optimize the number of contacts reported so as to maximize the number of true positives and true negatives—a goal that will be achieved using so-called consensus edges and a bootstrapping strategy.

A Priori Knowledge on Contacts—In a number of cases, the likelihood of a given contact may be known. On the experimental side, various assays have been developed to check whether two proteins interact, including yeast-two-hybrid, mammalian protein–protein interaction trap, luminescence-based mammalian interactome, yellow fluorescent protein complementation assay, co-immuno precipitation, etc (7, 8). But information obtained must be used with care for several reasons, notably because expression systems force promiscuity between proteins that may otherwise be located in different cellular compartments and also because affinity purification typically involves concentration beyond physiological levels. On the in-silico side, various interactions attributes can be used, such as gene expressions patterns (proteins with identical patterns are more likely to interact), domain interaction data (a known interaction between two domains hints at an interaction between proteins containing these domains), common neighbors in protein–protein interaction networks, or bibliographical data (number of publications providing evidence for a particular interaction). Here again, these pieces of information have a number of caveats. In particular, structural data from crystallography or mass spectrometry yield a bias toward stable interactions at the detriment of transient ones. For these reasons, strategies computing confidence scores usually resort to machine learning tools trained on the aforementioned data (11) and also (12).

In any case, being able to accommodate such previous knowledge is precisely another goal of the algorithms developed in this work.
Unveiling Contacts within Macromolecular Assemblies by Solving MWCI Problems

**Algorithms**—From a computer science perspective, solving CI problems is a hard task (supplemental section 11). Two algorithms targeting such problems have been developed so far. The first one is a two-stage heuristic method (13). First, random graphs meeting the connectivity constraint are generated by incrementally adding random edges. Second, a genetic algorithm is used to reduce the number of edges and also boost their diversity. Once the average size of the graphs stabilizes, the pool of graphs is analyzed to spot highly conserved edges.

The second one is our method solving MCI problems, based on a mixed integer linear program (14). On the one hand, this work delineates the combinatorial hardness of the CI problem and offers two algorithms. Of the particular interest is MILP since it delivers all optimal solutions of a given MCI problem. (As discussed in the supplemental section 11, this algorithm may require exponential time for hard instances, even though this behavior was not observed for the cases processed.) On the other hand, when assessed against contacts seen in crystal structures, the solutions of MILP suffer from two limitations. First, in all solutions, few false negatives are observed at the expenses of selected false positives. On the other hand, since all edges have a unit weight, one cannot favor or penalize some of them.

In this context, this paper makes two improvements. First, we introduce the minimum weight connectivity inference problem (MWCI), which allows computing optimal solutions incorporating a priori knowledge on the likelihood of edges. Second, we present algorithm MILP-W to solve MWCI problems, an algorithm aiming at maximizing the sensitivity and specificity of the set of contacts reported.

### 2 MINIMUM WEIGHT CONNECTIVITY INFERENCE: MATHEMATICAL MODEL

**Oligomers and Pools of Edges**—In solving CI problems, a valid edge set consists of edges such that each of them involves two subunits belonging to at least one oligomer. More precisely, consider an oligomer $O_i$. This oligomer defines a pool of candidate edges equal to all pairs of subunits found in $O_i$. Likewise, the pool of candidate edges $Pool_{\mathcal{E}}(\mathcal{O})$ defined by a set of oligomers $\mathcal{O}$ is obtained by taking the union of the pools defined by the individual oligomers. Note that one can also consider a restricted set of oligomers involving the oligomers whose size is bounded by an integer $s$, denoted $\mathcal{O}_{\leq s}$, the corresponding pool of candidate edges being denoted $Pool_{\mathcal{E}}(\mathcal{O}_{\leq s})$. The rationale for using small oligomers is that they favor local contacts. (Note that the extreme case is that of a dimer since the contact seen in a dimer must belong to every solution.) Note also that one can edit a pool of edges to enforce or forbid a given edge in all solutions. For example, if a cryo-electron microscopy map of the assembly is known and two proteins have been located far apart in the map, one can forbid the corresponding contact even though the two proteins appear in a common oligomer.

We now present two ways to solve CI problems. **Unweighted Case**—In the unweighted case, each edge from the pool is assigned a unit weight so that the weight of a solution is the number of its edges. The corresponding optimization problem is called MCI, and an algorithm solving it, MILP, has been proposed in (14).

**Weighted Case**—In the weighted case, each candidate edge $e$ from the pool $Pool_{\mathcal{E}}(\mathcal{O})$ is assigned a weight $w(e)$, namely a real number in range $[0,1]$. This number encodes the likelihood for the edge to be a true contact. Taking $G = 1/2$ as a baseline (i.e. no a priori on this contact), a value $F > G$ is meant to favor the inclusion of this edge in solutions, while a value $U < G$ is meant to penalize this edge.

**Unifying the Unweighted and Weighted Cases: MWCI Problems**—Depending on how much information is available on candidate contacts, one may wish to stress the number of contacts in a solution, or their total weight. Both options can actually be handled at once by interpolating between the previous two problems. Using a real number $\delta \in [0,1]$, we define a functional mixing the number of edges and their weights, this latter one being favored for values beyond the threshold 1/2. That is, we define the cost of a solution $S$ using two terms respectively corresponding to the number of edges and their weights:

$$C(S) = \alpha \sum_{e \in S} 1 + (1 - \alpha) \sum_{e \in S} \left( \frac{1}{2} - w(e) \right) = \sum_{e \in S} C_e(e)$$

(Eq. 1)

with

$$C_e(e) = \frac{\alpha + 1}{2} - (1 - \alpha)w(e)$$

(Eq. 2)

Eq. (1) corresponds to the objective of the optimization problem denoted MWCI in the sequel. The following comments are in order:

- In using $\alpha = 1$, which is the strategy used by algorithm MILP (14), the weights play no role, and the interchangeability of edges favors the exploration of a large pool of solutions.
- The situation is reversed for small values of $\alpha$. In particular, the conjunction $\alpha = 1$ and different weights for all edges typically yields a small number of solutions since ties between solutions are broken by the weights.
- A null weight does not prevent a given edge to appear in solutions. To forbid an edge, one should edit the pool of candidate edges, as explained above i.e. remove this edge from the pool.

**Remark 1**—Assume that each edge has a default weight $d$ instead of 1/2. Eq. (1) is a particular case of the following:

$$C_{\alpha,d}(e) = \alpha(1 - d) + d - (1 - \alpha)w(e)$$

(Eq. 3)
Setting $d = 1/2$ in Eq. (3) yields the edge cost of Eq. (1). On the other hand, setting $d = 1$ yields a constant term 1 instead of $(\alpha + 1)/2$. Since the default $d = 1$ yields a weighting criterion less sensitive to weights, we use $d = 1/2$.

We also observe that $dC_{\alpha}(e)/d\alpha = 1 - d + w(e)$. Thus, when varying, the edge weight prevails or not depending on its value with respect to the value $1 - d$. For $d = 1/2$, one gets $dC_{\alpha}(e)/d\alpha = 1 - d + w(e)$.

3 MINIMUM WEIGHT CONNECTIVITY INFERENCE: ALGORITHMS

3.1 Algorithm MILP-W—Algorithm MILP-W generalizes the unweighted version of MILP (14) and allows enumerating all optimal solutions with respect to the criterion of Eq. (1). The algorithm solves a mixed integer linear program, using constraints imposing the connectivity constraints inherent to all oligomers. Candidate edges are represented by binary variables taking the value 1 when edges belong to a specific solution (14) and 0 otherwise.

More precisely, algorithm MILP-W iteratively generates all optimal solutions and adds at each iteration extra constraints preventing finding the same solution twice. To this end, the method starts with a first resolution of the problem to get an optimal solution, if any. This solution defines a set of edges $S_{\text{OPT}}$ that are forced to be strictly less than the number of edges in solutions seen so far. The resolution is launched again, and the criterion value is compared with $\text{OPT}$. This process is iterated until the value of the solution exceeds $\text{OPT}$.

Remark 2—By picking the adequate combination of $\alpha$ and $w$, the individual edge cost of Eq. (2) can be null. Edges with null cost can create troubles in the enumeration problem since solutions with the same cost but nested sets of edges can be created. To get rid of spurious large edges, it is sufficient to build the Hasse diagram (for the inclusion) of all solutions, and remove the terminal nodes of this diagram.

3.2 Solutions and Consensus Solutions—The set of all optimal solutions reported by MILP-W is denoted $S_{\text{MILP-W}}$ and the set of contacts used in these solutions is denoted $e_{\text{MILP-W}}$. The size of a solution $S \in S_{\text{MILP-W}}$ denoted $|S|$, is its number of contacts. The score of a contact appearing in a solution $S \in S_{\text{MILP-W}}$ called contact score for short, is the number of solutions from $S_{\text{MILP-W}}$ containing it. The score of a solution $S \in S_{\text{MILP-W}}$ is the sum of the scores of its contacts. Finally, a consensus solution is a solution achieving the maximum score over $S_{\text{MILP-W}}$. The set of all such solutions being denoted $S_{\text{MILP-W}}^{\text{cons}}$. The contacts found in consensus solutions are called the consensus contacts, and define the set $e_{\text{MILP-W}}^{\text{cons}}$.

As noticed earlier, when $\alpha = 1$, algorithm MILP-W matches algorithm MILP. Therefore, for the sake of clarity, the solution set, consensus solutions, and the associated edge sets are respectively denoted $S_{\text{MILP}}$, $S_{\text{MILP-W}}^{\text{cons}}$, $e_{\text{MILP}}$, and $e_{\text{MILP-W}}^{\text{cons}}$. These notations are summarized in Table 1.

To further assess the quality of the solution set $S(= S_{\text{MILP}}, S_{\text{MILP-W}}^{\text{cons}})$, assume that a reference set of contacts $E_{\text{Ref}}$ is known. The ideal situation is that where a high-resolution crystal structure is known since then all pairwise contacts can be inferred (15). This reference set together with the pool Pool$_{\ell}$ defines positive (P), negative (N), and missed contacts (M) (Fig. 2). From these groups, one further classifies the edges of a predicted solution in set $s$ into four categories, namely true positive (TP), false positive (FP), true negative (TN), and false negative (FN).

Positives (P) and negatives (N) decompose as $P = TP + FN$, and $N = TN + FP$. From which one defines the sensitivity ROC$_{\text{spec}}$ and the specificity ROC$_{\text{spec}}$ as follows:

$$\text{ROC}_{\text{spec}} = \frac{|P|}{|P| + |N|} \quad \text{ROC}_{\text{spec}} = \frac{|P|}{|P| + |N|} \quad \text{Eq. (4)}$$

Note that specificity requires the set $N$ to be nonempty, which may not be the case if Pool$_{\ell} \subset E_{\text{Ref}}$.

We also combine the previous values to define the following coverage score, which favors true positives, penalizes false positives and false negatives, and scales the results with respect to the total number of reference contacts (since $P$ might be included into $E_{\text{Ref}}$ if the pool size is too small):

$$\text{Cvg}(S) = \frac{|TP| - (|FP| + |FN|)}{|E_{\text{Ref}}|} \quad \text{Eq. (5)}$$

Note that the maximum value is one and that the coverage score may be negative.

3.3 Algorithm MILP-W$_B$—The focus on consensus edges is quite natural since these may prosaically be seen as the backbone of the connectivity in the assembly. However, alternative edges of significant importance may exist too. To unveil such edges, we preclude one or more consensus edges, so as to trigger a rewiring of the connectivity of solutions, and check which novel consensus edges appear along...
Unveiling Contacts within Macromolecular Assemblies by Solving MWCI Problems

The exosome involves 10 protein units. The yeast exosome, an assembly consisting of 10 subunits. The connectivity inference problem consists of inferring contacts between the subunits from the composition of oligomers, i.e., connected blocks of the assembly. (left) Crystal structure. (right) The solid edges reported by the algorithm MILP, while the dashed edges are not present in the solution.

We test the performance of the algorithms MILP-W and MILP-WB on the following three systems for which reference contacts for validation are available either coming from crystal structure or from various biophysical experiments such as cryo-EM based reconstruction, cross-linking, and MS/MS dimers. See supplemental Section 8 for the input to the algorithms and the supplemental Section 9 for the reference contacts.

4 MATERIAL: TEST SYSTEMS

We test the performance of the algorithms MILP-W and MILP-WB on the following three systems for which reference contacts for validation are available either coming from crystal structure or from various biophysical experiments such as cryo-EM based reconstruction, cross-linking, and MS/MS dimers. See supplemental Section 8 for the input to the algorithms and the supplemental Section 9 for the reference contacts.

4.1 Yeast Exosome—The exosome involves 10 protein types (Figs. 3 and 4), and 19 oligomers have been reported (13), ranging in size from two to nine (Table II and supplemental Section 8). Note that originally 21 oligomers are reported in (13), including a trivial case of having all 10 protein types and one other oligomer of size 8 that has a duplicate, leaving behind 19 distinct oligomers.

Oligomers up to size five are required to encompass nine out of 10 proteins—the protein Cs4 is present in size nine oligomers only. In terms of contacts, classical interfaces modeling tools (15) applied to the crystal structure yield 26 contacts amid the 10 proteins, and 20 contacts in the assembly depleted of Cs4 (Fig. 4).

The status of Cs4 is interesting since, as discussed in Section 2, local contacts are favored by small oligomers. In the sequel, we therefore consider two settings, namely the full exosome, and the exosome without Cs4. In the former case, all oligomers define a pool $E_{\text{total}}$ of 45 candidate edges; in the latter, the pool $P$ contains 36 candidate edges.

4.2 Yeast 19S Proteasome Lid—Proteasomes are protein 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 nine distinct protein types each with unit stoichiometry (Fig. 9).

Series of overlapping oligomers were formed by mass spectrometry (MS), tandem MS, and cross-linking using BS3. In total, 14 complexes were obtained out of which eight came from MS, MS/MS and six came from cross-linking experiments (16) (Table III and supplemental Section 8).

4.3 Human 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, human eIF3 consists of 13 different protein types each with unit stoichiometry (Fig. 12). The eIF3 complex in this text refers to the human eIF3 unless otherwise stated.

The exosome involves 10 protein units. The yeast exosome, an assembly consisting of 10 subunits. The connectivity inference problem consists of inferring contacts between the subunits from the composition of oligomers, i.e., connected blocks of the assembly. (left) Crystal structure. (right) The solid edges reported by the algorithm MILP, while the dashed edges are not present in the solution.

We test the performance of the algorithms MILP-W and MILP-WB on the following three systems for which reference contacts for validation are available either coming from crystal structure or from various biophysical experiments such as cryo-EM based reconstruction, cross-linking, and MS/MS dimers. See supplemental Section 8 for the input to the algorithms and the supplemental Section 9 for the reference contacts.

4 MATERIAL: TEST SYSTEMS

We test the performance of the algorithms MILP-W and MILP-WB on the following three systems for which reference contacts for validation are available either coming from crystal structure or from various biophysical experiments such as cryo-EM based reconstruction, cross-linking, and MS/MS dimers. See supplemental Section 8 for the input to the algorithms and the supplemental Section 9 for the reference contacts.

4.1 Yeast Exosome—The exosome involves 10 protein types (Figs. 3 and 4), and 19 oligomers have been reported (13), ranging in size from two to nine (Table II and supplemental Section 8). Note that originally 21 oligomers are reported in (13), including a trivial case of having all 10 protein types and one other oligomer of size 8 that has a duplicate, leaving behind 19 distinct oligomers.

Oligomers up to size five are required to encompass nine out of 10 proteins—the protein Cs4 is present in size nine oligomers only. In terms of contacts, classical interfaces modeling tools (15) applied to the crystal structure yield 26 contacts amid the 10 proteins, and 20 contacts in the assembly depleted of Cs4 (Fig. 4).

The status of Cs4 is interesting since, as discussed in Section 2, local contacts are favored by small oligomers. In the sequel, we therefore consider two settings, namely the full exosome, and the exosome without Cs4. In the former case, all oligomers define a pool $E_{\text{total}}$ of 45 candidate edges; in the latter, the pool $P$ contains 36 candidate edges.

4.2 Yeast 19S Proteasome Lid—Proteasomes are protein 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 nine distinct protein types each with unit stoichiometry (Fig. 9).

Series of overlapping oligomers were formed by mass spectrometry (MS), tandem MS, and cross-linking using BS3. In total, 14 complexes were obtained out of which eight came from MS, MS/MS and six came from cross-linking experiments (16) (Table III and supplemental Section 8).

4.3 Human 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, human eIF3 consists of 13 different protein types each with unit stoichiometry (Fig. 12). The eIF3 complex in this text refers to the human eIF3 unless otherwise stated.
A total of 27 complexes were generated from the assembly by manipulating the ionic strength of the solution and using tandem mass spectrometry (17) (Table IV and supplemental Section 8). The subunit eIF3j is labile, and since none of 27 subcomplexes comprise this subunit, we exclude it from the list of protein types, leaving behind 12 protein types.

5 RESULTS

We first provide results for MILP-WB with $\alpha = 1$, namely when all edges have the same weight. In a second step, we illustrate the benefits of using weights.

5.1 Algorithm MILP-WB—As explained in Section 3.3, algorithm MILP-WB works by accumulating consensus contacts from MILP-W and those due to local rewiring as a result of precluding the initial consensus contacts one (or more) at a time. Consequently, our analysis focuses on the sensitivity, specificity and coverage statistics introduced in Section 3.2 in four settings:

- (C-1) The statistics for the edge set $e_{\text{MILP}}$, which serve as a baseline.
- (C-2) The statistics for the edge set $e_{\text{MILP-W}}$.
- (C-3) The statistics for the edge set $e_{\text{MILP-WB}}$ returned by MILP-WB after one iteration.
- (C-4) The statistics for the edge set $e_{\text{MILP-WB}}$ obtained when algorithm MILP-WB terminates.

The results are presented on Figs. 5, 6, 7, and 8 for the yeast exosome, Figs. 10 and 11 for the yeast proteasome, and Figs. 13 and 14 for human eIF3.

**FIG. 4.** Yeast exosome with and without Csl4: Contacts between subunits. Each edge corresponds to an interface between two subunits. The two numbers decorating an edge, respectively, refer to the number of atoms involved at that interface, and to the number of patches (connected components) of the interface. Interfaces were computed with the program INTERVOR, which implements the Voronoi model from (15). Note that a given subunit makes from three (e.g. Rrp40) to seven (e.g. Rrp45) interfaces.

**FIG. 5.** Yeast exosome: assessment of contacts yielded from different algorithms, MILP and MILP-WB. See supplemental Table 8 for the detailed statistics.

The following consistent observations can be made for the three systems:

- In comparing (C-1) against (C-2), the sensitivity decreases since consensus solutions have fewer edges. On the other hand, the specificity increases, indicating a
large number of true negatives or equivalently a small number of false positives—an observation in line with the high scores of edges in consensus solutions.

- In comparing (C-3) against (C-4), the sensitivity increases, while the specificity decreases. The variations observed for sensitivity and specificity actually depend on the number of contacts precluded in the bootstrap procedure. Indeed, precluding more contacts triggers more rewiring, which in turns yields a larger set of true positive edges (increased sensitivity), at the expense of more false positives (decreased specificity).
- Finally, for algorithm MILP-WB, one observes that a small number of iterations, typically in the range 1,…,3, is favorable to high coverage. This is because of the aforementioned counterbalance between sensitivity and specificity. In particular, when the bootstrap procedure halts, all contacts from the pool have been used, which entails a large number of true positives—high sensitivity—but also a large number of false positives—low specificity. Thus, the user may choose the risk level (in terms of false positives) he/she is willing to accept, depending on whether the focus is on sensitivity or specificity.
Comparison to Previous Work—The statistics just discussed compare favorably to previous work, obtained in particular with the heuristic network inference algorithm (13). We illustrate this fact with the results produced by MILP-WB after one iteration. On the yeast exosome with Csl4, the sensitivity of MILP-WB is \(1.67 (0.77/0.46)\) times that of network algorithm and Cvg. score increases from \(-0.08\) to \(0.35\) (lines T3 versus T0 in the supplemental Table 8).

For the yeast proteasome, one observes that the sensitivity for MILP-WB is \(1.76 (0.74/0.42)\) times that published earlier (13). Also, Cvg. score increases from \(-0.21\) to \(0\) (lines T3 versus T0 in the supplemental Table 11).

The comparison is not possible for eIF3, though, since the previously published contacts were computed manually using experimental information from various other sources (17). See, however, the supplemental Table 13 for the results using our algorithms.

5.2 Algorithm MILP-W—In this section, we illustrate the role of weights stemming from using Eq. (1) with \(a \neq 1\). One naturally expects a benefice in penalizing an edge that is a negative contact and that is thus predicted as a false positive or a true negative. But an improvement of statistics can also happen by merely favoring positive contacts. As an illustration, we consider the yeast exosome without Csl4 (specifications in Table II), using \(0.25\). We assign a weight of 0.6 to the following three contacts—(Rrp45, Rrp46), (Rrp40, Rrp46)
Unveiling Contacts within Macromolecular Assemblies by Solving MWCI Problems

FIG. 14. Human eIF3: variation of cumulative sensitivity, specificity, and coverage score with iteration index. Note that the iteration index also indicates number of contacts forbidden at a time. See supplemental Table 14 for the detailed statistics.

TABLE II
Yeast exosome: oligomers and associated statistics. The yeast exosome contains 10 proteins, with Csl4 found in size 9 oligomers only. (first column) Size of oligomers i.e. number of subunits (second column) Number of oligomers up to a given size (third column) size of the pool of contacts associated with the oligomers selected. Note also that for s = 8 and s = 9, the pool size is maximal i.e. contains all possible pairs of proteins: for s = 8: \( \binom{9}{2} = 36 \); for s = 9: \( \binom{10}{2} = 45 \). (4th column) The number of missed contacts, as defined on Fig. 2.

| Oligomer size s | \( O_{cs} \) | \( \text{Pool}_{E}(O_{cs}) \) | |M| |
|-----------------|--------------|---------------------------------|---|
| 2               | 3            | 3                              | 17 |
| 3               | 4            | 6                              | 14 |
| 4               | 6            | 13                             | 7  |
| 5               | 8            | 20                             | 3  |
| 6               | 9            | 21                             | 3  |
| 7               | 10           | 29                             | 3  |
| 8               | 15           | 36                             | 0  |
| 9               | 19           | 45                             | 0  |

and (Rrp41, Rrp42), the remaining contacts having the default weight of 0.5.

Upon moving from the instance without weights (i.e. \( \alpha=1 \)) to the instance with weights (i.e. \( \alpha=0.25 \)), we consider the changes in the sensitivity, specificity, and coverage, namely:

\[
\Delta = (\Delta \text{ROC}_{sens}, \Delta \text{ROC}_{spec}, \Delta \text{Cvg}) \tag{Eq. 6}
\]

Consider first \( e_{\text{MILP}} \). One observes four false positives instead of five (supplemental Fig. 15), improving the assessment tuple by \( (0, 0.06, 0.05) \). Thus, while none of the contacts has a weight less than 0.5, the relative value of weights has an incidence on the outcome.

Consider now \( e_{\text{MILP-W}} \). The union of consensus contacts has no false positive, and true positives are increased from 9 to 13, improving the assessment tuple by \( (0.2, 0.06, 0.45) \) (supplemental Fig. 16). For \( e_{\text{MILP-W}} \) the change in assessment tuple for \( n_{\text{MILP-W}} \) (first iteration) is \( (0.05, -0.12, 0) \). When more contacts are precluded, the trend seen is similar to earlier cases (supplemental Fig. 17).

The reader is referred to the supplemental Section 12 for a thorough assessment obtained upon varying the weights and the value of \( \alpha \).

6 DISCUSSION AND OUTLOOK

By giving access to a list of overlapping oligomers of a given macromolecular assembly, native mass spectrometry offers the possibility to infer pairwise contacts within that assembly, opening research avenues for systems beyond reach for other structural biology techniques. In this context, our work makes three contributions, based on state-of-the-art combinatorial optimization techniques.

First, we introduce the minimum weight connectivity inference problem (MWCI), which generalizes the minimum connectivity inference problem, by introducing weights associated with putative contacts. Second, we develop algorithm MILP-W to solve MWCI problems, taking into account a priori biological knowledge on the likelihood of contacts. Third, we
also develop algorithm MILP-\textsubscript{WB}, a bootstrap strategy aiming at enriching the solutions reported by MILP-W. Algorithm MILP-\textsubscript{WB} accumulates consensus contacts from MILP-W and those arising due to local rewiring as a result of precluding the initial consensus contacts one (or more) at a time. Our algorithms predict contacts with high specificity and sensitivity, yielding a very significant improvement over previous work, typically a twofold increase in sensitivity. Despite the combinatorial complexity of the problems addressed, all runs of algorithm MILP-\textsubscript{WB} terminated within a handful of seconds for all the cases processed in this work. Calculations with algorithm MILP-\textsubscript{WB} are more demanding, though, since the runtime depends on the combinatorics of the tuples to be precluded. These algorithms raise a number of opportunities and challenges.

In the context of native mass spectrometry, they offer the possibility to test various parameter sets, in particular regarding the number of contacts and their likelihood, and to compare the solutions obtained. More broadly, the ability to take into account confidence levels on putative edges should be key to incorporate scores currently being designed in proteomics in conjunction with various assays.

In terms of challenges, fully harnessing these algorithms raises difficult questions. On the practical side, one current difficulty is the lack of cases to learn from, namely assemblies for which a significant list of oligomers is known, and a high-resolution structure has been obtained. Such cases would be of high interest to tune the balance between the aforementioned two criteria (number of contacts and their likelihood). This would also aid in carrying out an in-depth study of incidence of weights on the solutions obtained from MILP-W runs, given true positives and false positives in the pool of contacts. Unfortunately, mass spectrometry studies are typically attempted on assemblies whose high-resolution structure are unknown and are likely to remain so, at least in the near future. On the theoretical side, outstanding questions remain open. The first one deals with the relationship between the set of oligomers processed and the solutions generated. Ideally, one would like to set up a correspondence between equivalence classes of oligomers yielding identical solutions. The ability to do so, coupled to the understanding of which oligomers are most likely generated, would be of invaluable interest. The second one relates to the generalization of our algorithms to accommodate cases where multiple copies of subunits are present. However, the multiple copies complicate matters significantly so that novel insights are called for not only computing solutions, but also representing them in a parsimonious fashion.

In any case, we anticipate that the implementations of our algorithms will prove its interest for the growing community of biologists using native mass spectrometry.

\footnote{This article contains supplemental material Tables 8 to 14 and Figs. 15 to 17.}

\[\] To whom correspondence should be addressed: Inria Sophia-Antipolis (Algorithms-Biology-Structure), 06902 Sophia Antipolis, France; Tel: +33 492 38 71 88; Fax: +33 497 15 53 95; E-mail: Frederic.Cazals@inria.fr.

REFERENCES

1. Schmidt, A. G., Xu, H., Khan, A. R., O’Donnell, T., Khurana, S., King, L. R., Manischewitz, J., Golding, H., Suphaphiphat, P., Carli, A., Settembre, E. C., Dormitzer, P. R., Keplinger, T. B., Zhang, R., Moody, M. A., Haynes, B. F., Liao, H.-X., Shaw, D. E., and Harrison, S. C. (2013) Preconfiguration of the antigen-binding site during affinity maturation of a broadly neutralizing influenza virus antibody. Proc. Natl. Acad. Sci. U.S.A. 110, 264–269

2. Alber, F., Dokudovskaya, S., Veenhoff, L. M., Zhang, W., Kipper, J., Devos, D., Suprapto, A., Karni-Schmidt, O., Williams, R., Chait, B. T., Rout, M. P., and Sali, A. (2007) Determining the architectures of macromolecular assemblies. Nature 450, 683–694

3. Alber, F., Forster, F., Korkin, D., Topf, M., and Sali, A. (2008) Integrating diverse data for structure determination of macromolecular assemblies. Ann. Rev. Biochem. 77, 11–11.35

4. Puig, O., Caspary, F., Rigaut, G., Rutz, B., Bouveret, E., Bragado-Nilsson, E., Wilm, M., Seraphin B. (2001) The tandem affinity purification method: A general procedure of protein complex purification. Methods 24, 218–229

5. Sharon, M., and Robinson, C. V. (2007) The role of mass spectrometry in structure elucidation of dynamic protein complexes. Annu. Rev. Biochem. 76, 167–193

6. Rose, R. J., Damoc, E., Denisov, E., Makarov, A., and Heck, A. J. R. (2012) High-sensitivity Orbitrap mass analysis of intact macromolecular assemblies. Nature Methods 9, 1084–1086

7. Berggård, T., Linse, S., and James, P. (2007) Methods for the detection and analysis of protein-protein interactions. Proteomics 7, 2833–2842

8. Braun, P., Tasan, M., Dreze, M., Barrios-Rodiles, M., Lemmens, I., Yu, H., Sahalie, J. J., Murray, R. R., Roncari, L., de Smet, A.-S., Venketesan, K., Rual, J.-F., Vandenhaute, J., Cusick, M. E., Pawson, T., Hill, D. E., Tavernier, J., Warna, J. L., Roth, P. F., and Vidal, M. (2008) An experimentally derived confidence score for binary protein-protein interactions. Nature Methods 6, 91–97

9. Dreyfus, T., Doye, V., and Cazals, F. (2012) Assessing the reconstruction of macro-molecular assemblies with tolerated models. Proteins 80, 2125–2136

10. Dreyfus, T., Doye, V., and Cazals, F. (2013) Probing a continuum of macro-molecular assembly models with graph templates of sub-complexes. Proteins 81, 2034–2044

11. Yu, J., Murrali, T., and Jr., R. R. (2012) Assigning confidence scores to protein-protein interactions. In Two Hybrid Technologies. Springer, pp. 161–174

12. Turner, B., Razick, S., Turinsky, A., Vlasblom, J., Crowdy, E. K., Cho, E., Morrison, K., Donaldson, I., and Wodak, S. (2012) iRefWeb: Interactive analysis of consolidated protein interaction data and their supporting evidence. Database 2010, baq023

13. Taverner, T., Hernandez, H., Sharon, M., Ruotolo, B., Matak-Vinkovic, D., Devos, D., Russell, R., and Robinson, C. (2008) Subunit architecture of intact protein complexes from mass spectrometry and homology modeling. Account Chem. Res. 41, 617–627

14. Agarwal, D., Araujo, J., Caillouet, C., Cazals, F., Coudert, D., and Perennes, M. P., and Sali, A. (2007) Determining the architectures of macromolecular assemblies. Nature 450, 683–694

15. Loriot, S., and Cazals, F. (2010) Modeling macro-molecular interfaces with Intervor. Bioinformatics 26, 964–965

16. Sharon, M., Taverner, T., Ambroggio, X. I., Deshaies, R. J., and Robinson, C. V. (2006) Structural organization of the 19s proteasome lid: Insights from ms of intact complexes. PLoS Biol. 4, e267

17. Zhou, H., Sandercock, A. M., Fraser, C., Ridlova, G., Stephens, E., Schenauer, M. R., Yokoii-Fong, T., Barsky, D., Leary, J. A., Hershey, J. W., Doudna, J. A., and Robinson, C. V. (2008) Mass spectrometry reveals modularity and a complete subunit interaction map of the eukaryotic translation factor elF3. Proc. Natl. Acad. Sci. U.S.A. 105, 18139–18144
18. Hernandez, H., Dziembowski, A., Taverner, T., Seraphin, B., and Robinson, C. (2006) Subunit architecture of multimeric complexes isolated directly from cells. *EMBO* 7, 605–610

19. Cazals, F., Proust, F., Bahadur, R. P., and Janin, J. (2006) Revisiting the Voronoi description of protein–protein interfaces. *Protein Sci.* 15, 2082–2092

20. Janin, J., Bahadur, R. P., and Chakrabarti, P. (2008) Protein-protein interaction and quaternary structure. *Qtr. Rev. Biophys.* 41, 133–180

21. Levy, E. D., Boeri Erba, E., Robinson, C. V., and Teichmann, S. A. (2008) Assembly reflects evolution of protein complexes. *Nature* 453, 1262–1265

22. Kao, A., Randall, A., Yang, Y., Patel, V. R., Kandur, W., Guan, S., Rychnovsky, S., D., Baldi, P., and Huang, L. (2012) Mapping the structural topology of the yeast 19s proteasomal regulatory particle using chemical cross-linking and probabilistic modeling. *Mol. Cell. Proteomics* 11, 1566–1577

23. Makino, D. L., Baumgartner, M., and Conti, E. (2013) Crystal structure of an RNA-bound 11-subunit eukaryotic exosome complex. *Nature* 495, 70–75

24. Lander, G. C., Estrin, E., Matysiak, M. E., Bashore, C., Nogales, E., and Martin, A. (2012) Complete subunit architecture of the proteasome regulatory particle. *Nature* 482, 186–191

25. Lasker, K., Förster, F., Bohn, S., Walzthoeni, T., Villa, E., Unverdorben, P., Beck, F., Aebersold, R., Sali, A., and Baumeister, W. (2012) Molecular architecture of the 26s protea-some holocomplex determined by an integrative approach. *Proc. Natl. Acad. Sci. U.S.A.* 109, 1380–1387

26. Querol-Audi, J., Sun, C., Vogan, J. M., Smith, M. D., Gu, Y., Cate, J. H., and Nogales, E. (2013) Architecture of human translation initiation factor 3. *Structure* 21, 920–928

27. Garey, M. R., and Johnson, D. S. (1979) Computers and Intractability: A Guide to the Theory of NP-Completeness. W. H. Freeman & Co., New York