HAL open science

Unveiling Contacts within Macro-molecular assemblies by solving Minimum Weight Connectivity Inference Problems

Deepesh Agarwal, Christelle Caillouet, David Coudert, Frédéric Cazals

▶ **To cite this version:**

Deepesh Agarwal, Christelle Caillouet, David Coudert, Frédéric Cazals. Unveiling Contacts within Macro-molecular assemblies by solving Minimum Weight Connectivity Inference Problems. [Research Report] RR-8622, Inria. 2014. hal-01078378v2

HAL Id: hal-01078378

https://hal.science/hal-01078378v2

Submitted on 25 Mar 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Unveiling Contacts within Macro-molecular assemblies by solving Minimum Weight Connectivity Inference Problems

D. Agarwal and C. Caillouet and D. Coudert and F. Cazals

RESEARCH REPORT
N° 8622
October 2014
Project-Team ABS and COATI

ISSN 0249-6399 ISRN INRIA/RR--8622--FR+ENG

Unveiling Contacts within Macro-molecular assemblies by solving Minimum Weight Connectivity Inference Problems

D. Agarwal* and C. Caillouet † and D. Coudert ‡ and F. Cazals §

Project-Team ABS and COATI

Research Report n° 8622 — version 2 — initial version October 2014 — revised version March 2015 — 52 pages

* Inria Sophia-Antipolis (Algorithms-Biology-Structure), 06902 Sophia Antipolis, France

† Inria Sophia Antipolis (COATI), and Univ. Nice Sophia Antipolis, CNRS, I3S, UMR 7271, 06900 Sophia Antipolis, France

‡ Inria Sophia Antipolis (COATI), and Univ. Nice Sophia Antipolis, CNRS, I3S, UMR 7271, 06900 Sophia Antipolis, France

§ Inria Sophia-Antipolis (Algorithms-Biology-Structure), 06902 Sophia Antipolis, France. Corresponding author: Frederic.Cazals@inria.fr

RESEARCH CENTRE
SOPHIA ANTIPOLIS – MÉDITERRANÉE

2004 route des Lucioles - BP 93
06902 Sophia Antipolis Cedex

Abstract: Consider a set of oligomers listing the subunits involved in sub-complexes 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-W`$_{\mathrm{B}}$. The former solves the *minimum weight connectivity inference* (MWCI), 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 the yeast exosome, for which both a high resolution crystal structure and a large set of oligomers is known, show that our algorithms predict contacts with high specificity and sensitivity, yielding a very significant improvement over previous work.
The software accompanying this paper is made available, and should prove of ubiquitous interest whenever connectivity inference from oligomers is faced.

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

Dévoilement de contacts au sein d'un assemblage macro-moléculaire, par résolution du problème d'inférence de connectivité de poids minimal

Résumé : Considérons un ensemble d'oligomères, obtenus e.g. par spectrométrie de masse native, listant les sous-unités contenues dans certains sous-complexes d'un assemblage macro-moléculaire.

Étant donnés ces oligomères, l'inférence de connectivité (CI) consiste à inférer les contacts les plus plausibles entre sous-unités. L'inférence de connectivité minimale (MCI) est la variante visant à trouver un ensemble de contacts de taille minimale. Les problèmes MCI évitent d'avoir à spéculer sur le nombre exact de contacts, mais ils conduisent à un sous ensemble de tous les contacts, et ne permettent pas d'exploiter une connaissance à priori sur la plausibilité de ces contacts. Dans ce contexte, nous présentons deux nouveaux algorithmes, `MILP-W` et $\texttt{MILP-W}_{\mathrm{B}}$. Le premier permet de résoudre les problèmes de type *inférence de connectivité à poids minimal* (MWCI), qui sont des problèmes d'optimisation où le critère fait intervenir le nombre de contacts mais aussi un poids sur chacun d'eux. Le second cascade à partir du premier, de façon à améliorer la sensitivité et la specificité des ensembles de solutions générées.

Des simulations sur l'exosome de la levure, système pour lequel sont connus un ensemble d'oligomères mais aussi un structure crystallographique, montrent que nos algorithmes prédisent les contacts avec une grande sensitivité et specificité, le gain par rapport à l'état de l'art étant très substantiel.

Le logiciel accompagnant ce travail est mis à la disposition de la communauté, et devrait s'avérer d'intérêt central pour tous les problèmes d'inférence de connectivité.

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

Contents

1 Connectivity Inference from Sets of Oligomers **6**

2 Minimum Weight Connectivity Inference: Mathematical Model **8**

3 Minimum Weight Connectivity Inference: Algorithms **9**
3.1 Algorithm `MILP-W` . . . 9
3.2 Solutions and consensus solutions . . . 9
3.3 Algorithm $\texttt{MILP-W}_{\mathrm{B}}$ . . . 10

4 Material: Test Systems **11**
4.1 Yeast exosome . . . 11
4.2 Yeast 19S Proteasome lid . . . 11
4.3 Human eIF3 . . . 11

5 Results **12**
5.1 Algorithm $\texttt{MILP-W}_{\mathrm{B}}$ . . . 12
5.2 Algorithm `MILP-W` . . . 13

6 Discussion and Outlook **13**

7 Artwork **15**
7.1 Methods . . . 15
7.2 Yeast Exosome . . . 17
7.3 Yeast Proteasome lid . . . 21
7.4 eIF3 . . . 23

8 Supplemental: Lists of Oligomers for the Assemblies Studied **27**
8.1 Yeast Exosome . . . 27
8.2 Yeast 19S Proteasome lid . . . 27
8.3 Eukaryotic Translation factor eIF3 . . . 28

9 Supplemental: Reference Contacts Within Assemblies **29**
9.1 Pairwise Contacts within Macro-molecular Oligomers . . . 29
9.2 Yeast Exosome . . . 31
9.3 Yeast Proteasome Lid . . . 32
9.4 eIF3 . . . 32

10 Supplemental: Results **33**
10.1 Yeast Exosome . . . 33
10.2 Yeast 19S Proteasome lid . . . 36
10.3 Eukaryotic Translation factor eIF3 . . . 37
10.4 Using Weights: an Illustration . . . 39

11 Supplemental: Algorithms and Programs **42**
11.1 Problem hardness, existing algorithms and contributions . . . 42
11.2 Algorithm $\texttt{MILP-W}_{\mathrm{B}}$: pseudo-code . . . 42
11.3 Implementation . . . 42

12 Supplemental: Using Weights: a Detailed Study **44**
12.1 Methods 44
12.1.1 Deterministic instances 44
12.1.2 Randomized instances 45
12.1.3 Overall recommendations 45
12.2 Results 46

1 Connectivity Inference from Sets of Oligomers

Structural inference from oligomers and contacts. Unraveling the function of macro-molecules and macro-molecular machines requires atomic level data, both in their static and dynamic dimensions, the latter coding for thermodynamic and kinetic properties [SXK+13]. 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), a strategy aiming at producing models of assemblies using complementary experimental data [ADV+07]. In its full generality, RDI accommodates both structural and purely combinatorial data [AFK+08]. 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., sub-complexes of the assembly) can be obtained by methods such as tandem affinity purification [ea01] or native mass spectrometry [SR07, RDD+12], and such oligomers can be complemented by information on pairwise protein - protein interactions [BLJ07, BTD+08]. 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 [SR07]. 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 sub-systems of the assembly under scrutiny [DDC12, DDC13].

Unweighted and weighted connectivity inference: MCI and MWCI. Consider a macro-molecular 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 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 non trivial, 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 [BLJ07, BTD$^+$08]. But information obtained must be used with care for several reasons, notably because expression systems force promiscuity between proteins which may otherwise be located in different cellular compartments, and also because affinity purification typically involves concentration beyond physiological levels. On the *in-silico* side, various interactions attributes can be used, such as gene expressions patterns (proteins with identical patterns are more likely to interact), domain interaction data (a known interaction between two domains hints at an interaction between proteins containing these domains), common neighbors in protein - protein interaction networks, or bibliographical data (number of publications providing evidence for a particular interaction). Here again, these pieces of information have a number of caveats. In particular, structural data from crystallography or mass spectrometry yield a bias towards stable interactions, at the detriment of transient ones. For these reasons, strategies computing confidence scores usually resort to machine learning tools trained on the aforementioned data [YMJ12] and also [TRT$^+$10].

In any case, being able to accommodate such a previous knowledge is precisely another goal of the algorithms developed in this work.

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 [THS$^+$08]. 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 [AAC$^+$13]. 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. (Following our discussion above, note that this algorithm may require exponential time for hard instances, event 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 $\mathrm{Pool_E}(\mathcal{O})$ 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 $\mathrm{Pool_E}(\mathcal{O}_{\leq s})$. 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 [AAC+13].

Weighted case. In the weighted case, each candidate edge e from the pool $\mathrm{Pool_E}(\mathcal{O})$ 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 $\alpha \in [0, 1]$, 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:

$$C(S) = \alpha \sum_{e \in S} 1 + (1 - \alpha) \sum_{e \in S} (1/2 - w(e)) = \sum_{e \in S} C_\alpha(e), \tag{1}$$

with

$$C_\alpha(e) = \frac{\alpha + 1}{2} - (1 - \alpha) w(e). \tag{2}$$

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` [AAC+13], the weights play no role, and the inter-changeability of edges favors the exploration of a large pool of solutions.

- The situation is reversed for small values of α. 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*

$$C_{\alpha,d}(e) = \alpha(1-d) + d - (1-\alpha)w(e). \tag{3}$$

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

3 Minimum Weight Connectivity Inference: Algorithms

3.1 Algorithm `MILP-W`

Algorithm `MILP-W` generalizes the unweighted version `MILP` [AAC+13], 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 [AAC+13] and 0 otherwise.

More precisely, algorithm `MILP-W` iteratively generates all optimal solutions, and adds at each iteration extra constraints preventing from 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 and the associated value OPT for the criterion of Eq. (1). To check whether another solution matching OPT exists, a new constraint preventing the concomitant selection of all edges from the first solution is added. More formally, the sum of the binary variables associated with the solution just produced is 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 to OPT. This process is iterated until the value of the solution exceeds OPT.

Remark 2 *By picking the adequate combination of* α *and* $w(\cdot)$*, 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 $\mathcal{S}_{\text{MILP-W}}$, and the set of contacts used in these solutions is denoted $\mathcal{E}_{\text{MILP-W}}$. The *size of a solution* $S \in \mathcal{S}_{\text{MILP-W}}$, denoted $\mid S \mid$, is its number of contacts. The *score of a contact* appearing in a solution $S \in \mathcal{S}_{\text{MILP-W}}$, called

contact score for short, is the number of solutions from $\mathcal{S}_{\text{MILP-W}}$ containing it. The highest scoring contacts are called the *consensus contacts*, and define the set $\mathcal{E}^{cons.}_{\text{MILP-W}}$. The *score of a solution* $S \in \mathcal{S}_{\text{MILP-W}}$ is the sum of the scores of its contacts. Finally, a *consensus solution* is a solution achieving the maximum score over $\mathcal{S}_{\text{MILP-W}}$. The set of all such solutions being denoted $\mathcal{S}^{cons.}_{\text{MILP-W}}$

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 $\mathcal{S}_{\text{MILP}}, \mathcal{S}^{cons.}_{\text{MILP}}, \mathcal{E}_{\text{MILP}}$ and $\mathcal{E}^{cons.}_{\text{MILP}}$. These notations are summarized in Table 1.

To further assess the quality of the solution set $\mathcal{S}(= \mathcal{S}_{\text{MILP}}, \mathcal{S}_{\text{MILP-W}})$, assume that a reference set of contacts E_{Ref} is known. The ideal situation is that where a high resolution crystal structure is known, since then, all pairwise contacts can be inferred [LC10]. This reference set together with the pool $\text{Pool}_{\text{E}}(\mathcal{O})$ define positive (P), negative (N), and missed contacts (M) (Fig. 2). From these groups, one further classifies the edges of a predicted solution in set $\mathcal{S}$ 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 $\text{ROC}_{sens.}$ and the specificity $\text{ROC}_{spec.}$ as follows:

$$\text{ROC}_{sens.} = \frac{\mid TP \mid}{\mid P \mid}, \qquad \text{ROC}_{spec.} = \frac{\mid TN \mid}{\mid N \mid}. \tag{4}$$

Note that specificity requires the set N to be non empty, which may not be the case if $\text{Pool}_{\text{E}}(\mathcal{O}) \subset \text{E}_{\text{Ref}}$.

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_{Ref} if the pool size is too small):

$$\text{Cvg}(\mathcal{S}) = \frac{\mid TP \mid - (\mid FP \mid + \mid FN \mid)}{\mid \text{E}_{\text{Ref}} \mid} \tag{5}$$

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

3.3 Algorithm `MILP-W`$_{\mathbf{B}}$

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 the way. Implementing this strategy requires two precautions, namely: (i) edges corresponding to dimers must be kept for a solution to be valid, and (ii) hindering too many edges may yield a connectivity inference problem without any solution.

More precisely, we start precluding the consensus contacts i.e. the initial consensus contacts $\mathcal{E}^{cons.}_{\text{MILP-W}}$ minus the dimers, one at a time from the pool of contacts to be explored. We subsequently report the union of consensus contacts (including the initial consensus contacts which we began with) yielded after all the `MILP-W` runs. The process can be iterated by precluding two or more contacts at a time. This strategy triggers rewiring of the system to a greater extent, at the risk of inducing more false positives. (See pseudo code in the supplemental section 11.)

4 Material: Test Systems

We test the performance of the algorithms `MILP-W` and `MILP-W`$_{\mathrm{B}}$ 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, and 19 oligomers [1] have been reported [THS+08], ranging in size from two to nine (Table 2 and supplemental section 8).

Oligomers up to size five are required to encompass 9 out of 10 proteins — the protein Csl4 is present in size nine oligomers only. In terms of contacts, classical interfaces modeling tools [LC10] applied to the crystal structure yield 26 contacts amidst the 10 proteins, and 20 contacts in the assembly depleted of Csl4 (Fig. 4).

The status of Csl4 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 Csl4. In the former case, all oligomers define a pool $\mathrm{Pool}_{\mathrm{E}}(9)$ of 45 candidate edges; in the latter, the pool $\mathrm{Pool}_{\mathrm{E}}(8)$ 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 9 distinct protein types each with unit stoichiometry.

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 8 came from MS, MS/MS and 6 came from cross-linking experiments [STA+06] (Table 3 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. 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 [ZSF+08] (Table 4 and supplemental section 8). The subunit eIF3j is labile and since none of 27 subcomplexes comprises of this subunit we exclude it from the list of protein types, leaving behind 12 protein types.

[1] Originally 21 oligomers are reported including a trivial case of having all 10 protein types and one other oligomer of size 8 has a duplicate, leaving behind 19 distinct oligomers.

5 Results

We first provide results for $\texttt{MILP-W}_\text{B}$ 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 $\texttt{MILP-W}_\text{B}$

As explained in section 3.3, algorithm $\texttt{MILP-W}_\text{B}$ works by accumulating consensus contacts (highest scoring 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 $\mathcal{E}_{\text{MILP}}$, which serve as a baseline.
- (C-2) The statistics for the edge set $\mathcal{E}^{cons.}_{\text{MILP}}$.
- (C-3) The statistics for the edge set $\mathcal{E}_{\text{MILP-W}_\text{B}}$ returned by $\texttt{MILP-W}_\text{B}$ after one iteration.
- (C-4) the statistics for the edge set $\mathcal{E}_{\text{MILP-W}_\text{B}}$ obtained when algorithm $\texttt{MILP-W}_\text{B}$ 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.

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 depends 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 $\texttt{MILP-W}_\text{B}$, one observes that a small number of iterations, typically in the range $1, \ldots, 3$, is favorable to high coverages. This owes to 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 positive – low specificity. Thus, the user may choose the risk level (in terms of false positives) he/she is willing to accepts, 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 [THS$^+$08]. We illustrate this fact with the results produced by $\texttt{MILP-W}_\text{B}$ after one iteration.

On the yeast exosome with Csl4, the sensitivity of $\texttt{MILP-W}_\text{B}$ is ~ 1.67 (=0.77/0.46) times that of network algorithm and *Cvg.* score increases from -0.08 to 0.35 (lines T3 vs T0 in the supplemental Table 8).

For the yeast proteasome, one observes that the sensitivity for $\mathcal{E}_{\text{MILP-W}_\text{B}}$ is 1.76 (=0.74/0.42) times that published earlier [THS$^+$08]. Also, *Cvg.* score increases from -0.21 to 0 (lines T3 vs 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 [ZSF+08]. 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 $\alpha \neq 1$. One naturally expects a benefice in penalizing an edge which is a negative contact and which 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 2), using $\alpha = 0.25$. We assign a weight of 0.6 to the following three contacts - (Rrp45, Rrp46), (Rrp40, Rrp46) 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:

$$\Delta = (\Delta\mathrm{ROC}_{sens.}, \Delta\mathrm{ROC}_{spec.}, \Delta\mathrm{Cvg}). \tag{6}$$

Consider first $\mathcal{E}_{\mathrm{MILP}}$. One observes 4 false positives instead of 5 (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 $\mathcal{E}^{cons.}_{\mathrm{MILP}}$. 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 $\mathcal{E}_{\mathrm{MILP\text{-}W_B}}$, the change in assessment tuple for $\mathcal{E}_{\mathrm{MILP\text{-}W_B}}$ (1st 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 α.

6 Discussion and Outlook

For these reasons, strategies computing confidence scores usually resort to machine learning tools trained on the aforementioned data [YMJ12] and also [TRT+10].

By giving access to a list of overlapping oligomers of a given macro-molecular 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 generalize 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-W`$_{\mathrm{B}}$, a bootstrap strategy aiming at enriching the solutions reported by `MILP-W`. Algorithm `MILP-W`$_{\mathrm{B}}$ accumulates consensus contacts (highest scoring 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-W` terminated within a hand-full of seconds for all the cases processed in this work. Calculations

with algorithm $\texttt{MILP-W}_{\mathrm{B}}$ are more demanding, though, since the run-time 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 is unknown and is 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 sub-units 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.

7 Artwork

7.1 Methods

Figure 1 (Minimum) Connectivity Inference from oligomers and a-priori information on contacts: illustration on a fictitious system. Given an assembly whose subunits are known but pairwise contacts are not, and for which the composition of a number of oligomers in terms of subunits is also known, the problem consists of inferring contacts between subunits. We consider a toy example involving 5 proteins and three oligomers (three trimers), as seen on panel **(A)**). As additional information, one may enforce and/or forbid contacts, and one may also weight contacts, depending on their likelihood. To connect each oligomer using as few edges as possible, two edges must be chosen, out of three possible (panels **(B, C, D)**). The *Minimum Connectivity Inference* consists of finding the overall smallest number of edges such that each each oligomer gets connected. Panel (**(E)**) shows a solution with 4 edges (bold edges). Note that these four edges from a subset of all pairwise contacts.

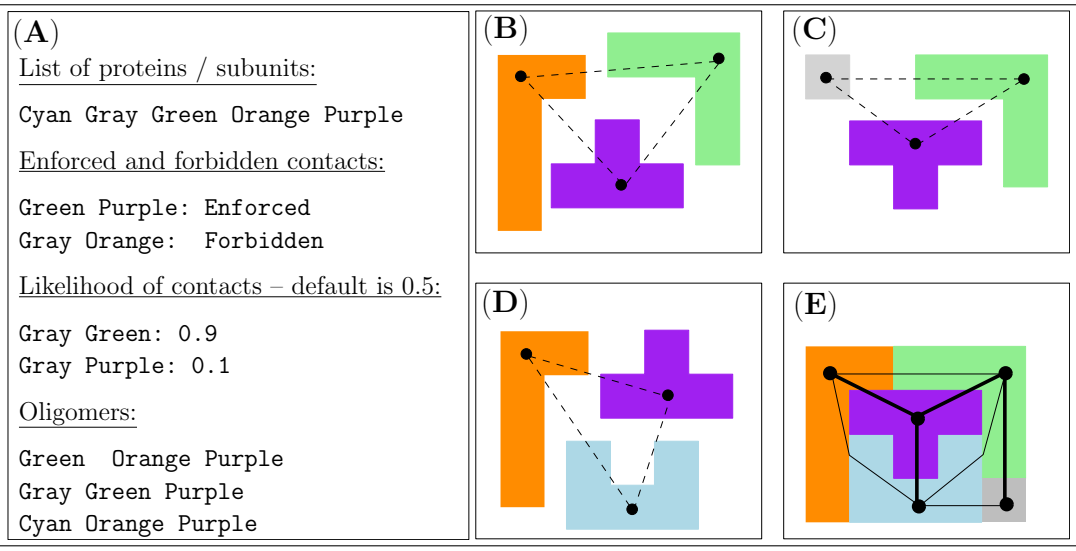
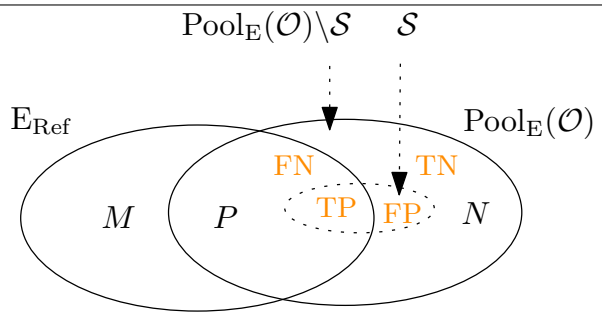


Table 1 Notations for (consensus) solutions and (consensus) edges returned by the algorithms `MILP`, `MILP-W` and `MILP-W`$_\mathbf{B}$.

	solutions	edges	consensus edges	consensus solutions
`MILP`	$\mathcal{S}_{\text{MILP}}$	$\mathcal{E}_{\text{MILP}}$	$\mathcal{E}^{cons.}_{\text{MILP}}$	$\mathcal{S}^{cons.}_{\text{MILP}}$
`MILP-W`	$\mathcal{S}_{\text{MILP-W}}$	$\mathcal{E}_{\text{MILP-W}}$	$\mathcal{E}^{cons.}_{\text{MILP-W}}$	$\mathcal{S}^{cons.}_{\text{MILP-W}}$
`MILP-W`$_\text{B}$	$\mathcal{S}_{\text{MILP-W}_\text{B}}$	$\mathcal{E}_{\text{MILP-W}_\text{B}}$	NA	NA

Figure 2 A pool of candidate $\mathbf{Pool_E}(\mathcal{O})$ and a set of reference contacts $\mathbf{E_{Ref}}$ define positive (P), negative (N), and missed contacts (M). Upon performing a prediction $\mathcal{S}$, $\mathcal{S}$ and its complement $\mathrm{Pool_E}(\mathcal{O})\backslash\mathcal{S}$ further split into true/false × positives/negatives (TP, FP, TN, FN).



7.2 Yeast Exosome

Figure 3 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.

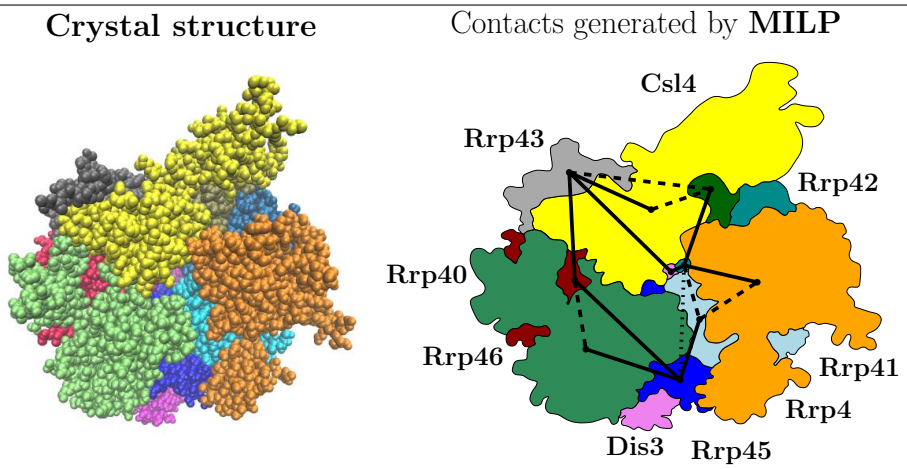


Figure 4 Yeast exosome with and without Csl4.: contacts between subunits. Each edge corresponds to an interface between two subunits. The two numbers decorating an edge respectively refer to the number of atoms involved at that interface, and to the number of patches (connected components) of the interface. Interfaces were computed with the program `intervor`, which implements the Voronoi model from [LC10]. Note that a given subunit makes from three (e.g. Rrp40) to seven (e.g. Rrp45) interfaces.

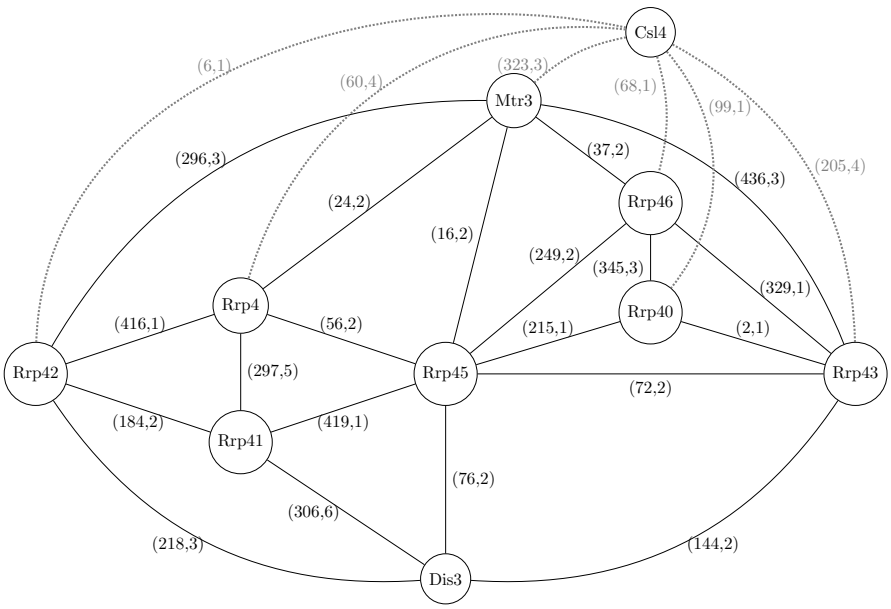


Table 2 Yeast exosome: oligomers and associated statistics. The yeast exosome contains 10 proteins, with Csl4 found in size 9 oligomers only. **(1st column)** Size of oligomers i.e. number of subunits **(2nd column)** Number of oligomers up to a given size **(3rd column)** size of the pool of contacts 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	$\mid \mathcal{O}_{\leq s} \mid$	$\mid \mathrm{Pool_E}(\mathcal{O}_{\leq s}) \mid$	$\lvert M \rvert$
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

Figure 5 Yeast Exosome: Assessment of contacts yielded from different algorithms, `MILP` and `MILP-W`$_{\mathbf{B}}$. See supplemental Table 8 for the detailed statistics.

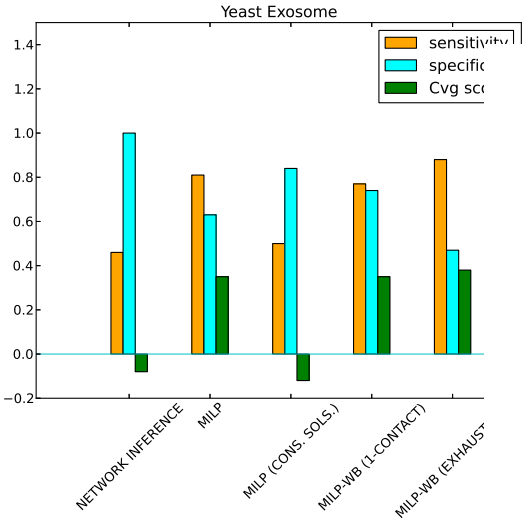


Figure 6 Yeast Exosome: Variation of cumulative sensitivity, specificity and coverage score with iteration index. Note that the iteration index also indicates number of contacts forbidden at a time. See supplemental Table 9 for the detailed statistics.

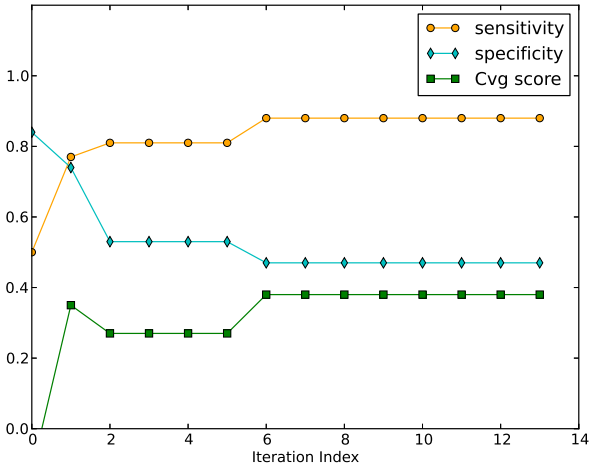


Figure 7 Yeast Exosome without Csl4: Assessment of contacts yielded from different algorithms, `MILP` and `MILP-W`$_{\mathbf{B}}$. See supplemental Table 8 for the detailed statistics.

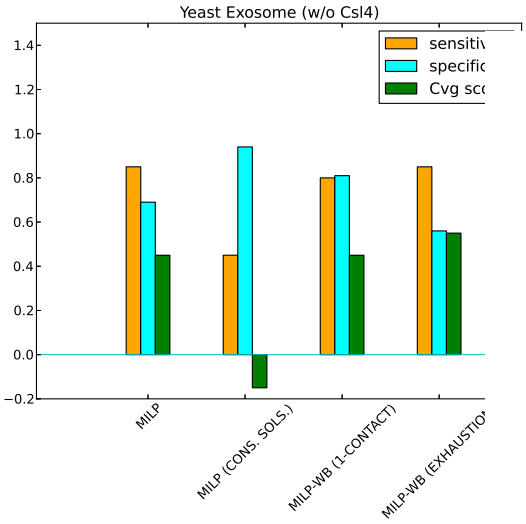
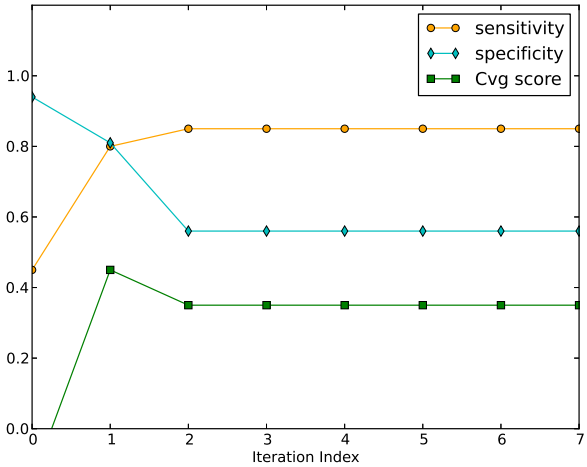


Figure 8 Yeast Exosome without Csl4: Variation of cumulative sensitivity, specificity and coverage score with iteration index. Note that the iteration index also indicates number of contacts forbidden at a time. See supplemental Table 10 for the detailed statistics.



7.3 Yeast Proteasome lid

Figure 9 Yeast proteasome lid: contacts between subunits. Each edge corresponds to an interface between two subunits.

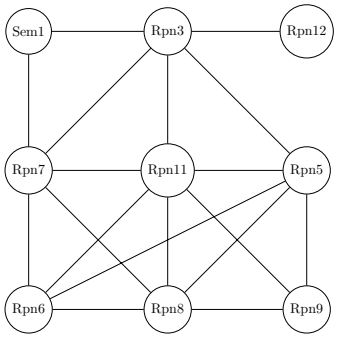


Table 3 Proteasome lid: oligomers and associated statistics. The yeast proteasome lid contains 9 proteins. Note that the pool size is maximal only for $s = 8$, the 14 oligomers yielding the 36 possible contacts. The value $s = 8$ also corresponds to a null number of missed contacts. See supplemental Table 2 for details on the notations.

Oligomer size s	$\mid \mathcal{O}_{\leq s} \mid$	$\mid \mathrm{Pool}_{\mathrm{E}}(\mathcal{O}_{\leq s}) \mid$	$\lvert M \rvert$
2	3	3	16
3	7	10	10
4	9	11	10
5	10	18	5
6	10	18	5
7	11	27	1
8	14	36	0

Figure 10 Yeast Proteasome Lid: Assessment of contacts yielded from different algorithms, `MILP` and `MILP-W`$_{\mathrm{B}}$. See supplemental Table 11 for the detailed statistics.

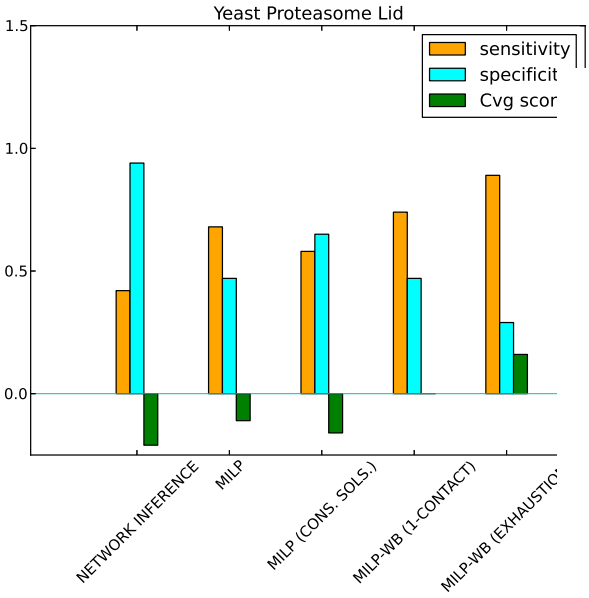
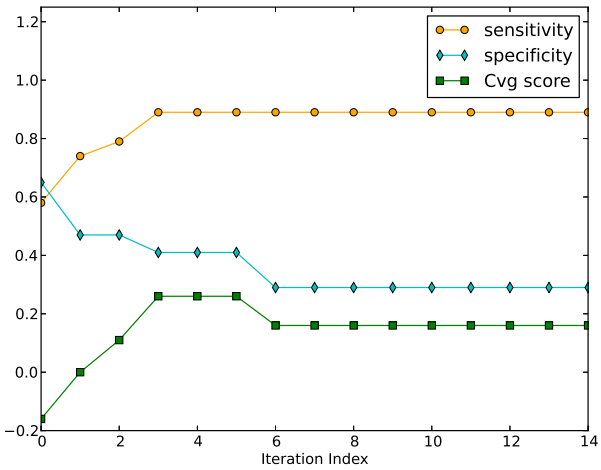


Figure 11 Yeast Proteasome Lid: Variation of cumulative sensitivity, specificity and coverage score with iteration index. Note that the iteration index also indicates number of contacts forbidden at a time. See supplemental Table 12 for the detailed statistics.



7.4 eIF3

Figure 12 Human eIF3: contacts between subunits. Each edge corresponds to an interface between two subunits.

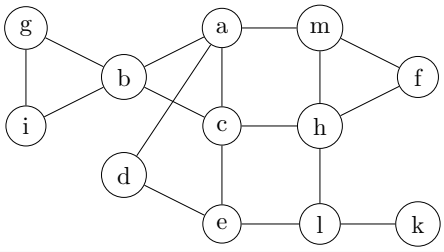


Table 4 Human eIF3: oligomers and associated statistics. The human eIF3 contains 12 proteins without eIF3j (a labile protein not present in the oligomers, see supplemental section 8). Note that the maximal pool size for 12 proteins is $\binom{12}{2} = 66$, however for $s = 11$, the pool size is 60, i.e. sub-maximal. The value $s = 11$ also lacks 2 reference contacts, i.e. $|M| = 2$ See Table 2 for details on the notations.

Oligomer size s	$\mid \mathcal{O}_{\leq s} \mid$	$\mid \mathrm{Pool_E}(\mathcal{O}_{\leq s}) \mid$	$\lvert M \rvert$
2	8	8	9
3	12	11	8
4	15	16	7
5	19	31	2
6	21	36	2
7	24	47	2
8	25	48	2
9	26	58	2
10	26	58	2
11	27	60	2

Figure 13 Human eIF3: Assessment of contacts yielded from different algorithms, `MILP` and `MILP-W`$_{\mathbf{B}}$. See supplemental Table 13 for the detailed statistics.

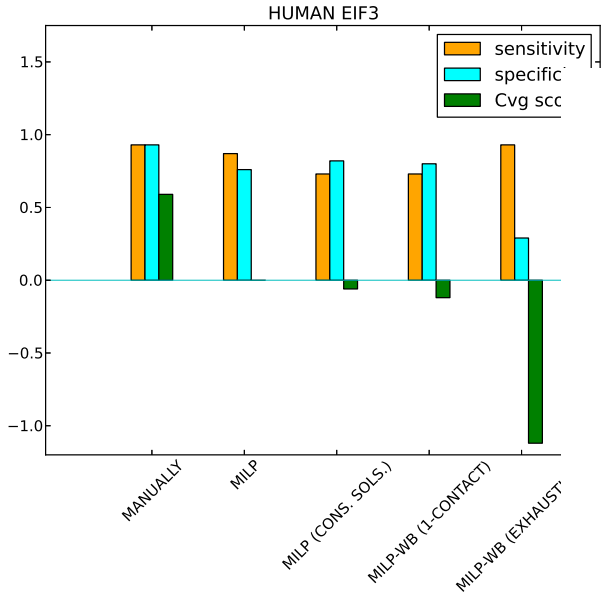
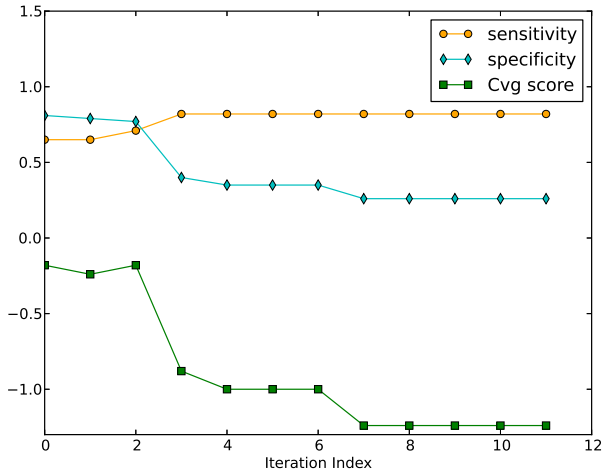


Figure 14 Human eIF3: Variation of cumulative sensitivity, specificity and coverage score with iteration index. Note that the iteration index also indicates number of contacts forbidden at a time. See supplemental Table 14 for the detailed statistics.



References

[AAC+13] D. Agarwal, J. Araujo, C. Caillouet, F. Cazals, D. Coudert, and S. Pérennes. Connectivity inference in mass spectrometry based structure determination. In H.L. Bodlaender and G.F. Italiano, editors, *European Symposium on Algorithms (LNCS 8125)*, pages 289–300, Sophia-Antipolis, France, 2013. Springer.

[ADV+07] 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.

[AFK+08] 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.

[BLJ07] Tord Berggård, Sara Linse, and Peter James. Methods for the detection and analysis of protein–protein interactions. *Proteomics*, 7(16):2833–2842, 2007.

[BTD+08] P. Braun, M. Tasan, M. Dreze, M. Barrios-Rodiles, I. Lemmens, H. Yu, J. Sahalie, R. Murray, L. Roncari, A-S. De Smet, K. Venketesan, J-F. Rual, J. Vandenhaute, M.E. Cusick, T. Pawson, D.E. Hill, J. Tavernier, J.L. Wrana, F.P. Roth, and M. Vidal. An experimentally derived confidence score for binary protein-protein interactions. *Nature methods*, 6(1):91–97, 2008.

[CPBJ06] F. Cazals, F. Proust, R. Bahadur, and J. Janin. Revisiting the Voronoi description of protein-protein interfaces. *Protein Science*, 15(9):2082–2092, 2006.

[DDC12] T. Dreyfus, V. Doye, and F. Cazals. Assessing the reconstruction of macro-molecular assemblies with toleranced models. *Proteins: structure, function, and bioinformatics*, 80(9):2125–2136, 2012.

[DDC13] T. Dreyfus, V. Doye, and F. Cazals. Probing a continuum of macro-molecular assembly models with graph templates of sub-complexes. *Proteins: structure, function, and bioinformatics*, 81(11):2034–2044, 2013.

[ea01] O. Puig et al. The tandem affinity purification method: A general procedure of protein complex purification. *Methods*, 24:218–229, 2001.

[GJ79] M. R. Garey and D. S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman & Co., New York, NY, USA, 1979.

[HDT+06] 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.

[JBC08] J. Janin, R. P. Bahadur, and P. Chakrabarti. Protein-protein interaction and quaternary structure. *Quarterly reviews of biophysics*, 41(2):133–180, 2008.

[KRY+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.

[LC10] S. Loriot and F. Cazals. Modeling macro-molecular interfaces with Intervor. *Bioinformatics*, 26(7):964–965, 2010.

[LEBRT08] E. Levy, E-B.Erba, C. Robinson, and S. Teichmann. Assembly reflects evolution of protein complexes. *Nature*, 453(7199):1262–1265, 2008.

[LEM+12] G. C. Lander, E. Estrin, M. E. Matyskiela, C. Bashore, E. Nogales, and A. Martin. Complete subunit architecture of the proteasome regulatory particle. *Nature*, 482(7384):186–191, 2012.

[LFB+12] 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.

[MBC13] 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.

[QASV+13] J. Querol-Audi, C. Sun, J. M. Vogan, M. D. Smith, Y. Gu, J. HD Cate, and E. Nogales. Architecture of human translation initiation factor 3. *Structure*, 21(6):920–928, 2013.

[RDD+12] R. J. Rose, E. Damoc, E. Denisov, A. Makarov, and A. J. R. Heck. High-sensitivity orbitrap mass analysis of intact macromolecular assemblies. *Nature Methods*, 9(11):1084–1086, 2012.

[SR07] 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.

[STA+06] 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.

[SXK+13] A. Schmidt, H. Xu, A. Khan, T. O'Donnell, S. Khurana, L. King, J. Manischewitz, H. Golding, P. Suphaphiphat, A. Carfi, E. Settembre, P. Dormitzer, T. Kepler, R. Zhang, A. Moody, B. Haynes, H-X. Liao, D. Shaw, and S. Harrison. Preconfiguration of the antigen-binding site during affinity maturation of a broadly neutralizing influenza virus antibody. *Proceedings of the National Academy of Sciences*, 110(1):264–269, 2013.

[THS+08] 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.

[TRT+10] B. Turner, S. Razick, A.L. Turinsky, J. Vlasblom, E. K. Crowdy, E. Cho, K. Morrison, I.M. Donaldson, and S.J. Wodak. irefweb: interactive analysis of consolidated protein interaction data and their supporting evidence. *Database*, 2010:baq023, 2010.

[YMJ12] J. Yu, T. Murali, and R.L. Finley Jr. Assigning confidence scores to protein–protein interactions. In *Two Hybrid Technologies*, pages 161–174. Springer, 2012.

[ZSF+08] 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. *Proc. of the National Academy of Sciences*, 105(47):18139–18144, 2008.

8 Supplemental: Lists of Oligomers for the Assemblies Studied

In this section, we list the composition of the **complexes**, also called **oligomers**, produced experimentally and used as input of our connectivity inference problems.

8.1 Yeast Exosome

The 19 oligomers generated using tandem mass spectrometry and subdenaturing concentration of organic solvents [HDT+06] are the following ones:

```
List of proteins
Csl4 Dis3 Mtr3 Rrp4 Rrp40 Rrp41 Rrp42 Rrp43 Rrp45 Rrp46

List of oligomers
Mtr3 Rrp42
Rrp41 Rrp45
Rrp43 Rrp46
Rrp40 Rrp45 Rrp46
Rrp4 Rrp41 Rrp42 Rrp45
Rrp40 Rrp43 Rrp45 Rrp46
Dis3 Rrp4 Rrp41 Rrp42 Rrp45
Mtr3 Rrp4 Rrp41 Rrp42 Rrp45
Dis3 Mtr3 Rrp4 Rrp41 Rrp42 Rrp45
Dis3 Rrp4 Rrp40 Rrp41 Rrp42 Rrp45 Rrp46
Dis3 Mtr3 Rrp4 Rrp40 Rrp41 Rrp42 Rrp43 Rrp45
Dis3 Mtr3 Rrp4 Rrp40 Rrp41 Rrp42 Rrp45 Rrp46
Dis3 Mtr3 Rrp4 Rrp41 Rrp42 Rrp43 Rrp45 Rrp46
Dis3 Mtr3 Rrp40 Rrp41 Rrp42 Rrp43 Rrp45 Rrp46
Mtr3 Rrp4 Rrp40 Rrp41 Rrp42 Rrp43 Rrp45 Rrp46
Csl4 Dis3 Mtr3 Rrp4 Rrp41 Rrp42 Rrp43 Rrp45 Rrp46
Csl4 Dis3 Mtr3 Rrp40 Rrp41 Rrp42 Rrp43 Rrp45 Rrp46
Csl4 Mtr3 Rrp4 Rrp40 Rrp41 Rrp42 Rrp43 Rrp45 Rrp46
Dis3 Mtr3 Rrp4 Rrp40 Rrp41 Rrp42 Rrp43 Rrp45 Rrp46
```

8.2 Yeast 19S Proteasome lid

The 14 oligomers obtained using MS and MS/MS (8 of them), and cross-linking experiments (6 of them) are as follows [STA+06]:

```
List of proteins
Rpn3 Rpn5 Rpn6 Rpn7 Rpn8 Rpn9 Rpn11 Rpn12 Sem1

List of oligomers
Rpn7 Sem1
Rpn3 Sem1
Rpn3 Rpn5
Rpn3 Rpn5 Rpn8
Rpn5 Rpn6 Rpn8
Rpn5 Rpn8 Rpn9
```

```
Rpn6 Rpn8 Rpn9
Rpn3 Rpn5 Rpn8 Rpn9
Rpn5 Rpn6 Rpn8 Rpn9
Rpn3 Rpn5 Rpn7 Rpn9 Rpn11
Rpn3 Rpn5 Rpn6 Rpn7 Rpn8 Rpn11 Sem1
Rpn3 Rpn5 Rpn6 Rpn7 Rpn8 Rpn9 Rpn11 Sem1
Rpn3 Rpn5 Rpn6 Rpn7 Rpn8 Rpn11 Rpn12 Sem1
Rpn3 Rpn5 Rpn7 Rpn8 Rpn9 Rpn11 Rpn12 Sem1
```

8.3 Eukaryotic Translation factor eIF3

The 27 oligomers obtained using tandem mass spectrometry [ZSF$^+$08] are the following ones:

```
List of proteins
a b c d e f g h i k l m

List of oligomers
b g
d e
e l
f h
f m
g i
h m
k l
b g i
d e l
e k l
f h m
c d e l
c e k l
d e k l
a b c g i
a b e g i
c d e h l
c d e k l
c d e f h l
c d e h k l
c d e f h k l
c d e f h l m
c d e g i k l
c d e f h k l m
b c d e f g h i m
b c d e f g h i k l m
```

Note that the subunit eIF3j is excluded from the list of protein types as there is no subcomplex (oligomer) yielded containing the same.

9 Supplemental: Reference Contacts Within Assemblies

In this section, 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.

9.1 Pairwise Contacts within Macro-molecular Oligomers

Crystal contacts: [$\mathbf{C_{Xtal}}$] 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 [2], two atoms from these pairs define an edge in the α-complex of the assembly for $\alpha = 0$, as classically done to define macro-molecular interfaces [CPBJ06, JBC08, LC10].)

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 [JBC08]. 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 (Table 5). 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 [LEBRT08, JBC08].

Cryo-electron microscopy reconstruction (set $\mathbf{C_{Cryo}}$). Cryo-electron microscopy is a technique to visualize and interpret unstained biological samples including macromolecular assemblies of 200 kDa and more. The biological sample is cryo-fixed to preserve the aqueous environment around the macromolecule thereby, preventing ultrastructural changes, redistribution of elements etc. The imaging is therefore done in near native conditions and using state of art computer controlled microscopy, image reconstruction software, sub-nanometer resolution structures of large biological macromolecular assemblies can be retrieved.

Cross-linking (set $\mathbf{C_{XL}}$). 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 [KRY+12], defining a *low-resolution topology.*

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

[2] Given a van der Waals model, 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) or Tandem Mass spectrometry (MS/MS): 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.

9.2 Yeast Exosome

Table 5 List of contacts determined from experiments for Yeast Exosome. Note in particular the crystal contacts (third column), determined from the crystal structure using the relative atomic positions, using a Voronoi based interface model [LC10]. In general, interfaces involving a large number of atoms are stable ones. Note also that small interfaces in the final product may correspond to interfaces which were large at an early stage of the assembly formation, and which got shrunk along the accretion of molecules.

C_{Xtal} (26 contacts)			C_{Dim} (7 contacts)
X-Ray Crystallography, 2.8 Å [MBC13]			TAP, MS, MS/MS
Chains	Subunits	#Interface atoms	*Partial Denaturation* [HDT+06] [THS+08]
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, MS/MS

Published previously in the panel C of Fig. 4 of [THS+08] a list of the contacts determined using *Network inference* algorithm for the set of oligomers for Yeast exosome in the Section 8. The 12 contacts are:

(Csl4, Rrp43)	(Dis3, Rrp45)	(Mtr3, Rrp42)	(Mtr3, Rrp43)	(Rrp4, Rrp41)
(Rrp4, Rrp42)	(Rrp40, Rrp45)	(Rrp40, Rrp46)	(Rrp41, Rrp42)	(Rrp41, Rrp45)
(Rrp43, Rrp46)	(Rrp45, Rrp46)			

9.3 Yeast Proteasome Lid

Table 6 List of contacts determined from experiments for Yeast 19S Proteasome Lid

C_{Cryo} (13 contacts)	C_{Dim} (3 contacts)	C_{XL} (14 contacts)	
[LEM+12]	MS, MS/MS analysis [STA+06]	$CX-DSSO, DSS, BS3$	*References*
(Rpn3, Rpn5)†	(Rpn5, Rpn8)	(Rpn3, Rpn7)	[KRY+12][LFB+12]
(Rpn3, Rpn8)†	(Rpn6, Rpn8)	(Rpn3, Rpn8)	[KRY+12]
(Rpn3, Rpn12)†	(Rpn8, Rpn9)	(Rpn3, Rpn12)	[KRY+12]
(Rpn5, Rpn6)†		(Rpn3, Sem1)	[KRY+12][STA+06]
(Rpn5, Rpn8)†		(Rpn5, Rpn6)	[KRY+12]
(Rpn5, Rpn9)†		(Rpn5, Rpn9)	[KRY+12][LFB+12]
(Rpn5, Rpn11)		(Rpn6, Rpn7)	[KRY+12]
(Rpn6, Rpn7)†		(Rpn6, Rpn11)	[KRY+12]
(Rpn6, Rpn11)†		(Rpn7, Rpn11)	[KRY+12]
(Rpn7, Rpn8)		(Rpn7, Sem1)	[KRY+12][STA+06]
(Rpn8, Rpn9)†		(Rpn8, Rpn9)	[KRY+12]
(Rpn8 Rpn11)†		(Rpn8, Rpn11)	[KRY+12]
(Rpn9, Rpn11)		(Rpn3, Rpn5)	[STA+06]
		(Rpn3, Rpn11)	[LFB+12]

† signifies those contacts in C_{Cryo} which are also recovered by other biophysical experiments, TAP, MS, MS/MS, *cross-links*

Number of distinct contacts, $|C_{Cryo} \cup C_{Dim} \cup C_{XL}| = 19$

Published previously in the panel B of Fig. 3 of [THS+08], a list of the contacts determined using *Network inference* algorithm for the set of oligomers for Yeast 19S Proteasome lid in the section 8. The 9 contacts are:

(Rpn3, Rpn5) (Rpn3, Rpn11) (Rpn3, Sem1) (Rpn5, Rpn7) (Rpn5, Rpn8)
(Rpn5, Rpn11) (Rpn6, Rpn8) (Rpn7, Sem1) (Rpn8, Rpn9)

9.4 eIF3

Table 7 List of contacts determined from experiments for eIF3.

Name	# contacts	Ref.	List of contacts
C_{Cryo}	15	[QASV+13]	(a, c) (a, d) (a, m) (b, c) (b, g)† (b, i)† (c, e) (c, h) (e, l)† (f, h)† (f, m)† (g, i)† (h, m)† (h, l) (k, l)†
C_{Dim}	10	[ZSF+08]	(a, b) (b, g) (b, i) (d, e) (e, l) (f, h) (f, m) (g, i) (h, m) (k, l)

† signifies those contacts in C_{Cryo} which are also recovered by other biophysical experiments, TAP, MS, MS/MS

Number of distinct contacts, $|C_{Cryo} \cup C_{Dim}| = 17$

Published previously in Fig. 4 of [ZSF+08], a list of the contacts determined manually for the set of oligomers for eIF3 in the Section 8. The 17 contacts are:

(a,b) (a,c) (b,c) (b,e) (b,f) (b,g) (b,i) (c,d) (c,e) (c,h) (d,e) (e,l) (f,h) (f,m) (g,i) (h,m) (k,l)

10 Supplemental: Results

10.1 Yeast Exosome

Results without Csl4. On solving the problem for yeast exosome (without Csl4) using `MILP` (or, `MILP-W` with $\alpha = 1$), one gets 10 consensus contacts in 2 consensus solutions (9 TP and 1 FP) (line with tag T6 in Table 8, and line with $n = 0$ in Table 9). We aim to enrich this initial set of consensus contacts, $\mathcal{E}^{cons.}_{\text{MILP}}$. Among these 10 contacts, we excluded 3 dimers in the set of oligomers (irreplaceable contacts), since, *ipso facto*, they are part of all the solutions, to launch the bootstrap procedure. The contacts from $\mathcal{E}^{cons.}_{\text{MILP}}$ are forbidden one at time by putting the label *forbidden* ('F'), yielding 7 different `MILP` problems each with different pool set (each having $\mid \text{Pool}_{\text{E}}(\mathcal{O}_{\leq s}) \mid = 35$). The union of consensus contacts from 7 runs (including initial consensus contacts) has 19 contacts having 3 FP. The ROC scores are $\text{ROC}_{sens.}$ of 0.80, $\text{ROC}_{spec.}$ of 0.81 and $Cvg.$ score of 0.45 (T7 in Table 8 and row with $n = 1$ in Table 9).

If one goes further by forbidding two contacts at a time, there are 21 possible `MILP` problems each with different pool set (each having $\mid \text{Pool}_{\text{E}}(\mathcal{O}_{\leq s}) \mid = 34$). The union of consensus contacts from these problems has 24 contacts having 7 FP. The cumulative number of contacts on taking the union from the first step of forbidding one contact at a time is also 24. Therefore, ROC scores are – $\text{ROC}_{sens.}$ of 0.85, $\text{ROC}_{spec.}$ of 0.56 and $Cvg.$ score of 0.35 ($n = 2$ in Table 9). These scores remain unchanged when we forbid 3,4,5 contacts at a time. When 6 or 7 contacts are forbidden, there is no solution since the pool set is insufficient ($n = 3, \ldots, 7$ in Table 9). Therefore, the final cumulative ROC score when all possible combinations of the contacts are precluded is $\text{ROC}_{sens.}$ of 0.85, $\text{ROC}_{spec.}$ of 0.56 and $Cvg.$ score of 0.35 (T8 in Table 8).

Results with Csl4. The complete system involves 10 proteins and 19 set of oligomers. The initial consensus set has 13 TP and 3 FP (T2 in Table 9) out of which 3 are dimers (irreplaceable contacts). On forbidding one contact at a time union of consensus contacts from 13 different `MILP` runs is 25 contacts with 20 TP and 5 FP. The corresponding ROC scores are i.e. $\text{ROC}_{sens.}$ of 0.77, $\text{ROC}_{spec.}$ of 0.74 and $Cvg.$ score of 0.35 (T3 in Table 8).

On forbidding further upto 11 contacts, the union of consensus contacts is 33 with 23 TP and 10 FP. Forbidding 12 and 13 contacts do not yield any solutions due to insufficient number of contacts. The ROC scores at the end, therefore, are $\text{ROC}_{sens.}$ of 0.88, $\text{ROC}_{spec.}$ of 0.47 and $Cvg.$ score of 0.38 (T4 in Table 8).

Assessment. Precluding the initial consensus contacts simultaneously yields new consensus contacts. We observe that the bootstrapping procedure has served its purpose which is to enrich the initial consensus contacts, $\mathcal{E}^{cons.}_{\text{MILP}}$ by possibly further sampling more TP and less FP. For $s = 8$, number of TP increases from 9 to 16 with 2 additional FP (T6 and T7 in Table 8). Similarly, for $s = 9$, TP increases from 13 to 20 with additional 2 FP (T2 and T3 in Table 8).

We also see that the bootstrapping procedure, $\texttt{MILP-W}_{\text{B}}$, which essentially extends the initial consensus contact set by including consensus contacts (high scoring contacts with high specificity) that are found in the way, in the end have comparable sensitivity and improved sensitivity to that of the contact sets yielded by `MILP`. For $s = 8$, for $\mathcal{E}_{\text{MILP}}$ and $\mathcal{E}_{\text{MILP-W}_{\text{B}}}$, respectively, the sensitivities are 0.85 and 0.80, whereas, the specificities are 0.69 and 0.81 (T5 vs T7 in Table 8). These numbers for $s = 9$ are 0.81 and 0.77 and 0.63 and 0.74 (T1 vs T3 in Table 8).

Note, that precluding more number of contacts though increase sensitivity as more TP are sampled but hurt the specificity as well as a result of sampling of FP (T4 and T8 in Table 8). This is due to rewiring of the system to a larger extent on forbidding large number of consensus contacts.

Finally, the performances are excellent when compared against those of the heuristic network algorithm [THS$^+$08]. On the yeast exosome with Csl4, the sensitivity of `MILP-W`$_{\mathrm{B}}$ is ~ 1.67 times that of network algorithm and *Cvg*. score increases from -0.08 to 0.35 (T3 vs T0 in Table 8).

Table 8 Yeast exosome: sensitivity, specificity and coverage for various edge sets generated by `MILP` and `MILP-W`$_{\mathbf{B}}$. Results from T0-T4 corresponds to Yeast exosome including Csl4 and from T5-T8 corresponds to Yeast exosome without Csl4. For a given run (each line), all edges predicted get distributed into TP and FP. In the following paragraph, the content in the bracket correspond to yeast exosome without Csl4. Out of a pool of candidate edges of size 45 (36), the edge set $\mathcal{E}_{\text{MILP-W}_\text{B}}$ (1st iteration) contains all true positives but six (but four), and five false positives (three false positives). It is to be noted that tag T3 (T7) corresponds to the 1st iteration of the Table 9 (Table 10), while tag T4 (T8) corresponds to results obtained upon precluding all possible combinations of initial consensus contacts.

Tag	algo	s	$\lvert \text{Pool}_\text{E}(\mathcal{O}_{\leq s}) \rvert$	M	P	TP	FN	N	TN	FP	ROC$_{sens.}$	ROC$_{spec.}$	Cvg
(T0)	*Network inference* [THS+08]	9	45	0	26	12	14	19	19	0	0.46	1	-0.08
(T1)	$\mathcal{E}_{\text{MILP}}$	9	45	0	26	21	5	19	12	7	0.81	0.63	0.35
(T2)	$\mathcal{E}^{cons.}_{\text{MILP}}$	9	45	0	26	13	13	19	16	3	0.50	0.84	-0.12
(T3)	$\mathcal{E}_{\text{MILP-W}_\text{B}}$, 1-contact	9	45	0	26	20	6	19	14	5	0.77	0.74	0.35
(T4)	$\mathcal{E}_{\text{MILP-W}_\text{B}}$, after exhaustion	9	45	0	26	23	3	19	9	10	0.88	0.47	0.38
(T5)	$\mathcal{E}_{\text{MILP}}$	8	36	0	20	17	3	16	11	5	0.85	0.69	0.45
(T6)	$\mathcal{E}^{cons.}_{\text{MILP}}$	8	36	0	20	9	11	16	15	1	0.45	0.94	-0.15
(T7)	$\mathcal{E}_{\text{MILP-W}_\text{B}}$, 1-contact	8	36	0	20	16	4	16	13	3	0.80	0.81	0.45
(T8)	$\mathcal{E}_{\text{MILP-W}_\text{B}}$, after exhaustion	8	36	0	20	17	3	16	9	7	0.85	0.56	0.35

Table 9 Yeast exosome: sensitivity, specificity and coverage of enriched consensus set on forbidding a number of initial consensus contacts by `MILP-W`$_{\mathbf{B}}$. Note that the cumulative statistics for row n is computed by considering union of all the consensus edge sets, $\mathcal{E}^{cons.}_{\text{MILP-W}_\text{B}}$ from 0 to $n = 13$.

#contacts fobidden, n	#combinations, $(\binom{13}{n})$	$\lvert\mathcal{E}^{cons.}_{\text{MILP-W}_\text{B}}\rvert$	$\lvert\mathcal{E}^{cons.}_{\text{MILP-W}_\text{B}}\rvert^{Cum.}$	individual ROC$_{sens.}$	individual ROC$_{spec.}$	individual Cvg	cumulative ROC$_{sens.}$	cumulative ROC$_{spec.}$	cumulative Cvg
0	1	16	16	0.50	0.84	-0.12	0.50	0.84	-0.12
1	13	25	25	0.77	0.74	0.35	0.77	0.74	0.35
2	78	30	30	0.81	0.53	0.27	0.81	0.53	0.27
3	286	30	30	0.81	0.53	0.27	0.81	0.53	0.27
4	715	30	30	0.81	0.53	0.27	0.81	0.53	0.27
5	1287	30	30	0.81	0.53	0.27	0.81	0.53	0.27
6	1716	33	33	0.88	0.47	0.38	0.88	0.47	0.38
7	1716	33	33	0.88	0.47	0.38	0.88	0.47	0.38
8	1287	33	33	0.88	0.47	0.38	0.88	0.47	0.38
9	715	33	33	0.88	0.47	0.38	0.88	0.47	0.38
10	286	33	33	0.88	0.47	0.38	0.88	0.47	0.38
11	78	25	33	0.77	0.74	0.35	0.88	0.47	0.38
12	13	0	33	-	-	-	0.88	0.47	0.38
13	1	0	24	-	-	-	0.88	0.47	0.38

Table 10 Yeast exosome without Csl4: sensitivity, specificity and coverage of enriched consensus set on forbidding a number of initial consensus contacts by `MILP-W`$_{\mathbf{B}}$. Note that the cumulative statistics for row n is computed by considering union of all the consensus edge sets, $\mathcal{E}^{cons.}_{\text{MILP-W}_\text{B}}$ from 0 to n.

#contacts fobidden, n	#combinations, $(\binom{7}{n})$	$\lvert\mathcal{E}^{cons.}_{\text{MILP-W}_\text{B}}\rvert$	$\lvert\mathcal{E}^{cons.}_{\text{MILP-W}_\text{B}}\rvert^{Cum.}$	individual			cumulative		
				ROC$_{sens.}$	ROC$_{spec.}$	Cvg	ROC$_{sens.}$	ROC$_{spec.}$	Cvg
0	1	10	10	0.45	0.94	-0.15	0.45	0.94	-0.15
1	7	19	19	0.80	0.81	0.45	0.80	0.81	0.45
2	21	24	24	0.85	0.56	0.35	0.85	0.56	0.35
3	35	24	24	0.85	0.56	0.35	0.85	0.56	0.35
4	35	24	24	0.85	0.56	0.35	0.85	0.56	0.35
5	21	24	24	0.85	0.56	0.35	0.85	0.56	0.35
6	7	0	24	-	-	-	0.85	0.56	0.35
7	1	0	24	-	-	-	0.85	0.56	0.35

10.2 Yeast 19S Proteasome lid

Results. This system involves 9 proteins and 14 oligomers. The initial consensus set using `MILP` (or `MILP-W` with $\alpha = 1$) has 11 TP and 6 FP out of which 3 are dimers (irreplaceable contacts) (line with tag T2 of Table 11, line with $n = 0$ in Table 12). On forbidding the contacts, $\mathcal{E}^{cons.}_{\text{MILP}}$ one at a time, one has 14 different `MILP` problems each with different pool set of size of 35. The union of consensus contacts of all such problems has 23 contacts having 14 TP and 9 FP. The ROC scores, therefore, are ROC$_{sens.}$ of 0.74, ROC$_{spec.}$ of 0.47 and $Cvg.$ score of 0 (T3 of the Table 11 and row with $n = 1$ of the Table 12). Proceeding further by precluding two contacts at a time, the size of union of consensus contacts yielded is 24 and cumulative union size (taking into account the previous step) is also 24. The ROC scores are - ROC$_{sens.}$ of 0.79, ROC$_{spec.}$ of 0.47 and $Cvg.$ score of 0.11 ($n = 2$ in Table 12). When three to five contacts are precluded then the following cumulative scores for 27 contacts are – ROC$_{sens.}$ of 0.89, ROC$_{spec.}$ of 0.41 and $Cvg.$ score of 0.26 ($n = 3, \ldots, 5$ in Table 12). When six contacts are precluded more FP are induced yielding scores – ROC$_{sens.}$ of 0.89, ROC$_{spec.}$ of 0.29 and $Cvg.$ score of 0.16. Beyond this point the cumulative scores do not change when number of contacts are precluded from 4 to 14 at a time ($n = 6, \ldots, 14$ in Table 12).

Assessment. The bootstrapping algorithm `MILP-W`$_{\text{B}}$ enriched the initial consensus contacts $\mathcal{E}^{cons.}_{\text{MILP}}$ by augmenting TP from 11 to 14 with 3 additional FP (T2 and T3 of the Table 11). On comparing with $\mathcal{E}_{\text{MILP}}$, we observe that for $\mathcal{E}_{\text{MILP-W}_\text{B}}$ the number of FP are same but number of TP is one more, thus improved sensitivity.

We again observe that on precluding more consensus contacts at a time, the specificity is dropped. The final cumulative score is ROC$_{sens.}$ of 0.89, ROC$_{spec.}$ of 0.29 and $Cvg.$ score of 0.16 (T4 of the Table 11 and Table 12).

Finally, when we compare with previously published contacts by *Network inference* algorithm in [THS$^+$08], we observe that the sensitivity for $\mathcal{E}_{\text{MILP-W}_\text{B}}$ is 1.76 higher than those published earlier. Also, $Cvg.$ score increases from -0.21 to 0 (T3 vs T0 in the Table 11).

Table 11 Yeast proteasome lid: sensitivity, specificity and coverage for various edge sets generated by `MILP` and `MILP-W`$_\mathbf{B}$. For a given run (each line), all edges predicted get distributed into TP and FP. Out of a pool of candidate edges of size 36, the edge set $\mathcal{E}_{\text{MILP-W}_\text{B}}$ (1st iteration) contains all true positive but five, and nine false positives. It is to be noted that tag T3 corresponds to the 1st iteration of the Table 12, while tag T4 corresponds to results obtained upon precluding all possible combinations of initial consensus contacts.

Tag	algo	s	$\lvert \text{Pool}_\text{E}(\mathcal{O}_{\leq s}) \rvert$	M	P	TP	FN	N	TN	FP	$\text{ROC}_{sens.}$	$\text{ROC}_{spec.}$	Cvg
(T0)	*Network inference* [THS+08]	8	36	0	19	8	11	17	16	1	0.42	0.94	-0.21
(T1)	$\mathcal{E}_{\text{MILP}}$	8	36	0	19	13	6	17	8	9	0.68	0.47	-0.11
(T2)	$\mathcal{E}^{cons.}_{\text{MILP}}$	8	36	0	19	11	8	17	11	6	0.58	0.65	-0.16
(T3)	$\mathcal{E}_{\text{MILP-W}_\text{B}}$, 1-contact	8	36	0	19	14	5	17	8	9	0.74	0.47	0
(T4)	$\mathcal{E}_{\text{MILP-W}_\text{B}}$, After exhaustion	8	36	0	19	17	2	17	5	12	0.89	0.29	0.16

Table 12 Yeast proteasome assembly: sensitivity, specificity and coverage of enriched consensus set on forbidding a number of initial consensus contacts by `MILP-W`$_\mathbf{B}$. Note that the cumulative statistics for row n is computed by considering union of all the consensus edge sets, $\mathcal{E}^{cons.}_{\text{MILP-W}_\text{B}}$ from 0 to n.

#contacts fobidden, n	#combinations, $(\binom{14}{n})$	$\lvert\mathcal{E}^{cons.}_{\text{MILP-W}_\text{B}}\rvert$	$\lvert\mathcal{E}^{cons.}_{\text{MILP-W}_\text{B}}\rvert^{Cum.}$	individual			cumulative		
				$\text{ROC}_{sens.}$	$\text{ROC}_{spec.}$	Cvg	$\text{ROC}_{sens.}$	$\text{ROC}_{spec.}$	Cvg
0	1	17	17	0.58	0.65	-0.16	0.58	0.65	-0.16
1	14	23	23	0.74	0.47	0	0.74	0.47	0
2	91	24	24	0.79	0.47	0.11	0.79	0.47	0.11
3	364	27	27	0.89	0.41	0.26	0.89	0.41	0.26
4	1001	27	27	0.89	0.41	0.26	0.89	0.41	0.26
5	2002	27	27	0.89	0.41	0.26	0.89	0.41	0.26
6	3003	29	29	0.89	0.29	0.16	0.89	0.29	0.16
7	3432	29	29	0.89	0.29	0.16	0.89	0.29	0.16
8	3003	29	29	0.89	0.29	0.16	0.89	0.29	0.16
9	2002	29	29	0.89	0.29	0.16	0.89	0.29	0.16
10	1001	29	29	0.89	0.29	0.16	0.89	0.29	0.16
11	364	29	29	0.89	0.29	0.16	0.89	0.29	0.16
12	91	28	29	0.84	0.29	0.05	0.89	0.29	0.16
13	14	0	29	-	-	-	0.89	0.29	0.16
14	1	0	29	-	-	-	0.89	0.29	0.16

10.3 Eukaryotic Translation factor eIF3

Results and Assessment. Regarding the reference contacts for human eIF3 we have cryo-EM reconstruction and MS, MS/MS dimers. We do not have cross-linking contacts for human eIF3. Therefore, the set of reference contacts is possibly not exhaustive. Also, pool set of contacts is sub-maximal since the size is 60 instead of 66 (for 12 vertices) and maximum 15 out of 17 positives could be sampled from the pool set (Table 4).

However, the behavior of $\mathcal{E}_{\text{MILP-W}_\text{B}}$ viz-a-viz $\mathcal{E}^{cons.}_{\text{MILP}}$ and $\mathcal{E}_{\text{MILP}}$ resembles that of the previous two systems. Using bootstrapping procedure we report 20 contacts with 11 TP and 9 FP. The contact set $\mathcal{E}_{\text{MILP-W}_\text{B}}$ has one more additional FP than that of $\mathcal{E}^{cons.}_{\text{MILP}}$ (T3 vs T2 in the Table 13). However, the specificity is still better than that of $\mathcal{E}_{\text{MILP}}$ (T3 vs T1 in the Table 13). We observe that precluding more than 1 contact at a time in `MILP-W`$_\text{B}$ yields low specificity (T3 vs T4 in the Table 13 and Table 14).

The previously published contacts in [ZSF$^+$08] are computed manually using experimental information from various other sources (T0 in the Table 13).

Table 13 Sensitivity, specificity and coverage for various edge sets generated by `MILP` and `MILP-W`$_\mathbf{B}$ for human eIF3 assembly. For a given run (each line), all edges predicted get distributed into TP and FP. Out of a pool of candidate edges of size 60, the edge set $\mathcal{E}_{\text{MILP-W}_\text{B}}$ (1st iteration) contains all true positive but four, and nine false positives. It is to be noted that tag T3 corresponds to the 1st iteration of the Table 14, while tag T4 corresponds to results obtained upon precluding all possible combinations of initial consensus contacts.

Tag	algo	s	$\lvert \text{Pool}_\text{E}(\mathcal{O}_{\leq s}) \rvert$	M	P	TP	FN	N	TN	FP	ROC$_{sens.}$	ROC$_{spec.}$	Cvg
(T0)	*Manually* [ZSF$^+$08]	11	60	2	15	14	1	45	42	3	0.93	0.93	0.59
(T1)	$\mathcal{E}_{\text{MILP}}$	11	60	2	15	13	2	45	34	11	0.87	0.76	0
(T2)	$\mathcal{E}^{cons.}_{\text{MILP}}$	11	60	2	15	11	4	45	37	8	0.73	0.82	-0.06
(T3)	$\mathcal{E}_{\text{MILP-W}_\text{B}}$, 1-contact	11	60	2	15	11	4	45	36	9	0.73	0.80	-0.12
(T4)	$\mathcal{E}_{\text{MILP-W}_\text{B}}$, After exhaustion	11	60	2	15	14	1	45	13	32	0.93	0.29	-1.12

Table 14 Sensitivity, specificity and coverage of enriched consensus set on forbidding a number of initial consensus contacts by `MILP-W`$_\mathbf{B}$ for human eIF3 assembly. Note that the cumulative statistics for row n is computed by considering union of all the consensus edge sets, $\mathcal{E}^{cons.}_{\text{MILP-W}_\text{B}}$ from 0 to n.

#contacts fobidden, n	#combinations, $(\binom{11}{n})$	$\lvert\mathcal{E}^{cons.}_{\text{MILP-W}_\text{B}}\rvert$	$\lvert\mathcal{E}^{cons.}_{\text{MILP-W}_\text{B}}\rvert^{Cum.}$	individual			cumulative		
				ROC$_{sens.}$	ROC$_{spec.}$	Cvg	ROC$_{sens.}$	ROC$_{spec.}$	Cvg
0	1	19	19	0.65	0.81	-0.18	0.65	0.81	-0.18
1	11	20	20	0.65	0.79	-0.24	0.65	0.79	-0.24
2	55	22	22	0.71	0.77	-0.18	0.71	0.77	-0.18
3	165	40	40	0.82	0.40	-0.88	0.82	0.40	-0.88
4	330	42	42	0.82	0.35	-1.00	0.82	0.35	-1.00
5	462	42	42	0.82	0.35	-1.00	0.82	0.35	-1.00
6	462	42	42	0.82	0.35	-1.00	0.82	0.35	-1.00
7	330	46	46	0.82	0.26	-1.24	0.82	0.26	-1.24
8	165	46	46	0.82	0.26	-1.24	0.82	0.26	-1.24
9	55	46	46	0.82	0.26	-1.24	0.82	0.26	-1.24
10	11	46	46	0.82	0.26	-1.24	0.82	0.26	-1.24
11	1	46	46	0.82	0.26	-1.24	0.82	0.26	-1.24

10.4 Using Weights: an Illustration

Figure 15 Yeast Exosome: Contact distribution with [Left] $\alpha = 0.25$ **and [Right]** $\alpha = 1.0$**.** Note that the contacts (Rrp45, Rrp46), (Rrp40, Rrp46) and (Rrp41, Rrp42) are assigned weights of 0.6 and rest are left at default (0.5).

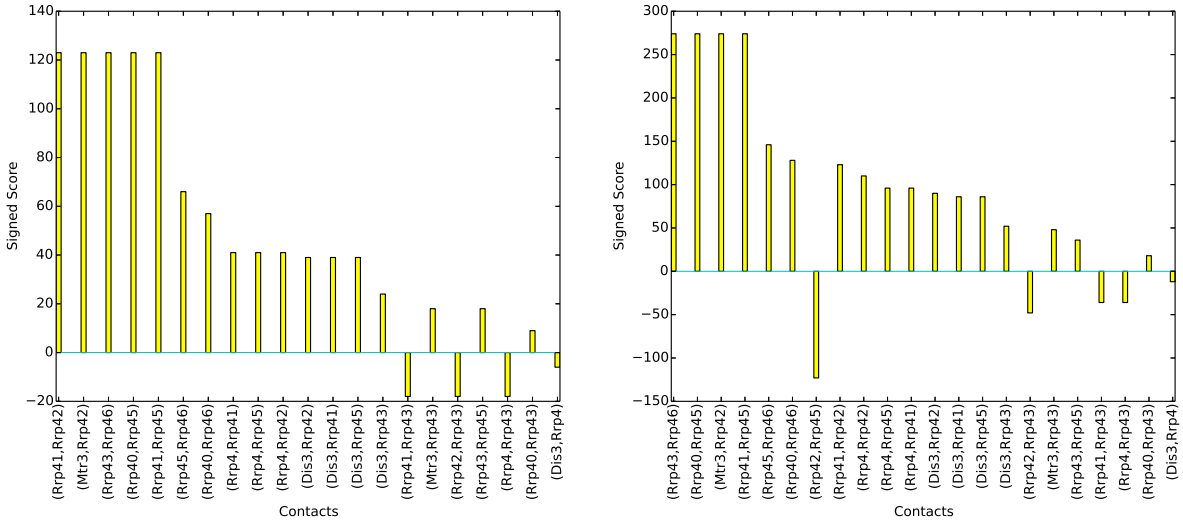


Figure 16 Yeast Exosome (without Csl4): Assessment of contacts yielded from different algorithms, `MILP` and `MILP-W`$_{\mathbf{B}}$ with $\alpha = 0.25$. See supplemental Table 15 for the detailed statistics. Note that the contacts (Rrp45, Rrp46), (Rrp40, Rrp46) and (Rrp41, Rrp42) are assigned weights of 0.6 and rest are left at default (0.5).

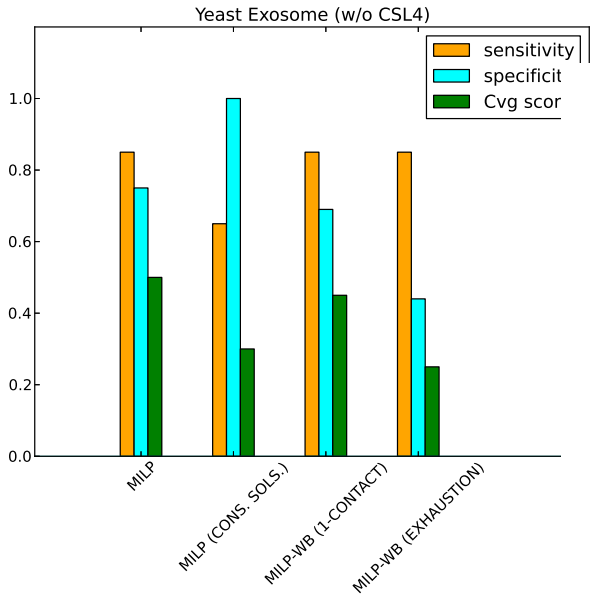


Figure 17 Yeast Exosome (without Csl4): Evolution of cumulative sensitivity, specificity and coverage score with iteration index; $\alpha = 0.25$. Note that the iteration index also indicates number of contacts forbidden at a time. See supplemental Table 16 for the detailed statistics. Note that the contacts (Rrp45, Rrp46), (Rrp40, Rrp46) and (Rrp41, Rrp42) are assigned weights of 0.6 and rest are left at default (0.5).

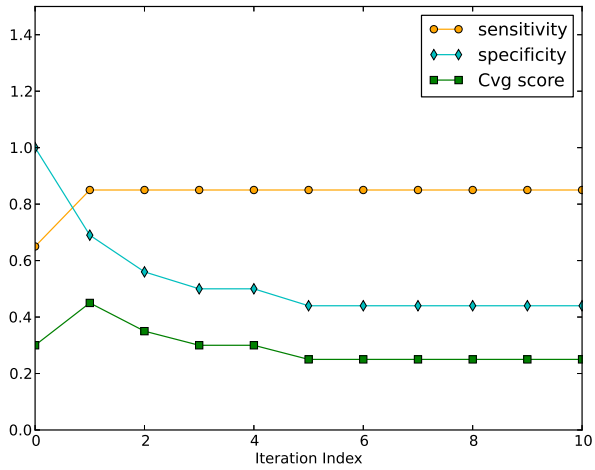


Table 15 Sensitivity, specificity and coverage for various edge sets generated by `MILP` and `MILP-W`$_{\mathbf{B}}$ for yeast exosome (without Csl4) assembly; $\alpha = 0.25$. Note that the contacts (Rrp45, Rrp46), (Rrp40, Rrp46) and (Rrp41, Rrp42) are assigned weights of 0.6 and rest are left at default (0.5). For a given run (each line), all edges predicted get distributed into TP and FP. Out of a pool of candidate edges of size 36, the edge set $\mathcal{E}_{\text{MILP-W}_\text{B}}$ (1st iteration) contains all true positive but three, and five false positives. It is to be noted that tag T3 corresponds to the 1st iteration of the Table 16, while tag T4 corresponds to results obtained upon precluding all possible combinations of initial consensus contacts.

Tag	algo	s	$\lvert \text{Pool}_\text{E}(\mathcal{O}_{\leq s}) \rvert$	M	P	TP	FN	N	TN	FP	$\text{ROC}_{sens.}$	$\text{ROC}_{spec.}$	Cvg
(T1)	$\mathcal{E}_{\text{MILP}}$	8	36	0	20	17	3	16	12	4	0.85	0.75	0.5
(T2)	$\mathcal{E}^{cons.}_{\text{MILP}}$	8	36	0	20	13	7	16	16	0	0.65	1.0	0.3
(T3)	$\mathcal{E}_{\text{MILP-W}_\text{B}}$, 1-contact	8	36	0	20	17	3	16	11	5	0.85	0.69	0.45
(T4)	$\mathcal{E}_{\text{MILP-W}_\text{B}}$, After exhaustion	8	36	0	20	17	3	16	7	9	0.85	0.44	0.25

Table 16 Yeast exosome (without Csl4): sensitivity, specificity and coverage of enriched consensus set on forbidding a number of initial consensus contacts by `MILP-W`$_{\mathbf{B}}$; $\alpha = 0.25$. The contacts (Rrp45, Rrp46), (Rrp40, Rrp46) and (Rrp41, Rrp42) are assigned weights of 0.6 and rest are left at default (0.5). Note that the cumulative statistics for row n is computed by considering union of all the consensus edge sets, $\mathcal{E}^{cons.}_{\text{MILP-W}_\text{B}}$ from 0 to n.

#contacts fobidden, n	#combinations, $(\binom{10}{n})$	$\lvert\mathcal{E}^{cons.}_{\text{MILP-W}_\text{B}}\rvert$	$\lvert\mathcal{E}^{cons.}_{\text{MILP-W}_\text{B}}\rvert^{Cum.}$	individual			cumulative		
				ROC$_{sens.}$	ROC$_{spec.}$	Cvg	ROC$_{sens.}$	ROC$_{spec.}$	Cvg
0	1	13	13	0.65	1	0.3	0.65	1	0.3
1	10	22	22	0.85	0.69	0.45	0.85	0.69	0.45
2	45	24	24	0.85	0.56	0.35	0.85	0.56	0.35
3	120	25	25	0.85	0.50	0.30	0.85	0.50	0.30
4	210	23	25	0.85	0.63	0.40	0.85	0.50	0.30
5	252	24	26	0.85	0.56	0.35	0.85	0.44	0.25
6	210	24	26	0.85	0.56	0.35	0.85	0.44	0.25
7	120	24	26	0.85	0.56	0.35	0.85	0.44	0.25
8	45	24	26	0.85	0.56	0.35	0.85	0.44	0.25
9	10	-	-	-	-	-	0.85	0.44	0.25
10	1	-	-	-	-	-	0.85	0.44	0.25

11 Supplemental: Algorithms and Programs

11.1 Problem hardness, existing algorithms and contributions

Assessing the intrinsic difficulty of a combinatorial problem requires inspecting the *decision* and the *optimization* versions of the problem [GJ79]. In our case, deciding whether a MCI problem admits a solution using a pre-defined number of edges k is **NP**-complete, while finding the solution of smallest size is APX-hard. This latter result is of special interest since we aim at finding an edge set of minimal size. It stipulates that unless $\mathbf{P} = \mathbf{NP}$, there does not exist any polynomial time approximation scheme [AAC+13], that is, a polynomial time algorithm reporting an edge set *as close as desired*, in terms of size, from the optimum. It should be stressed that these facts do not exploit any peculiar property of real data, and only show the existence of hard i.e. difficult to solve instances.

11.2 Algorithm MILP-$\mathrm{W_B}$: pseudo-code

Algorithm 1 Algorithm MILP-$\mathrm{W_B}$, with initial call MILP-$\mathrm{W_B}$ ($\mathcal{E}^{cons.}_{\mathbf{MILP}} \backslash I$), with I standing for the list of dimers in the list of oligomers. The algorithm bootstraps from consensus edges, and collects novel consensus edges which appear upon precluding already found consensus edges.

Require: $\mathcal{E}^{cons.}_{\mathrm{MILP}}$ – initial consensus edges.
Require: I – irreplaceable contacts (dimers).
Require: $spec_0$ – the initial connectivity inference specification.
Require: $\mathcal{E}_{\text{MILP-W}_\text{B}}$ – the set storing all consensus edges.
Parameter: B – consensus edges to be precluded $:= \mathcal{E}^{cons.}_{\mathrm{MILP}} \setminus I$

Algorithm: MILP-$\mathrm{W_B}$ (B)

```
1: E_MILP-WB ← E^cons._MILP
2: /* NB: an iteration precludes all possible ℓ-tuples */
3: for ℓ from 1 to |B| do
4:    Get fsets_ℓ: all ℓ-tuples from the set B, namely (B choose ℓ)
5:    /* C_ℓ: A set of consensus contacts that will be generated for all ℓ-tuples is initiated to
      E^cons._MILP */
6:    C_ℓ ← ∅
7:    for each fset ∈ fsets_ℓ do
8:       /* Edit the initial specification spec_0 to take into account the annotations */
9:       Assign label forbidden ('F') to all contacts in fset
10:      Run MILP-W for this novel specification
11:      Get consensus contacts associated with fset, denoted C_fset
12:      C_ℓ ← C_ℓ ∪ C_fset
13:   E_MILP-WB = E_MILP-WB ∪ C_ℓ
```

11.3 Implementation

Our algorithms have been implemented using IBM CPLEX solver 12.6. The typical running time required to solve an instance presented in this paper is circa 30 seconds, on a standard laptop computer (2.80GHz Intel(R) Xeon(R) CPU E5-1603 0).

Upon publication of this paper, the programs implementing `MILP`, MILP-W, and MILP-W_B will be distributed within the *Structural Bioinformatics Library* (`http://structural-bioinformatics-library.org/`).

12 Supplemental: Using Weights: a Detailed Study

12.1 Methods

In the sequel, we present a thorough evaluation of `MILP-W` upon varying weights and the value of α – Eq. (1).

To this end, we challenge algorithm `MILP-W` with two classes of instances. While *deterministic instances* are meant to assess the behavior of the algorithm under controlled conditions, *randomized instances* are meant to investigate scenarios where no a priori information on the contacts is known.

12.1.1 Deterministic instances

Specification. The input specification of a MWCI problem depends to three ingredients, namely the set of oligomers $\mathcal{O}_{\leq s}$, the value of α, and the individual weights $w(\cdot)$ for the candidate edges in $\mathrm{Pool}_{\mathrm{E}}(\mathcal{O}_{\leq s})$. We design MWCI instances to assess the relative importance of these ingredients. To this end, consider two values $F > G = 0.5 > U$, respectively meant to favor and penalize contacts. Note that the value $G = 0.5$ is a default value for contacts for which there is no a priori. The gap between these two values is defined by $\Delta = F - U$. Practically, we consider three cases, namely $(F, U) = (0.9, 0.1)$, $(F, U) = (0.75, 0.25)$, and $(F, U) = (0.6, 0.4)$,

The first set of instances involves the two weights F and U applied to the edges of the pool. The instance FU is obtained by assigning the weight F to all TP, and the weight U to all FP. To define a control, we define the UF instance by swapping the weights i.e. by favoring FP and penalizing TP. Note that instances of the type FF or UU, where true and false positives are given the same weight, are irrelevant since they are covered by the case $\alpha = 1$.

We first report basic facts observed for deterministic instances FU and UF defined by oligomers of size $s = 5, 8$, since the cases $s = 6, 7$ match $s = 5$ (supplemental Tables 17, 18, and 19).

Results. We examine successively the roles of α and of the individual weights.

Parameter α. When α increases, two striking facts are observed. First, the number of solutions increases, since one has up to 9 solutions when $\alpha = 0.25$, but up to 274 solutions when $\alpha = 1$ (supp. Table 17, $s = 8$). This solution set uses 22 contacts out of a pool of size 36. These 22 contacts involves 17 TP and 5 FP, resulting in a coverage of 0.45. The maximal number of solutions for $\alpha = 1$ owes to the fact that ties between contacts cannot be broken thanks to the weights, so that all solutions with the same number of contacts are equivalent. Second, the size of solutions decreases (up to 22 contacts for $\alpha = 0.25$ but nine only for $\alpha = 1$). This owes to the modest constant overhead in Eq. (2) for small values of α.

Weights. The configuration yielding the maximum number of solutions comes with an average (0.45) coverage (supp. Table 17, $s = 8$). Improving this score requires optimized combinations of α and weights, which is observed for the FU instance and $\alpha = 0.25$. In that case, the 20 TP are reported, while no FP is found, resulting in perfect unit values for the sensitivity, the specificity, and the coverage. This is admittedly a contrived experiments since TP are promoted while FP are hindered. Reverting odds, the control setup UF yields the expected, since penalizing TP and promoting FP results in a poor coverage (from one for FU to -0.90 for UF). It is also noticed that the difference in coverage decreases when α increases. For example, considering oligomers of size five, one gets $0.95 (= 0.85 - (-0.1))$, $0.5 (= 0.45 - (-0.05))$, and $0 (= 0.35 - 0.35)$ for $\alpha = 0.25, 0.5, 1$ respectively (supp. Table 17,

$s = 5$). This owes to the decreasing prevalence of weights when α increases. In a similar vein, larger values of Δ, or equivalently large values of the weight F favor high coverages (for the FU case, $\alpha = 0.25$ and $s = 8$, the coverage drops from one to 0.7 in moving from $\Delta = 0.8$ to $\Delta = 0.2$.)

All versus consensus solutions. Consensus solutions, which form a subset of all solutions, are characterized by two main properties. First, the number of consensus solutions varies in the range 1 to 48, that is, one get a 6 fold reduction with respect to the max number of total solutions. Second, the number of solutions is accompanied by a smaller set of edges used out of the pool of size 36, and also a smaller number (often null) of false positives. The former number decreases faster than the later, whence, overall, lower coverages.

12.1.2 Randomized instances

Specification. In designing deterministic instances involving the weights F and U, some a priori knowledge on the individual contacts is required to favor contacts standing a better chance to be true positives. If such information is not available, one could use favorable or unfavorable weights only. However, from the analysis carried out on deterministic instances, one gets that the FF scenario yields large solutions with false positives, while the UU scenario yields poor statistics — and in the extreme case connectivity inference problems without any solution. We therefore design a new class of instances also involving the intermediate weight G.

To specify these instances, we start from a deterministic instance, and use randomization. Consider e.g. the assignment of weights TP $\leftrightarrow F$ and FP $\leftrightarrow U$. For each contact from FP, we toss a fair coin and proceed as follows: if head is obtained, the contact keeps the weight F; if not, its weight is changed to G. We proceed likewise for false positive contacts, which may then be re-assigned a weight of G instead of the initial weight U. Note that for a given set of contacts (TP or FP), the expectation of the number of contacts whose weight is changed is half of the size of that set since the coin is fair. To avoid random bias, we generate 20 such instances.

Results. We noticed above that the FF and UU cases in the deterministic setting actually correspond to the case $\alpha = 1$. In comparing the results for randomized FF and UU instances against the case $\alpha = 1$, one first notices a drastic decrease of the number of solutions (2 for FF and $\alpha = 0.25$, 7 for UU and $\alpha = 0.25$, versus 274 for $\alpha = 1$) (Table 20). Solution size, however, are coherent with the deterministic case, and depend on the weights (large solutions for F weights, small solutions for U weights). Most interesting is the analysis of UU instances. On the one hand, a satisfactory sensitivity is obtained (for $\alpha = 0.25$: $\mathrm{ROC}_{sens.} = 0.55$ for randomized instances, versus $\mathrm{ROC}_{sens.} = 0.85$ for deterministic instances). On the other hand, an excellent specificity is observed (for $\alpha = 0.25$: $\mathrm{ROC}_{spec.} = 0.91$ for randomized instances, versus $\mathrm{ROC}_{spec.} = 0.69$ for deterministic instances).

12.1.3 Overall recommendations

We summarize the insights gained from the previous experiments on deterministic and randomized instances:

(i) Low values of α are sensitive to weights on the edges, as large solutions arise from favored edges.

(ii) Consensus solutions strongly hint at contacts which are true positives. However, modest coverage may stem from many false negatives.

(iii) High coverage scores are observed in two cases, namely when large solutions are obtained, or when a large number of solutions are obtained.

(iv) The scenario consisting of hindering a fraction of true contacts (by unfavorable weights or removing them from the pool) may trigger the discovery of alternative contacts also satisfying the connectivity constraints of oligomers. This finding, which stems from the analysis of randomized instances, underlies the strategy used in Algorithm `MILP-W`$_{\mathrm{B}}$ (section 5.1).

12.2 Results

The following tables present statistics to assess the incidence of weights, as explained in the main text. The following comments are in order:

- In the tables, the coverage values of Eq. (5) are color coded with a heat map, from blue (0-0.1) to red (0.9 - 1).
- The values reported in Tables 20, 21, 22 were obtained on 20 runs. The statistics reported correspond to the median of the values. For example, the number of solutions and the solution size are the median of the values obtained for all runs.

Table 17 Yeast exosome: statistics for U=0.1, F=0.9.

					$\alpha = 0.25$		$\alpha = 0.50$		$\alpha = 0.75$		$\alpha = 1$	
oligomer size, s	$\mid \mathrm{Pool_E}(\mathcal{O}_{\leq s}) \mid$	$\mid M \mid$	mode	*sols type*	($\mathrm{ROC}_{sens.}$, $\mathrm{ROC}_{spec.}$, Cvg)	Solutions (#, size)	($\mathrm{ROC}_{sens.}$, $\mathrm{ROC}_{spec.}$, Cvg)	Solutions (#, size)	($\mathrm{ROC}_{sens.}$, $\mathrm{ROC}_{spec.}$, Cvg)	Solutions (#, size)	($\mathrm{ROC}_{sens.}$, $\mathrm{ROC}_{spec.}$, Cvg)	Solutions (#, size)
5	20	3	FU	all	(1.00, 1.00, 0.85)	(1, 17)	(0.76, 1.00, 0.45)	(45, 8)	(0.76, 1.00, 0.45)	(45, 8)	(0.76, 0.33, 0.35)	(96, 8)
5	20	3	FU	cons	(1.00, 1.00, 0.85)	(1, 17)	(0.65, 1.00, 0.25)	(9, 8)	(0.65, 1.00, 0.25)	(9, 8)	(0.76, 0.33, 0.35)	(48, 8)
5	20	3	UF	all	(0.53, 0.00, -0.10)	(9, 9)	(0.53, 0.33, -0.05)	(9, 8)	(0.53, 0.33, -0.05)	(9, 8)	(0.76, 0.33, 0.35)	(96, 8)
5	20	3	UF	cons	(0.53, 0.00, -0.10)	(9, 9)	(0.53, 0.33, -0.05)	(9, 8)	(0.53, 0.33, -0.05)	(9, 8)	(0.76, 0.33, 0.35)	(48, 8)
6	21	3	FU	all	(1.00, 1.00, 0.85)	(1, 17)	(0.76, 1.00, 0.45)	(45, 8)	(0.76, 1.00, 0.45)	(45, 8)	(0.76, 0.50, 0.35)	(96, 8)
6	21	3	FU	cons	(1.00, 1.00, 0.85)	(1, 17)	(0.65, 1.00, 0.25)	(9, 8)	(0.65, 1.00, 0.25)	(9, 8)	(0.76, 0.50, 0.35)	(48, 8)
6	21	3	UF	all	(0.53, 0.00, -0.15)	(9, 10)	(0.53, 0.50, -0.05)	(9, 8)	(0.53, 0.50, -0.05)	(9, 8)	(0.76, 0.50, 0.35)	(96, 8)
6	21	3	UF	cons	(0.53, 0.00, -0.15)	(9, 10)	(0.53, 0.50, -0.05)	(9, 8)	(0.53, 0.50, -0.05)	(9, 8)	(0.76, 0.50, 0.35)	(48, 8)
7	29	3	FU	all	(1.00, 1.00, 0.85)	(1, 17)	(0.76, 1.00, 0.45)	(45, 8)	(0.76, 1.00, 0.45)	(45, 8)	(0.76, 0.83, 0.35)	(96, 8)
7	29	3	FU	cons	(1.00, 1.00, 0.85)	(1, 17)	(0.65, 1.00, 0.25)	(9, 8)	(0.65, 1.00, 0.25)	(9, 8)	(0.76, 0.83, 0.35)	(48, 8)
7	29	3	UF	all	(0.53, 0.00, -0.55)	(9, 18)	(0.53, 0.83, -0.05)	(9, 8)	(0.53, 0.83, -0.05)	(9, 8)	(0.76, 0.83, 0.35)	(96, 8)
7	29	3	UF	cons	(0.53, 0.00, -0.55)	(9, 18)	(0.53, 0.83, -0.05)	(9, 8)	(0.53, 0.83, -0.05)	(9, 8)	(0.76, 0.83, 0.35)	(48, 8)
8	36	0	FU	all	(1.00, 1.00, 1.00)	(1, 20)	(0.85, 1.00, 0.70)	(79, 9)	(0.85, 1.00, 0.70)	(79, 9)	(0.85, 0.69, 0.45)	(274, 9)
8	36	0	FU	cons	(1.00, 1.00, 1.00)	(1, 20)	(0.45, 1.00, -0.10)	(1, 9)	(0.45, 1.00, -0.10)	(1, 9)	(0.45, 0.94, -0.15)	(2, 9)
8	36	0	UF	all	(0.45, 0.00, -0.90)	(9, 22)	(0.45, 0.50, -0.50)	(63, 10)	(0.65, 0.69, 0.05)	(60, 9)	(0.85, 0.69, 0.45)	(274, 9)
8	36	0	UF	cons	(0.45, 0.00, -0.90)	(9, 22)	(0.45, 0.63, -0.40)	(18, 10)	(0.60, 0.75, 0.00)	(54, 9)	(0.45, 0.94, -0.15)	(2, 9)

Table 18 Yeast exosome: statistics for U=0.25, F=0.75.

					$\alpha = 0.25$		$\alpha = 0.50$		$\alpha = 0.75$		$\alpha = 1$	
oligomer size, s	$\mid \mathrm{Pool}_{\mathrm{E}}(\mathcal{O}_{\leq s}) \mid$	$\mid M \mid$	mode	*sols type*	($\mathrm{ROC}_{sens.}$, $\mathrm{ROC}_{spec.}$, Cvg)	Solutions (#, size)	($\mathrm{ROC}_{sens.}$, $\mathrm{ROC}_{spec.}$, Cvg)	Solutions (#, size)	($\mathrm{ROC}_{sens.}$, $\mathrm{ROC}_{spec.}$, Cvg)	Solutions (#, size)	($\mathrm{ROC}_{sens.}$, $\mathrm{ROC}_{spec.}$, Cvg)	Solutions (#, size)
5	20	3	FU	all	(0.76, 1.00, 0.45)	(45, 8)	(0.76, 1.00, 0.45)	(45, 8)	(0.76, 1.00, 0.45)	(45, 8)	(0.76, 0.33, 0.35)	(96, 8)
5	20	3	FU	cons	(0.65, 1.00, 0.25)	(9, 8)	(0.65, 1.00, 0.25)	(9, 8)	(0.65, 1.00, 0.25)	(9, 8)	(0.76, 0.33, 0.35)	(48, 8)
5	20	3	UF	all	(0.53, 0.33, -0.05)	(9, 8)	(0.53, 0.33, -0.05)	(9, 8)	(0.53, 0.33, -0.05)	(9, 8)	(0.76, 0.33, 0.35)	(96, 8)
5	20	3	UF	cons	(0.53, 0.33, -0.05)	(9, 8)	(0.53, 0.33, -0.05)	(9, 8)	(0.53, 0.33, -0.05)	(9, 8)	(0.76, 0.33, 0.35)	(48, 8)
6	21	3	FU	all	(0.76, 1.00, 0.45)	(45, 8)	(0.76, 1.00, 0.45)	(45, 8)	(0.76, 1.00, 0.45)	(45, 8)	(0.76, 0.50, 0.35)	(96, 8)
6	21	3	FU	cons	(0.65, 1.00, 0.25)	(9, 8)	(0.65, 1.00, 0.25)	(9, 8)	(0.65, 1.00, 0.25)	(9, 8)	(0.76, 0.50, 0.35)	(48, 8)
6	21	3	UF	all	(0.53, 0.50, -0.05)	(9, 8)	(0.53, 0.50, -0.05)	(9, 8)	(0.53, 0.50, -0.05)	(9, 8)	(0.76, 0.50, 0.35)	(96, 8)
6	21	3	UF	cons	(0.53, 0.50, -0.05)	(9, 8)	(0.53, 0.50, -0.05)	(9, 8)	(0.53, 0.50, -0.05)	(9, 8)	(0.76, 0.50, 0.35)	(48, 8)
7	29	3	FU	all	(0.76, 1.00, 0.45)	(45, 8)	(0.76, 1.00, 0.45)	(45, 8)	(0.76, 1.00, 0.45)	(45, 8)	(0.76, 0.83, 0.35)	(96, 8)
7	29	3	FU	cons	(0.65, 1.00, 0.25)	(9, 8)	(0.65, 1.00, 0.25)	(9, 8)	(0.65, 1.00, 0.25)	(9, 8)	(0.76, 0.83, 0.35)	(48, 8)
7	29	3	UF	all	(0.53, 0.83, -0.05)	(9, 8)	(0.53, 0.83, -0.05)	(9, 8)	(0.53, 0.83, -0.05)	(9, 8)	(0.76, 0.83, 0.35)	(96, 8)
7	29	3	UF	cons	(0.53, 0.83, -0.05)	(9, 8)	(0.53, 0.83, -0.05)	(9, 8)	(0.53, 0.83, -0.05)	(9, 8)	(0.76, 0.83, 0.35)	(48, 8)
8	36	0	FU	all	(0.85, 1.00, 0.70)	(79, 9)	(0.85, 1.00, 0.70)	(79, 9)	(0.85, 1.00, 0.70)	(79, 9)	(0.85, 0.69, 0.45)	(274, 9)
8	36	0	FU	cons	(0.45, 1.00, -0.10)	(1, 9)	(0.45, 1.00, -0.10)	(1, 9)	(0.45, 1.00, -0.10)	(1, 9)	(0.45, 0.94, -0.15)	(2, 9)
8	36	0	UF	all	(0.45, 0.50, -0.50)	(63, 10)	(0.65, 0.69, 0.05)	(60, 9)	(0.65, 0.69, 0.05)	(60, 9)	(0.85, 0.69, 0.45)	(274, 9)
8	36	0	UF	cons	(0.45, 0.63, -0.40)	(18, 10)	(0.60, 0.75, 0.00)	(54, 9)	(0.60, 0.75, 0.00)	(54, 9)	(0.45, 0.94, -0.15)	(2, 9)

Table 19 Yeast exosome: statistics for U=0.4, F=0.6.

					$\alpha = 0.25$		$\alpha = 0.50$		$\alpha = 0.75$		$\alpha = 1$	
oligomer size, s	$\mid \mathrm{Pool_E}(\mathcal{O}_{\leq s}) \mid$	$\mid M \mid$	mode	*sols type*	(ROC$_{sens.}$, ROC$_{spec.}$, Cvg)	Solutions (#, size)	(ROC$_{sens.}$, ROC$_{spec.}$, Cvg)	Solutions (#, size)	(ROC$_{sens.}$, ROC$_{spec.}$, Cvg)	Solutions (#, size)	(ROC$_{sens.}$, ROC$_{spec.}$, Cvg)	Solutions (#, size)
5	20	3	FU	all	(0.76, 1.00, 0.45)	(45, 8)	(0.76, 1.00, 0.45)	(45, 8)	(0.76, 1.00, 0.45)	(45, 8)	(0.76, 0.33, 0.35)	(96, 8)
5	20	3	FU	cons	(0.65, 1.00, 0.25)	(9, 8)	(0.65, 1.00, 0.25)	(9, 8)	(0.65, 1.00, 0.25)	(9, 8)	(0.76, 0.33, 0.35)	(48, 8)
5	20	3	UF	all	(0.53, 0.33, -0.05)	(9, 8)	(0.53, 0.33, -0.05)	(9, 8)	(0.53, 0.33, -0.05)	(9, 8)	(0.76, 0.33, 0.35)	(96, 8)
5	20	3	UF	cons	(0.53, 0.33, -0.05)	(9, 8)	(0.53, 0.33, -0.05)	(9, 8)	(0.53, 0.33, -0.05)	(9, 8)	(0.76, 0.33, 0.35)	(48, 8)
6	21	3	FU	all	(0.76, 1.00, 0.45)	(45, 8)	(0.76, 1.00, 0.45)	(45, 8)	(0.76, 1.00, 0.45)	(45, 8)	(0.76, 0.50, 0.35)	(96, 8)
6	21	3	FU	cons	(0.65, 1.00, 0.25)	(9, 8)	(0.65, 1.00, 0.25)	(9, 8)	(0.65, 1.00, 0.25)	(9, 8)	(0.76, 0.50, 0.35)	(48, 8)
6	21	3	UF	all	(0.53, 0.50, -0.05)	(9, 8)	(0.53, 0.50, -0.05)	(9, 8)	(0.53, 0.50, -0.05)	(9, 8)	(0.76, 0.50, 0.35)	(96, 8)
6	21	3	UF	cons	(0.53, 0.50, -0.05)	(9, 8)	(0.53, 0.50, -0.05)	(9, 8)	(0.53, 0.50, -0.05)	(9, 8)	(0.76, 0.50, 0.35)	(48, 8)
7	29	3	FU	all	(0.76, 1.00, 0.45)	(45, 8)	(0.76, 1.00, 0.45)	(45, 8)	(0.76, 1.00, 0.45)	(45, 8)	(0.76, 0.83, 0.35)	(96, 8)
7	29	3	FU	cons	(0.65, 1.00, 0.25)	(9, 8)	(0.65, 1.00, 0.25)	(9, 8)	(0.65, 1.00, 0.25)	(9, 8)	(0.76, 0.83, 0.35)	(48, 8)
7	29	3	UF	all	(0.53, 0.83, -0.05)	(9, 8)	(0.53, 0.83, -0.05)	(9, 8)	(0.53, 0.83, -0.05)	(9, 8)	(0.76, 0.83, 0.35)	(96, 8)
7	29	3	UF	cons	(0.53, 0.83, -0.05)	(9, 8)	(0.53, 0.83, -0.05)	(9, 8)	(0.53, 0.83, -0.05)	(9, 8)	(0.76, 0.83, 0.35)	(48, 8)
8	36	0	FU	all	(0.85, 1.00, 0.70)	(79, 9)	(0.85, 1.00, 0.70)	(79, 9)	(0.85, 1.00, 0.70)	(79, 9)	(0.85, 0.69, 0.45)	(274, 9)
8	36	0	FU	cons	(0.45, 1.00, -0.10)	(1, 9)	(0.45, 1.00, -0.10)	(1, 9)	(0.45, 1.00, -0.10)	(1, 9)	(0.45, 0.94, -0.15)	(2, 9)
8	36	0	UF	all	(0.65, 0.69, 0.05)	(60, 9)	(0.65, 0.69, 0.05)	(60, 9)	(0.65, 0.69, 0.05)	(60, 9)	(0.85, 0.69, 0.45)	(274, 9)
8	36	0	UF	cons	(0.60, 0.75, 0.00)	(54, 9)	(0.60, 0.75, 0.00)	(54, 9)	(0.60, 0.75, 0.00)	(54, 9)	(0.45, 0.94, -0.15)	(2, 9)

Table 20 Yeast exosome: statistics for U=0.1, F=0.9, G=0.5. 20 intances each.

					$\alpha = 0.25$		$\alpha = 0.50$		$\alpha = 0.75$		$\alpha = 1$	
oligomer size, s	$\lvert \text{Pool}_{\text{E}}(\mathcal{O}_{\leq s}) \rvert$	$\lvert M \rvert$	mode	*sols type*	(ROC$_{sens.}$, ROC$_{spec.}$, Cvg, σ_{Cvg})	Solutions (#, size)	(ROC$_{sens.}$, ROC$_{spec.}$, Cvg, σ_{Cvg})	Solutions (#, size)	(ROC$_{sens.}$, ROC$_{spec.}$, Cvg, σ_{Cvg})	Solutions (#, size)	(ROC$_{sens.}$, ROC$_{spec.}$, Cvg, σ_{Cvg})	Solutions (#, size)
5	20	3	FU	all	(0.76, 1.00, 0.45, 0.18)	(3, 11)	(0.65, 1.00, 0.20, 0.13)	(8, 8)	(0.65, 1.00, 0.20, 0.13)	(8, 8)	(0.76, 0.33, 0.35, 0.00)	(96, 8)
5	20	3	FU	cons	(0.76, 1.00, 0.45, 0.18)	(3, 11)	(0.65, 1.00, 0.20, 0.13)	(8, 8)	(0.65, 1.00, 0.20, 0.13)	(8, 8)	(0.76, 0.33, 0.35, 0.00)	(48, 8)
5	20	3	FF	all	(0.71, 0.33, 0.20, 0.15)	(2, 13)	(0.59, 0.67, 0.10, 0.14)	(12, 8)	(0.59, 0.67, 0.10, 0.14)	(12, 8)	(0.76, 0.33, 0.35, 0.00)	(96, 8)
5	20	3	FF	cons	(0.71, 0.33, 0.20, 0.15)	(2, 13)	(0.59, 0.67, 0.10, 0.14)	(12, 8)	(0.59, 0.67, 0.10, 0.14)	(12, 8)	(0.76, 0.33, 0.35, 0.00)	(48, 8)
5	20	3	UF	all	(0.53, 0.33, -0.10, 0.15)	(6, 9)	(0.53, 0.33, -0.05, 0.14)	(6, 8)	(0.53, 0.33, -0.05, 0.14)	(6, 8)	(0.76, 0.33, 0.35, 0.00)	(96, 8)
5	20	3	UF	cons	(0.53, 0.33, -0.10, 0.15)	(6, 9)	(0.53, 0.33, -0.05, 0.14)	(6, 8)	(0.53, 0.33, -0.05, 0.14)	(6, 8)	(0.76, 0.33, 0.35, 0.00)	(48, 8)
5	20	3	UU	all	(0.59, 0.67, 0.10, 0.13)	(8, 8)	(0.59, 0.67, 0.10, 0.13)	(8, 8)	(0.59, 0.67, 0.10, 0.13)	(8, 8)	(0.76, 0.33, 0.35, 0.00)	(96, 8)
5	20	3	UU	cons	(0.59, 0.67, 0.10, 0.13)	(8, 8)	(0.59, 0.67, 0.10, 0.13)	(8, 8)	(0.59, 0.67, 0.10, 0.13)	(8, 8)	(0.76, 0.33, 0.35, 0.00)	(48, 8)
6	21	3	FU	all	(0.71, 1.00, 0.35, 0.19)	(2, 11)	(0.59, 1.00, 0.15, 0.13)	(4, 8)	(0.59, 1.00, 0.15, 0.13)	(4, 8)	(0.76, 0.33, 0.35, 0.00)	(96, 8)
6	21	3	FU	cons	(0.71, 1.00, 0.35, 0.19)	(2, 11)	(0.59, 1.00, 0.15, 0.13)	(4, 8)	(0.59, 1.00, 0.15, 0.13)	(4, 8)	(0.76, 0.33, 0.35, 0.00)	(48, 8)
6	21	3	FF	all	(0.71, 0.25, 0.20, 0.20)	(2, 13)	(0.59, 0.50, 0.05, 0.16)	(8, 8)	(0.59, 0.50, 0.05, 0.16)	(8, 8)	(0.76, 0.33, 0.35, 0.00)	(96, 8)
6	21	3	FF	cons	(0.71, 0.25, 0.20, 0.20)	(2, 13)	(0.59, 0.50, 0.05, 0.16)	(8, 8)	(0.59, 0.50, 0.05, 0.16)	(8, 8)	(0.76, 0.33, 0.35, 0.00)	(48, 8)
6	21	3	UF	all	(0.47, 0.25, -0.15, 0.17)	(6, 9)	(0.47, 0.50, -0.13, 0.15)	(6, 8)	(0.47, 0.50, -0.13, 0.15)	(6, 8)	(0.76, 0.33, 0.35, 0.00)	(96, 8)
6	21	3	UF	cons	(0.47, 0.25, -0.15, 0.17)	(6, 9)	(0.47, 0.50, -0.13, 0.15)	(6, 8)	(0.47, 0.50, -0.13, 0.15)	(6, 8)	(0.76, 0.33, 0.35, 0.00)	(48, 8)
6	21	3	UU	all	(0.59, 0.50, 0.13, 0.16)	(9, 8)	(0.59, 0.50, 0.13, 0.16)	(9, 8)	(0.59, 0.50, 0.13, 0.16)	(9, 8)	(0.76, 0.33, 0.35, 0.00)	(96, 8)
6	21	3	UU	cons	(0.59, 0.50, 0.13, 0.16)	(9, 8)	(0.59, 0.50, 0.13, 0.16)	(9, 8)	(0.59, 0.50, 0.13, 0.16)	(9, 8)	(0.76, 0.33, 0.35, 0.00)	(48, 8)
7	29	3	FU	all	(0.76, 1.00, 0.45, 0.15)	(2, 12)	(0.65, 1.00, 0.25, 0.11)	(6, 8)	(0.65, 1.00, 0.25, 0.11)	(6, 8)	(0.76, 0.33, 0.35, 0.00)	(96, 8)
7	29	3	FU	cons	(0.76, 1.00, 0.45, 0.15)	(2, 12)	(0.65, 1.00, 0.25, 0.11)	(6, 8)	(0.65, 1.00, 0.25, 0.11)	(6, 8)	(0.76, 0.33, 0.35, 0.00)	(48, 8)
7	29	3	FF	all	(0.73, 0.50, 0.08, 0.19)	(2, 17)	(0.59, 0.92, 0.15, 0.14)	(8, 8)	(0.59, 0.92, 0.15, 0.14)	(8, 8)	(0.76, 0.33, 0.35, 0.00)	(96, 8)
7	29	3	FF	cons	(0.73, 0.50, 0.08, 0.19)	(2, 17)	(0.59, 0.92, 0.15, 0.14)	(8, 8)	(0.59, 0.92, 0.15, 0.14)	(8, 8)	(0.76, 0.33, 0.35, 0.00)	(48, 8)
7	29	3	UF	all	(0.47, 0.42, -0.38, 0.16)	(6, 13)	(0.47, 0.83, -0.15, 0.14)	(6, 8)	(0.47, 0.83, -0.15, 0.14)	(6, 8)	(0.76, 0.33, 0.35, 0.00)	(96, 8)
7	29	3	UF	cons	(0.47, 0.42, -0.38, 0.16)	(6, 13)	(0.47, 0.83, -0.15, 0.14)	(6, 8)	(0.47, 0.83, -0.15, 0.14)	(6, 8)	(0.76, 0.33, 0.35, 0.00)	(48, 8)
7	29	3	UU	all	(0.62, 0.92, 0.15, 0.11)	(12, 8)	(0.62, 0.92, 0.15, 0.11)	(12, 8)	(0.62, 0.92, 0.15, 0.11)	(12, 8)	(0.76, 0.33, 0.35, 0.00)	(96, 8)
7	29	3	UU	cons	(0.62, 0.92, 0.15, 0.11)	(12, 8)	(0.62, 0.92, 0.15, 0.11)	(12, 8)	(0.62, 0.92, 0.15, 0.11)	(12, 8)	(0.76, 0.33, 0.35, 0.00)	(48, 8)
8	36	0	FU	all	(0.70, 1.00, 0.35, 0.18)	(4, 12)	(0.60, 1.00, 0.17, 0.19)	(6, 9)	(0.60, 1.00, 0.17, 0.19)	(6, 9)	(0.85, 0.69, 0.45, 0.00)	(274, 9)
8	36	0	FU	cons	(0.70, 1.00, 0.35, 0.18)	(4, 12)	(0.60, 1.00, 0.17, 0.19)	(6, 9)	(0.60, 1.00, 0.17, 0.19)	(6, 9)	(0.45, 0.94, -0.15, 0.00)	(2, 9)
8	36	0	FF	all	(0.70, 0.44, -0.05, 0.20)	(2, 21)	(0.55, 0.88, 0.00, 0.18)	(16, 9)	(0.55, 0.88, 0.00, 0.18)	(16, 9)	(0.85, 0.69, 0.45, 0.00)	(274, 9)
8	36	0	FF	cons	(0.70, 0.44, -0.05, 0.20)	(2, 21)	(0.55, 0.88, 0.00, 0.18)	(16, 9)	(0.55, 0.88, 0.00, 0.18)	(16, 9)	(0.45, 0.94, -0.15, 0.00)	(2, 9)
8	36	0	UF	all	(0.45, 0.38, -0.63, 0.16)	(6, 16)	(0.45, 0.81, -0.23, 0.15)	(7, 9)	(0.50, 0.81, -0.17, 0.13)	(8, 9)	(0.85, 0.69, 0.45, 0.00)	(274, 9)
8	36	0	UF	cons	(0.45, 0.38, -0.63, 0.16)	(6, 16)	(0.45, 0.81, -0.23, 0.15)	(7, 9)	(0.50, 0.81, -0.17, 0.13)	(8, 9)	(0.45, 0.94, -0.15, 0.00)	(2, 9)
8	36	0	UU	all	(0.55, 0.91, 0.05, 0.16)	(7, 9)	(0.55, 0.91, 0.05, 0.17)	(8, 9)	(0.55, 0.91, 0.05, 0.17)	(8, 9)	(0.85, 0.69, 0.45, 0.00)	(274, 9)
8	36	0	UU	cons	(0.55, 0.91, 0.05, 0.16)	(7, 9)	(0.55, 0.91, 0.05, 0.17)	(8, 9)	(0.55, 0.91, 0.05, 0.17)	(8, 9)	(0.45, 0.94, -0.15, 0.00)	(2, 9)

Table 21 Yeast exosome: statistics for U=0.25, F=0.75, G=0.5. 20 instances each.

					$\alpha = 0.25$		$\alpha = 0.50$		$\alpha = 0.75$		$\alpha = 1$	
oligomer size, s	$\lvert \text{Pool}_{\text{E}}(\mathcal{O}_{\leq s}) \rvert$	$\lvert M \rvert$	mode	*sols type*	(ROC$_{sens.}$, ROC$_{spec.}$, Cvg, σ_{Cvg})	Solutions (#, size)	(ROC$_{sens.}$, ROC$_{spec.}$, Cvg, σ_{Cvg})	Solutions (#, size)	(ROC$_{sens.}$, ROC$_{spec.}$, Cvg, σ_{Cvg})	Solutions (#, size)	(ROC$_{sens.}$, ROC$_{spec.}$, Cvg, σ_{Cvg})	Solutions (#, size)
5	20	3	FU	all	(0.59, 1.00, 0.15, 0.12)	(3, 8)	(0.59, 1.00, 0.15, 0.12)	(3, 8)	(0.59, 1.00, 0.15, 0.12)	(3, 8)	(0.76, 0.33, 0.35, 0.00)	(96, 8)
5	20	3	FU	cons	(0.59, 1.00, 0.15, 0.12)	(3, 8)	(0.59, 1.00, 0.15, 0.12)	(3, 8)	(0.59, 1.00, 0.15, 0.12)	(3, 8)	(0.76, 0.33, 0.35, 0.00)	(48, 8)
5	20	3	FF	all	(0.59, 0.67, 0.10, 0.15)	(6, 8)	(0.59, 0.67, 0.10, 0.15)	(6, 8)	(0.59, 0.67, 0.10, 0.15)	(6, 8)	(0.76, 0.33, 0.35, 0.00)	(96, 8)
5	20	3	FF	cons	(0.59, 0.67, 0.10, 0.15)	(6, 8)	(0.59, 0.67, 0.10, 0.15)	(6, 8)	(0.59, 0.67, 0.10, 0.15)	(6, 8)	(0.76, 0.33, 0.35, 0.00)	(48, 8)
5	20	3	UF	all	(0.47, 0.33, -0.15, 0.12)	(4, 8)	(0.47, 0.33, -0.15, 0.12)	(4, 8)	(0.47, 0.33, -0.15, 0.12)	(4, 8)	(0.76, 0.33, 0.35, 0.00)	(96, 8)
5	20	3	UF	cons	(0.47, 0.33, -0.15, 0.12)	(4, 8)	(0.47, 0.33, -0.15, 0.12)	(4, 8)	(0.47, 0.33, -0.15, 0.12)	(4, 8)	(0.76, 0.33, 0.35, 0.00)	(48, 8)
5	20	3	UU	all	(0.56, 0.67, 0.05, 0.14)	(6, 8)	(0.56, 0.67, 0.05, 0.14)	(6, 8)	(0.56, 0.67, 0.05, 0.14)	(6, 8)	(0.76, 0.33, 0.35, 0.00)	(96, 8)
5	20	3	UU	cons	(0.56, 0.67, 0.05, 0.14)	(6, 8)	(0.56, 0.67, 0.05, 0.14)	(6, 8)	(0.56, 0.67, 0.05, 0.14)	(6, 8)	(0.76, 0.33, 0.35, 0.00)	(48, 8)
6	21	3	FU	all	(0.62, 1.00, 0.20, 0.13)	(6, 8)	(0.62, 1.00, 0.20, 0.13)	(6, 8)	(0.62, 1.00, 0.20, 0.13)	(6, 8)	(0.76, 0.33, 0.35, 0.00)	(96, 8)
6	21	3	FU	cons	(0.62, 1.00, 0.20, 0.13)	(6, 8)	(0.62, 1.00, 0.20, 0.13)	(6, 8)	(0.62, 1.00, 0.20, 0.13)	(6, 8)	(0.76, 0.33, 0.35, 0.00)	(48, 8)
6	21	3	FF	all	(0.59, 0.75, 0.05, 0.12)	(4, 8)	(0.59, 0.75, 0.05, 0.12)	(4, 8)	(0.59, 0.75, 0.05, 0.12)	(4, 8)	(0.76, 0.33, 0.35, 0.00)	(96, 8)
6	21	3	FF	cons	(0.59, 0.75, 0.05, 0.12)	(4, 8)	(0.59, 0.75, 0.05, 0.12)	(4, 8)	(0.59, 0.75, 0.05, 0.12)	(4, 8)	(0.76, 0.33, 0.35, 0.00)	(48, 8)
6	21	3	UF	all	(0.50, 0.50, -0.10, 0.16)	(7, 8)	(0.50, 0.50, -0.10, 0.16)	(7, 8)	(0.50, 0.50, -0.10, 0.16)	(7, 8)	(0.76, 0.33, 0.35, 0.00)	(96, 8)
6	21	3	UF	cons	(0.50, 0.50, -0.10, 0.16)	(7, 8)	(0.50, 0.50, -0.10, 0.16)	(7, 8)	(0.50, 0.50, -0.10, 0.16)	(7, 8)	(0.76, 0.33, 0.35, 0.00)	(48, 8)
6	21	3	UU	all	(0.59, 0.75, 0.10, 0.14)	(9, 8)	(0.59, 0.75, 0.10, 0.14)	(9, 8)	(0.59, 0.75, 0.10, 0.14)	(9, 8)	(0.76, 0.33, 0.35, 0.00)	(96, 8)
6	21	3	UU	cons	(0.59, 0.75, 0.10, 0.14)	(9, 8)	(0.59, 0.75, 0.10, 0.14)	(9, 8)	(0.59, 0.75, 0.10, 0.14)	(9, 8)	(0.76, 0.33, 0.35, 0.00)	(48, 8)
7	29	3	FU	all	(0.59, 1.00, 0.15, 0.11)	(5, 8)	(0.59, 1.00, 0.15, 0.11)	(5, 8)	(0.59, 1.00, 0.15, 0.11)	(5, 8)	(0.76, 0.33, 0.35, 0.00)	(96, 8)
7	29	3	FU	cons	(0.59, 1.00, 0.15, 0.11)	(5, 8)	(0.59, 1.00, 0.15, 0.11)	(5, 8)	(0.59, 1.00, 0.15, 0.11)	(5, 8)	(0.76, 0.33, 0.35, 0.00)	(48, 8)
7	29	3	FF	all	(0.59, 0.92, 0.13, 0.11)	(8, 8)	(0.59, 0.92, 0.13, 0.11)	(8, 8)	(0.59, 0.92, 0.13, 0.11)	(8, 8)	(0.76, 0.33, 0.35, 0.00)	(96, 8)
7	29	3	FF	cons	(0.59, 0.92, 0.13, 0.11)	(8, 8)	(0.59, 0.92, 0.13, 0.11)	(8, 8)	(0.59, 0.92, 0.13, 0.11)	(8, 8)	(0.76, 0.33, 0.35, 0.00)	(48, 8)
7	29	3	UF	all	(0.53, 0.83, -0.05, 0.16)	(6, 8)	(0.53, 0.83, -0.05, 0.16)	(6, 8)	(0.53, 0.83, -0.05, 0.16)	(6, 8)	(0.76, 0.33, 0.35, 0.00)	(96, 8)
7	29	3	UF	cons	(0.53, 0.83, -0.05, 0.16)	(6, 8)	(0.53, 0.83, -0.05, 0.16)	(6, 8)	(0.53, 0.83, -0.05, 0.16)	(6, 8)	(0.76, 0.33, 0.35, 0.00)	(48, 8)
7	29	3	UU	all	(0.59, 0.92, 0.10, 0.13)	(12, 8)	(0.59, 0.92, 0.10, 0.13)	(12, 8)	(0.59, 0.92, 0.10, 0.13)	(12, 8)	(0.76, 0.33, 0.35, 0.00)	(96, 8)
7	29	3	UU	cons	(0.59, 0.92, 0.10, 0.13)	(12, 8)	(0.59, 0.92, 0.10, 0.13)	(12, 8)	(0.59, 0.92, 0.10, 0.13)	(12, 8)	(0.76, 0.33, 0.35, 0.00)	(48, 8)
8	36	0	FU	all	(0.55, 1.00, 0.10, 0.18)	(4, 9)	(0.55, 1.00, 0.10, 0.18)	(4, 9)	(0.55, 1.00, 0.10, 0.18)	(4, 9)	(0.85, 0.69, 0.45, 0.00)	(274, 9)
8	36	0	FU	cons	(0.55, 1.00, 0.10, 0.18)	(4, 9)	(0.55, 1.00, 0.10, 0.18)	(4, 9)	(0.55, 1.00, 0.10, 0.18)	(4, 9)	(0.45, 0.94, -0.15, 0.00)	(2, 9)
8	36	0	FF	all	(0.55, 0.88, -0.05, 0.19)	(7, 9)	(0.55, 0.88, 0.03, 0.19)	(8, 9)	(0.55, 0.88, 0.03, 0.19)	(8, 9)	(0.85, 0.69, 0.45, 0.00)	(274, 9)
8	36	0	FF	cons	(0.55, 0.88, -0.05, 0.19)	(7, 9)	(0.55, 0.88, 0.03, 0.19)	(8, 9)	(0.55, 0.88, 0.03, 0.19)	(8, 9)	(0.45, 0.94, -0.15, 0.00)	(2, 9)
8	36	0	UF	all	(0.35, 0.75, -0.50, 0.18)	(4, 10)	(0.45, 0.81, -0.20, 0.17)	(4, 9)	(0.45, 0.81, -0.20, 0.17)	(4, 9)	(0.85, 0.69, 0.45, 0.00)	(274, 9)
8	36	0	UF	cons	(0.35, 0.75, -0.50, 0.18)	(4, 10)	(0.45, 0.81, -0.20, 0.17)	(4, 9)	(0.45, 0.81, -0.20, 0.17)	(4, 9)	(0.45, 0.94, -0.15, 0.00)	(2, 9)
8	36	0	UU	all	(0.55, 0.88, 0.08, 0.17)	(10, 9)	(0.55, 0.88, 0.08, 0.17)	(10, 9)	(0.55, 0.88, 0.08, 0.17)	(10, 9)	(0.85, 0.69, 0.45, 0.00)	(274, 9)
8	36	0	UU	cons	(0.55, 0.88, 0.08, 0.17)	(10, 9)	(0.55, 0.88, 0.08, 0.17)	(10, 9)	(0.55, 0.88, 0.08, 0.17)	(10, 9)	(0.45, 0.94, -0.15, 0.00)	(2, 9)

Table 22 Yeast exosome: statistics for U=0.4, F=0.6, G=0.5. 20 instances each.

					$\alpha = 0.25$		$\alpha = 0.50$		$\alpha = 0.75$		$\alpha = 1$	
oligomer size, s	$\lvert \text{Pool}_{\text{E}}(\mathcal{O}_{\leq s}) \rvert$	$\lvert M \rvert$	mode	*sols type*	(ROC$_{sens.}$, ROC$_{spec.}$, Cvg, σ_{Cvg})	Solutions (#, size)	(ROC$_{sens.}$, ROC$_{spec.}$, Cvg, σ_{Cvg})	Solutions (#, size)	(ROC$_{sens.}$, ROC$_{spec.}$, Cvg, σ_{Cvg})	Solutions (#, size)	(ROC$_{sens.}$, ROC$_{spec.}$, Cvg, σ_{Cvg})	Solutions (#, size)
5	20	3	FU	all	(0.59, 1.00, 0.15, 0.12)	(8, 8)	(0.59, 1.00, 0.15, 0.12)	(8, 8)	(0.59, 1.00, 0.15, 0.12)	(8, 8)	(0.76, 0.33, 0.35, 0.00)	(96, 8)
5	20	3	FU	cons	(0.59, 1.00, 0.15, 0.12)	(8, 8)	(0.59, 1.00, 0.15, 0.12)	(8, 8)	(0.59, 1.00, 0.15, 0.12)	(8, 8)	(0.76, 0.33, 0.35, 0.00)	(48, 8)
5	20	3	FF	all	(0.59, 0.67, 0.10, 0.13)	(9, 8)	(0.59, 0.67, 0.10, 0.13)	(9, 8)	(0.59, 0.67, 0.10, 0.13)	(9, 8)	(0.76, 0.33, 0.35, 0.00)	(96, 8)
5	20	3	FF	cons	(0.59, 0.67, 0.10, 0.13)	(9, 8)	(0.59, 0.67, 0.10, 0.13)	(9, 8)	(0.59, 0.67, 0.10, 0.13)	(9, 8)	(0.76, 0.33, 0.35, 0.00)	(48, 8)
5	20	3	UF	all	(0.47, 0.33, -0.15, 0.14)	(4, 8)	(0.47, 0.33, -0.15, 0.14)	(4, 8)	(0.47, 0.33, -0.15, 0.14)	(4, 8)	(0.76, 0.33, 0.35, 0.00)	(96, 8)
5	20	3	UF	cons	(0.47, 0.33, -0.15, 0.14)	(4, 8)	(0.47, 0.33, -0.15, 0.14)	(4, 8)	(0.47, 0.33, -0.15, 0.14)	(4, 8)	(0.76, 0.33, 0.35, 0.00)	(48, 8)
5	20	3	UU	all	(0.59, 0.67, 0.05, 0.12)	(6, 8)	(0.59, 0.67, 0.05, 0.12)	(6, 8)	(0.59, 0.67, 0.05, 0.12)	(6, 8)	(0.76, 0.33, 0.35, 0.00)	(96, 8)
5	20	3	UU	cons	(0.59, 0.67, 0.05, 0.12)	(6, 8)	(0.59, 0.67, 0.05, 0.12)	(6, 8)	(0.59, 0.67, 0.05, 0.12)	(6, 8)	(0.76, 0.33, 0.35, 0.00)	(48, 8)
6	21	3	FU	all	(0.59, 1.00, 0.15, 0.13)	(6, 8)	(0.59, 1.00, 0.15, 0.13)	(6, 8)	(0.59, 1.00, 0.15, 0.13)	(6, 8)	(0.76, 0.33, 0.35, 0.00)	(96, 8)
6	21	3	FU	cons	(0.59, 1.00, 0.15, 0.13)	(6, 8)	(0.59, 1.00, 0.15, 0.13)	(6, 8)	(0.59, 1.00, 0.15, 0.13)	(6, 8)	(0.76, 0.33, 0.35, 0.00)	(48, 8)
6	21	3	FF	all	(0.59, 0.75, 0.10, 0.15)	(11, 8)	(0.59, 0.75, 0.10, 0.15)	(11, 8)	(0.59, 0.75, 0.10, 0.15)	(11, 8)	(0.76, 0.33, 0.35, 0.00)	(96, 8)
6	21	3	FF	cons	(0.59, 0.75, 0.10, 0.15)	(11, 8)	(0.59, 0.75, 0.10, 0.15)	(11, 8)	(0.59, 0.75, 0.10, 0.15)	(11, 8)	(0.76, 0.33, 0.35, 0.00)	(48, 8)
6	21	3	UF	all	(0.47, 0.50, -0.15, 0.13)	(4, 8)	(0.47, 0.50, -0.15, 0.13)	(4, 8)	(0.47, 0.50, -0.15, 0.13)	(4, 8)	(0.76, 0.33, 0.35, 0.00)	(96, 8)
6	21	3	UF	cons	(0.47, 0.50, -0.15, 0.13)	(4, 8)	(0.47, 0.50, -0.15, 0.13)	(4, 8)	(0.47, 0.50, -0.15, 0.13)	(4, 8)	(0.76, 0.33, 0.35, 0.00)	(48, 8)
6	21	3	UU	all	(0.59, 0.75, 0.13, 0.14)	(9, 8)	(0.59, 0.75, 0.13, 0.14)	(9, 8)	(0.59, 0.75, 0.13, 0.14)	(9, 8)	(0.76, 0.33, 0.35, 0.00)	(96, 8)
6	21	3	UU	cons	(0.59, 0.75, 0.13, 0.14)	(9, 8)	(0.59, 0.75, 0.13, 0.14)	(9, 8)	(0.59, 0.75, 0.13, 0.14)	(9, 8)	(0.76, 0.33, 0.35, 0.00)	(48, 8)
7	29	3	FU	all	(0.59, 1.00, 0.15, 0.12)	(6, 8)	(0.59, 1.00, 0.15, 0.12)	(6, 8)	(0.59, 1.00, 0.15, 0.12)	(6, 8)	(0.76, 0.33, 0.35, 0.00)	(96, 8)
7	29	3	FU	cons	(0.59, 1.00, 0.15, 0.12)	(6, 8)	(0.59, 1.00, 0.15, 0.12)	(6, 8)	(0.59, 1.00, 0.15, 0.12)	(6, 8)	(0.76, 0.33, 0.35, 0.00)	(48, 8)
7	29	3	FF	all	(0.59, 0.92, 0.10, 0.12)	(7, 8)	(0.59, 0.92, 0.10, 0.12)	(7, 8)	(0.59, 0.92, 0.10, 0.12)	(7, 8)	(0.76, 0.33, 0.35, 0.00)	(96, 8)
7	29	3	FF	cons	(0.59, 0.92, 0.10, 0.12)	(7, 8)	(0.59, 0.92, 0.10, 0.12)	(7, 8)	(0.59, 0.92, 0.10, 0.12)	(7, 8)	(0.76, 0.33, 0.35, 0.00)	(48, 8)
7	29	3	UF	all	(0.53, 0.83, -0.05, 0.15)	(6, 8)	(0.53, 0.83, -0.05, 0.15)	(6, 8)	(0.53, 0.83, -0.05, 0.15)	(6, 8)	(0.76, 0.33, 0.35, 0.00)	(96, 8)
7	29	3	UF	cons	(0.53, 0.83, -0.05, 0.15)	(6, 8)	(0.53, 0.83, -0.05, 0.15)	(6, 8)	(0.53, 0.83, -0.05, 0.15)	(6, 8)	(0.76, 0.33, 0.35, 0.00)	(48, 8)
7	29	3	UU	all	(0.53, 0.92, 0.05, 0.14)	(7, 8)	(0.53, 0.92, 0.05, 0.14)	(7, 8)	(0.53, 0.92, 0.05, 0.14)	(7, 8)	(0.76, 0.33, 0.35, 0.00)	(96, 8)
7	29	3	UU	cons	(0.53, 0.92, 0.05, 0.14)	(7, 8)	(0.53, 0.92, 0.05, 0.14)	(7, 8)	(0.53, 0.92, 0.05, 0.14)	(7, 8)	(0.76, 0.33, 0.35, 0.00)	(48, 8)
8	36	0	FU	all	(0.55, 1.00, 0.10, 0.19)	(4, 9)	(0.55, 1.00, 0.10, 0.19)	(4, 9)	(0.55, 1.00, 0.10, 0.19)	(4, 9)	(0.85, 0.69, 0.45, 0.00)	(274, 9)
8	36	0	FU	cons	(0.55, 1.00, 0.10, 0.19)	(4, 9)	(0.55, 1.00, 0.10, 0.19)	(4, 9)	(0.55, 1.00, 0.10, 0.19)	(4, 9)	(0.45, 0.94, -0.15, 0.00)	(2, 9)
8	36	0	FF	all	(0.55, 0.88, -0.05, 0.17)	(8, 9)	(0.55, 0.88, -0.05, 0.17)	(8, 9)	(0.55, 0.88, -0.05, 0.17)	(8, 9)	(0.85, 0.69, 0.45, 0.00)	(274, 9)
8	36	0	FF	cons	(0.55, 0.88, -0.05, 0.17)	(8, 9)	(0.55, 0.88, -0.05, 0.17)	(8, 9)	(0.55, 0.88, -0.05, 0.17)	(8, 9)	(0.45, 0.94, -0.15, 0.00)	(2, 9)
8	36	0	UF	all	(0.50, 0.81, -0.15, 0.14)	(9, 9)	(0.50, 0.81, -0.15, 0.14)	(9, 9)	(0.50, 0.81, -0.15, 0.14)	(9, 9)	(0.85, 0.69, 0.45, 0.00)	(274, 9)
8	36	0	UF	cons	(0.50, 0.81, -0.15, 0.14)	(9, 9)	(0.50, 0.81, -0.15, 0.14)	(9, 9)	(0.50, 0.81, -0.15, 0.14)	(9, 9)	(0.45, 0.94, -0.15, 0.00)	(2, 9)
8	36	0	UU	all	(0.60, 0.88, 0.08, 0.13)	(12, 9)	(0.60, 0.88, 0.08, 0.13)	(12, 9)	(0.60, 0.88, 0.08, 0.13)	(12, 9)	(0.85, 0.69, 0.45, 0.00)	(274, 9)
8	36	0	UU	cons	(0.60, 0.88, 0.08, 0.13)	(12, 9)	(0.60, 0.88, 0.08, 0.13)	(12, 9)	(0.60, 0.88, 0.08, 0.13)	(12, 9)	(0.45, 0.94, -0.15, 0.00)	(2, 9)

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