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

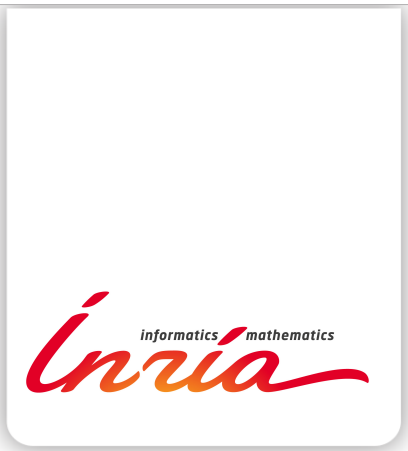
Deepesh Agarwal, Christelle Caillouet, David Coudert, Frédéric Cazals


# 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 — October 2014 — 23 pages

**Abstract:** Consider a set of oligomers listing the subunits involved in sub-complexes of a macro-molecular 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

* 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

# 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 $\texttt{MILP-W}_{\text{B}}$. 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

**Structural inference from oligomers.** 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 snapshot of large systems remains *tour de force*, so that alternative methods are being developed. Given a large assembly, of particular interest are methods producing oligomers of varying size of the assembly, such as tandem affinity purification [ea01] or native mass spectrometry [SR07]. Oligomers of varying size can indeed 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], a problem which we formalize now.

**Unweighted and weighted connectivity inference: MCI and MWCI.** Consider a macromolecular assembly consisting of subunits (typically proteins or nucleic acids). Assume that these subunits are known, but that the pairwise contacts between them are not. 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. Note that mathematically, the subunits may be seen as the nodes of a graph whose edges are defined by the contacts.

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 instances 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 contacts $S$. 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 to the vertices of an oligomer formula yields a connected graph.

If no a priori on contacts exists, the *size* of a solution is its number of edges. Mastering this size is non trivial, since the number of interfaces between the assembly is unknown, and the trivial solution involving all edges is admissible. To avoid speculating on this number, the *Minimum Connectivity Inference* problem (MCI) is the variant of CI seeking valid edge sets of minimum cardinality.

In a variety of setting, knowledge on contacts exists. 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, etc. [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 involve 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 (rather than transient) interactions. For these reasons, strategies computing confidence scores usually resort to machine learning tools trained on the

aforementioned data [YMJ12] and also [TRT+10].

**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 admit a solution using a predefined number of edges $k$ is **NP**-complete, while finding the solution of smallest size is APX-hard (unless $\mathbf{P} \neq \mathbf{NP}$, there does not exist any polynomial time approximation scheme) [AAC+13]. It should be stressed that these facts do not exploit any peculiar property of real data, and only show the existence of hard instances.

Two algorithms targeting CI problems have been developed so far. The first one is the two-stage heuristic method reported in [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 the diversity of the connectivity. 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, in particular `MILP` solving MCI problems. 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 opposite, in consensus solutions, few false positives are observed, at the detriment of more false negatives. 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, and use it to report contacts with high sensitivity and specificity.

# 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 $\mathrm{Pool}_{\mathrm{E}}(\mathcal{O})$ 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 $\mathrm{Pool}_{\mathrm{E}}(\mathcal{O}_{\leq s})$. The rationale for doing so is that smaller oligomers favor local contacts, the extreme case being that of dimers – 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]. In fact, connecting an oligomer merely requires a tree, whose number of edges is the number of vertices minus one, so that solving a MCI problem consists of efficiently *combining* the trees associated to all oligomers.

**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 for a solution $S$, we define:

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

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 $\alpha$. In particular, $\alpha < 1$ and different weights for all edges favor 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.

**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 $\alpha$, the edge weight prevails or not depending on its value with respect to the value $1 - d$. For $d = 1/2$, one gets $dC_{\alpha,d}(e)/d\alpha = 1/2 + w(e)$. From this observation, one also gets that in increasing the weight $\alpha$ to $\alpha' = \alpha + \varepsilon$, one has*

$$C_{\alpha'}(e) = C_\alpha(e) + \varepsilon(\frac{1}{2} + w(e)). \tag{4}$$

# 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 $\alpha$ 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 $\mathcal{S}_{\text{MILP-W}}$. 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 *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 $\mathcal{S}^{cons.}_{\text{MILP-W}}$ The union of edges seen is the solutions of $\mathcal{S}_{\text{MILP-W}}$ is denoted $\mathcal{E}_{\text{MILP-W}}$, while the edges associated with the consensus solutions is denoted $\mathcal{E}^{cons.}_{\text{MILP-W}}$.

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 $\text{C}_{\text{Ref}}$ is known. The ideal situation is that where a high resolution crystal structure is known, since then, all pairwise contacts can be inferred [LC10]. This reference set together with the pool $\text{Pool}_{\text{E}}(\mathcal{O})$ define positive ($P$), negative ($N$), and missed contacts ($M$) (Fig. 1). 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:

$$\text{ROC}_{sens.} = \frac{\mid TP \mid}{\mid P \mid}, \qquad \text{ROC}_{spec.} = \frac{\mid TN \mid}{\mid N \mid}. \tag{5}$$

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

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 $\mathrm{C_{Ref}}$ if the pool size is too small):

$$\mathrm{Cvg}(\mathcal{S}) = \frac{\mid TP \mid -(\mid FP \mid + \mid FN \mid)}{\mid \mathrm{C_{Ref}} \mid} \tag{6}$$

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* or the *highway* of the connectivity in the complex. However, alternative edges of significant importance may exist too. To unveil such edges, we forbid consensus edges 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, though.

First, edges corresponding to dimers must be kept for a solution to be valid. Second, hindering too many edges may yield a connectivity inference problem without any solution. Therefore, starting from the maximum number of hindered edges (the initial set of consensus solutions $\mathcal{E}^{cons.}_{\mathrm{MILP}}$, minus the dimers), we incrementally relax the constraints by considering hindered sets of smaller size (Algorithm 1).

## 4 Results

### 4.1 Test System: the Yeast Exosome

In recent work dedicated to the unweighted case [AAC+13], results were reported for several systems, including the yeast 19S proteasome lid, the eukaryotic translation factor eIF3, and the yeast exosome. The statistics obtained for these systems in terms of (consensus) solutions and (consensus) edges were comparable. On the other hand, solutions could only be assessed precisely for the yeast exosome, since a crystal structure is only known for this system. Therefore, we focus on this system in the sequel to investigate the role of weights in solving MWCI problems.

The exosome involves 10 protein types, and 20 oligomers have been reported [THS+08], ranging in size from two to nine (Table 2 and supplemental Table 4).

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. 3).

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 $\mathrm{Pool_E}(9)$ of 45 candidate edges; in the latter, the pool $\mathrm{Pool_E}(8)$ contains 36 candidate edges.

### 4.2 Algorithm `MILP-W`

In the sequel, 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, *random-*

*ized instances* are meant to investigate scenarios where no a priori information on the contacts is known.

### 4.2.1 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 $\alpha$, 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 5, 6, and 7).

**Results.** We examine successively the roles of $\alpha$ and of the individual weights.

**Parameter $\alpha$.** When $\alpha$ 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 5, $s = 8$). 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 $\alpha$.

**Weights.** The configuration yielding the maximum number of solutions comes with an average (0.45) coverage (supp. Table 5, $s = 8$). Improving this score requires optimized combinations of $\alpha$ 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 $\alpha$ 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 5, $s = 5$). This owes to the decreasing prevalence of weights when $\alpha$ increases. In a similar vein, larger values of $\Delta$, 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.

### 4.2.2 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 8). 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).

### 4.2.3 Overall recommendations

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

(i) Low values of $\alpha$ 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 4.3).

### 4.3 Algorithm MILP-$W_B$

**Yeast exosome 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). We aim to enrich this initial set of consensus contacts. 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. On switching off 7 remaining contacts simultaneously, in $\mathcal{E}^{cons.}_{\text{MILP}}$, we did not get any solution due to the fact that the pool set of contacts is not sufficient anymore to solve the problem. Holding onto the idea that new set of consensus contacts are to be found on switching off maximum number of contacts in $\mathcal{E}^{cons.}_{\text{MILP}}$, initially, we removed the label *forbidden* ('F') from one of the seven contacts. There are 7 such possibilities. No new consensus contacts are found either due to insufficient pool set or no solutions altogether. We then removed 'F' labels from two contacts at a time. The union of consensus contacts from 21 possible specifications, $\mathcal{E}_{\text{MILP-W}_\text{B}}$, has 17 TP, 7 FP, i.e. $\text{ROC}_{sens.}$ of 0.85, $\text{ROC}_{spec.}$ of 0.56 and $Cvg.$ score, 0.35 (T7 in Table 2 and Table 3). Thus, the sensitivity and coverage scores improve (respectively from 0.45 to 0.85, and from -0.15 to 0.35; T6 in Table 2; Table 3).

We also ran tests for switching off 5,4,3,2 and 1 contact(s), simultaneously. We find that cumulative set of consensus contacts does not change. On switching off 1 contact at a time, one has 7 options. The union of consensus contacts, has 16 TP and 3 FP, thus yielding, $\text{ROC}_{sens.}$, $\text{ROC}_{spec.}$ and $Cvg.$ score, respectively, 0.80, 0.81 and 0.45 (T8 in Table 2, Table 3). Interestingly, these results are better than those obtained on switching off 5 contacts simultaneously.

**Yeast exosome with Csl4.** The complete system involves 10 proteins and 20 set of oligomers. The initial consensus set has 13 TP and 3 FP out of which 3 are dimers (irreplaceable contacts). On switching off 13 contacts, one does not have any solution. On switching off 12 contacts simultaneously, the union of consensus solutions for 13 possible specifications, $\mathcal{E}_{\text{MILP-W}_\text{B}}$ has 20 TP and 6 FP, yielding $\text{ROC}_{sens.}$ of 0.77, $\text{ROC}_{spec.}$ of 0.68 and $Cvg.$ score of 0.31 (T3 in Table 2) over, initial numbers, respectively, 0.50, 0.84 and -0.12 (T2 in Table 2). When one contact at a time is switched off for this case, the triplet observed is 0.69, 0.79 and 0.23 (T4 in Table 2). Unlike in the absence of Csl4, the statistics do not improve, a fact likely related to the presence of larger oligomers.

In any case, performances are excellent when compared against those of the heuristic network algorithm [THS+08]. On the yeast exosome with Csl4, the sensitivity of MILP-$W_B$ is $\sim 1.67$ times that of network algorithm and $Cvg.$ score shows an increase of $\sim 500\%$ than the later (T1,T3 vs T0 in Table 2).

**Assessment.** Switching off the initial consensus contacts simultaneously yields new consensus contacts. However, a comparison of the edge sets $\mathcal{E}_{\text{MILP-W}_\text{B}}$ against $\mathcal{E}_{\text{MILP}}$ shows different behaviors, depending on the pool of oligomers used. In our case, the number of TP does not increase beyond that of $\mathcal{E}_{\text{MILP}}$ and the number of FP varies while remaining comparable to that of $\mathcal{E}_{\text{MILP}}$ (T1 vs T3 and T5 vs T7 in the Table 2). On this example, the bootstrapping procedure validates the contacts in $\mathcal{E}_{\text{MILP}}$, and puts confidence on the edge set, though, qualitatively.

## 5 Discussion and Outlook

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 con-

text, 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. Third, we also develop algorithm `MILP-W`$_{\mathrm{B}}$, a bootstrap strategy aiming at enriching the solutions reported by `MILP-W`. Our algorithms performance shows an increase of almost 500% in coverage score w.r.t. competing heuristic approaches. Despite the combinatorial complexity of the problems addressed, our algorithms require a hand-full of seconds for all the cases processed in this work. 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). 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.

# 6 Artwork

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

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

**Figure 1 A pool of candidate Pool$_\mathbf{E}(\mathcal{O})$ and a set of reference contacts C$_\mathbf{Ref}$ define positive ($P$), negative ($N$), and missed contacts ($M$).** Upon performing a prediction $\mathcal{S}$, $\mathcal{S}$ and its complement $\text{Pool}_\text{E}(\mathcal{O})\backslash\mathcal{S}$ further split into true/false × positives/negatives (TP, FP, TN, FN).

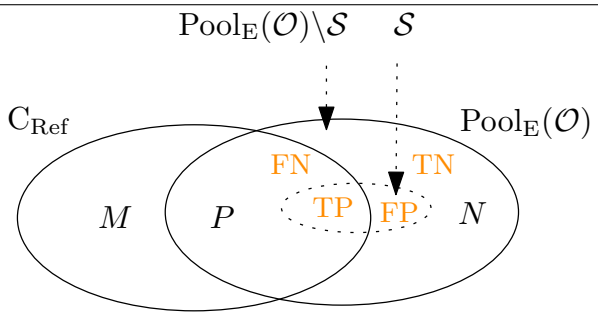
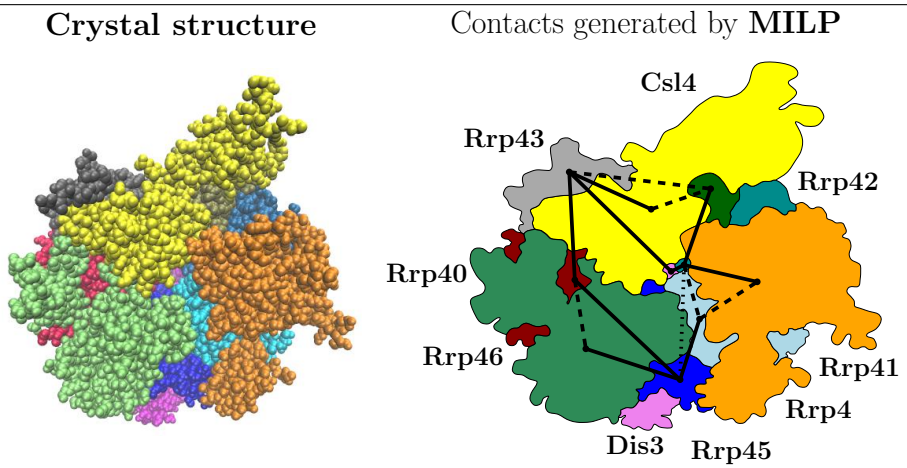



**Figure 2 The yeast exosome, an assembly consisting of 10 subunits.** The Connectivity Inference problem consists of inferring contacts between the subunits from the composition of oligomers, i.e. connected blocks of the assembly. **(Left)** Crystal structure **(Right)** The solid edges reported by the algorithm `MILP`, while the dashed edges are not present in the solution.



**Algorithm 1 Algorithm MILP-W$_\mathbf{B}$, with initial call MILP-W$_\mathbf{B}$ ($\mathcal{E}^{cons.}_{\mathbf{MILP}} \backslash I$).** The algorithm bootstraps from consensus edges, and collects novel consensus edges which appear upon precluding already found consensus edges.

**Algorithm** MILP-W$_\mathrm{B}$ ($B$)
{**Require** $\mathcal{E}^{cons.}_{\mathrm{MILP}}$: initial consensus edges}
{**Require** I: irreplaceable contacts (dimers)}
{**Require:** $spec_0$: the initial connectivity inference specification.}
{**Require:** $\mathcal{E}_{\text{MILP-W}_\mathrm{B}}$: the set storing all consensus edges, initialized to $\mathcal{E}^{cons.}_{\mathrm{MILP}}$.}
{**Parameters** B: consensus edges to be challenged := $\mathcal{E}^{cons.}_{\mathrm{MILP}} \setminus I$}
**for** $l$ from $|B|$ to 0 by step of -1 **do**
  Get $fsets_l$: all $l$-tuples from the set $B$, namely $\binom{B}{l}$
  {$C_l$: A set of consensus contacts that will be generated for all l-tuples is initiated to $\mathcal{E}^{cons.}_{\mathrm{MILP}}$}
  $C_l \leftarrow \mathcal{E}^{cons.}_{\mathrm{MILP}}$
  **for each** $fset \in fsets_l$ **do**
    {Edit the initial specification $spec_0$ to take into account the annotations}
    Assign label *forbidden* ('F') to all contacts in $fset$
    Run MWCI for this novel specification
    Get consensus contacts, $C^i_l$
    $C_l \leftarrow C_l \cup C^i_l$
  $\mathcal{E}_{\text{MILP-W}_\mathrm{B}} = \mathcal{E}_{\text{MILP-W}_\mathrm{B}} \cup C_l$

**Figure 3 Contacts between subunits of the yeast exosome without Csl4.** 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.

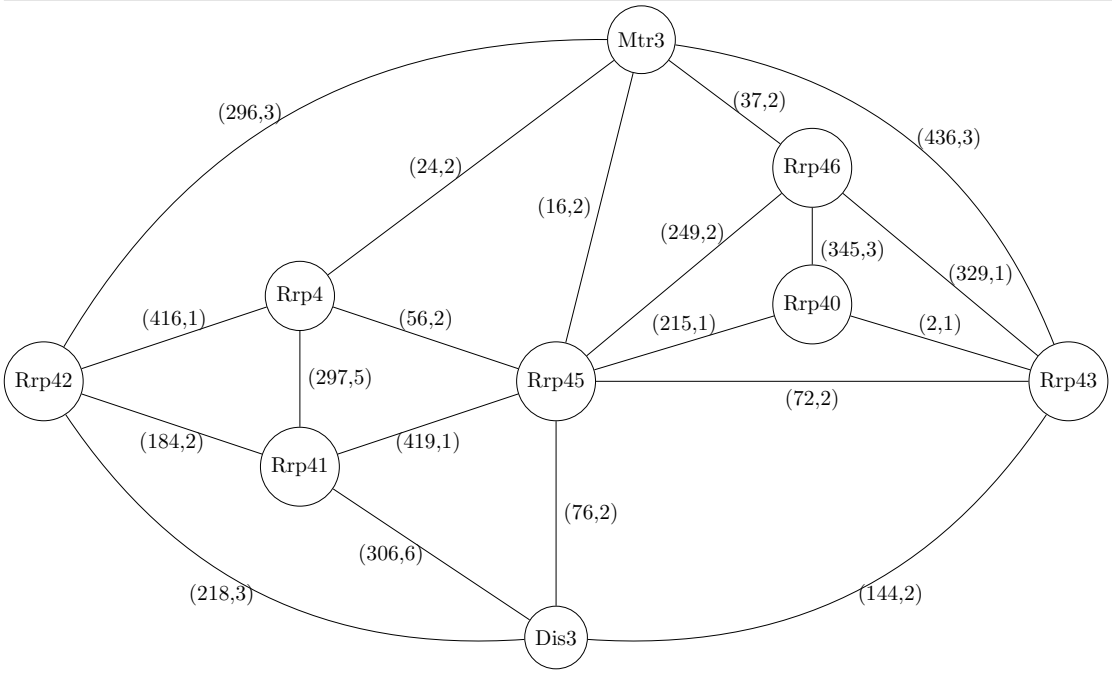


**Table 2 Sensitivity, specificity and coverage for various edge sets generated by `MILP` and `MILP-W`$_\mathbf{B}$.** Out of a pool of candidate edges of size 36, the edge set $\mathcal{E}_{\text{MILP-W}_\text{B}}$ contains all true positive but three, and ten false positives.

| 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] | 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) | $\mathcal{E}_{\text{MILP-W}_\text{B}}$ | 9 | 45 | 0 | 26 | 20 | 6 | 19 | 13 | 6 | 0.77 | 0.68 | 0.31 |
| (T4) | $\mathcal{E}_{\text{MILP-W}_\text{B}}$ | 9 | 45 | 0 | 26 | 18 | 8 | 19 | 15 | 4 | 0.69 | 0.79 | 0.23 |
| (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) | $\mathcal{E}_{\text{MILP-W}_\text{B}}$ | 8 | 36 | 0 | 20 | 17 | 3 | 16 | 9 | 7 | 0.85 | 0.56 | 0.35 |
| (T8) | $\mathcal{E}_{\text{MILP-W}_\text{B}}$ | 8 | 36 | 0 | 20 | 16 | 4 | 16 | 13 | 3 | 0.80 | 0.81 | 0.45 |

**Table 3 Sensitivity, specificity and coverage of enriched consensus set on forbidding a number of initial consensus contacts by `MILP-W`$_\mathbf{B}$.**

| #contacts fobidden, n | #combinations, $\left(\binom{7}{n}\right)$ | (size, n) $\text{ROC}_{sens.}$ | (size, n) $\text{ROC}_{spec.}$ | (size, n) Cvg | cumulative $\text{ROC}_{sens.}$ | cumulative $\text{ROC}_{spec.}$ | cumulative Cvg |
|---|---|---|---|---|---|---|---|
| 0 | 1 | 0.45 | 0.94 | -0.15 | 0.45 | 0.94 | -0.15 |
| 7 | 1 | - | - | - | 0.45 | 0.94 | -0.15 |
| 6 | 7 | - | - | - | 0.45 | 0.94 | -0.15 |
| 5 | 21 | 0.85 | 0.56 | 0.35 | 0.85 | 0.56 | 0.35 |
| 4 | 35 | 0.85 | 0.56 | 0.35 | 0.85 | 0.56 | 0.35 |
| 3 | 35 | 0.85 | 0.56 | 0.35 | 0.85 | 0.56 | 0.35 |
| 2 | 21 | 0.85 | 0.56 | 0.35 | 0.85 | 0.56 | 0.35 |
| 1 | 7 | 0.80 | 0.81 | 0.45 | 0.85 | 0.56 | 0.35 |

# 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.

[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.

[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.

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

[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.

[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.

# 7 Supplemental

## 7.1 Programs

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/`).

## 7.2 Yeast Exosome: Oligomers

**Table 4 Yeast exosome: oligomers and associated statistics. (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. NB: one protein, Csl4, is found in size 9 oligomers only. Note also that for $s = 9$ and $s = 8$, 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. 1.

| Oligomer size $s$ | $\mid \mathcal{O}_{\leq s} \mid$ | $\mid \mathrm{Pool}_{\mathrm{E}}(\mathcal{O}_{\leq s}) \mid$ | M |
|---|---|---|---|
| 2 | 3 | 3 | 17 |
| 3 | 4 | 6 | 14 |
| 4 | 6 | 13 | 7 |
| 5 | 8 | 20 | 3 |
| 6 | 9 | 21 | 3 |
| 7 | 10 | 29 | 3 |
| 8 | 15 | 36 | 0 |
| 9 | 21 | 45 | 0 |

## 7.3 Assessing the Importance of Weights

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. (6) are color coded with a heat map, from blue (0-0.1) to red (0.9 - 1).
- The values reported in Tables 8, 9, 10 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 5 Yeast exosome: statistics for U=0.1, F=0.9.**

| | | | | | $\alpha = 0.25$ | | $\alpha = 0.50$ | | $\alpha = 0.75$ | | $\alpha = 1$ | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| oligomer size, s | $\lvert \text{Pool}_{\text{E}}(\mathcal{O}_{\leq s}) \rvert$ | $\lvert M \rvert$ | mode | *sols type* | ($\text{ROC}_{sens.}$, $\text{ROC}_{spec.}$, $Cvg$) | Solutions (#, size) | ($\text{ROC}_{sens.}$, $\text{ROC}_{spec.}$, $Cvg$) | Solutions (#, size) | ($\text{ROC}_{sens.}$, $\text{ROC}_{spec.}$, $Cvg$) | Solutions (#, size) | ($\text{ROC}_{sens.}$, $\text{ROC}_{spec.}$, $Cvg$) | 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 6 Yeast exosome: statistics for U=0.25, F=0.75.**

| | | | | $\alpha = 0.25$ | | $\alpha = 0.50$ | | $\alpha = 0.75$ | | $\alpha = 1$ | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| oligomer size, s | $\lvert \text{Pool}_{\text{E}}(\mathcal{O}_{\leq s}) \rvert$ | $\lvert M \rvert$ | mode | *sols type* | $(\text{ROC}_{sens.}, \text{ROC}_{spec.}, Cvg)$ | Solutions (#, size) | $(\text{ROC}_{sens.}, \text{ROC}_{spec.}, Cvg)$ | Solutions (#, size) | $(\text{ROC}_{sens.}, \text{ROC}_{spec.}, Cvg)$ | Solutions (#, size) | $(\text{ROC}_{sens.}, \text{ROC}_{spec.}, Cvg)$ | 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 7 Yeast exosome: statistics for U=0.4, F=0.6.**

| | | | | | $\alpha = 0.25$ | | $\alpha = 0.50$ | | $\alpha = 0.75$ | | $\alpha = 1$ | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| oligomer size, s | $\lvert \text{Pool}_{\text{E}}(\mathcal{O}_{\leq s}) \rvert$ | $\lvert M \rvert$ | mode | *sols type* | $(\text{ROC}_{sens.}, \text{ROC}_{spec.}, Cvg)$ | Solutions (#, size) | $(\text{ROC}_{sens.}, \text{ROC}_{spec.}, Cvg)$ | Solutions (#, size) | $(\text{ROC}_{sens.}, \text{ROC}_{spec.}, Cvg)$ | Solutions (#, size) | $(\text{ROC}_{sens.}, \text{ROC}_{spec.}, Cvg)$ | 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 8 Yeast exosome: statistics for U=0.1, F=0.9, G=0.5.** 20 intances each.

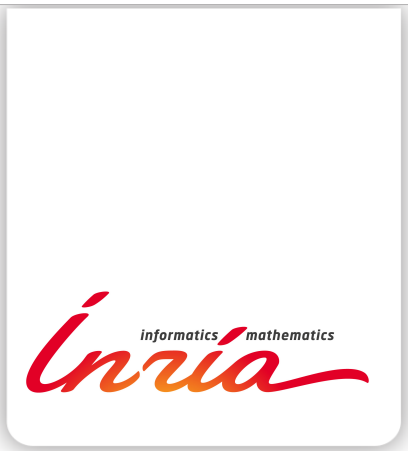
| | | | | | $\alpha = 0.25$ | | $\alpha = 0.50$ | | $\alpha = 0.75$ | | $\alpha = 1$ | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| oligomer size, s | $\lvert \text{Pool}_E(\mathcal{O}_{\leq s}) \rvert$ | $\lvert M \rvert$ | mode | *sols type* | $(\text{ROC}_{sens}, \text{ROC}_{spec}, Cvg, \sigma_{Cvg})$ | Solutions (#, size) | $(\text{ROC}_{sens}, \text{ROC}_{spec}, Cvg, \sigma_{Cvg})$ | Solutions (#, size) | $(\text{ROC}_{sens}, \text{ROC}_{spec}, Cvg, \sigma_{Cvg})$ | Solutions (#, size) | $(\text{ROC}_{sens}, \text{ROC}_{spec}, Cvg, \sigma_{Cvg})$ | 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 9 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 | $\lvert \text{Pool}_E(\mathcal{O}_{\leq s}) \rvert$ | $\lvert M \rvert$ | mode | *sols type* | $(\text{ROC}_{sens.}, \text{ROC}_{spec.}, Ctg, \sigma_{Ctg})$ | Solutions (#, size) | $(\text{ROC}_{sens.}, \text{ROC}_{spec.}, Ctg, \sigma_{Ctg})$ | Solutions (#, size) | $(\text{ROC}_{sens.}, \text{ROC}_{spec.}, Ctg, \sigma_{Ctg})$ | Solutions (#, size) | $(\text{ROC}_{sens.}, \text{ROC}_{spec.}, Ctg, \sigma_{Ctg})$ | 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 10 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 | $\lvert \text{Pool}_E(\mathcal{O}_{\preceq s}) \rvert$ | $\lvert M \rvert$ | mode | *sols type* | $(\text{ROC}_{sens}, \text{ROC}_{spec}, Cvg, \sigma_{Cvg})$ | Solutions (#, size) | $(\text{ROC}_{sens}, \text{ROC}_{spec}, Cvg, \sigma_{Cvg})$ | Solutions (#, size) | $(\text{ROC}_{sens}, \text{ROC}_{spec}, Cvg, \sigma_{Cvg})$ | Solutions (#, size) | $(\text{ROC}_{sens}, \text{ROC}_{spec}, Cvg, \sigma_{Cvg})$ | 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