SLIDER: Mining correlated motifs in protein-protein interaction networks

SLIDER: Mining correlated motifs in protein-protein interaction networks

Peter Boyen, Frank Neven, Dries Van Dyck
Hasselt University, Transnational University of Limburg
{*peter.boyen, frank.neven, dries.vandyck*}*@uhasselt.be*

Aalt D.J. van Dijk, Roeland C.H.J. van Ham
Applied Bioinformatics - Plant Research International
Wageningen UR

***Abstract*—Correlated motif mining (CMM) is the problem to find overrepresented pairs of patterns, called motif pairs, in interacting protein sequences. Algorithmic solutions for CMM thereby provide a computational method for predicting binding sites for protein interaction. In this paper, we adopt a motif-driven approach where the support of candidate motif pairs is evaluated in the network. We experimentally establish the superiority of the Chi-square-based support measure over other support measures. Furthermore, we obtain that CMM is an NP-hard problem for a large class of support measures (including Chi-square) and reformulate the search for correlated motifs as a combinatorial optimization problem. We then present the method SLIDER which uses local search with a neighborhood function based on sliding motifs and employs the Chi-square-based support measure. We show that SLIDER outperforms existing motif-driven CMM methods and scales to large protein-protein interaction networks.**

***Keywords*-correlated motifs; PPI networks; local search;**

I. Introduction

Large-scale biological networks describing interactions between proteins are available now for several organisms [13]. Such data demonstrate how proteins function as part of an interaction network, but provide no insight into how interactions are encoded in protein sequences. In particular, it is unknown which part of the sequences correspond with physical interaction sites. Unfortunately, the discovery of these sites requires laborious and expensive biological experiments. In fact, it is estimated that it would take 20 years to determine all interactions types using current experimental techniques [2]. Therefore, several computational approaches have been proposed to locate binding sites by mining overrepresented pairs of patterns (called motifs) in the sequences of interacting proteins [8], [9], [10], [11], [14]. Correlated motif mining (CMM) is an approach to identify binding sites by looking for a consensus pattern in one set of proteins which interact with (almost) all proteins which contain another consensus pattern. If so, both patterns are likely to represent a part of the surface of the molecules which makes interactions possible through a physical binding. For instance, in Fig. 1 the patterns $\{1, A\}$ and $\{2, B\}$ represent two such correlated motifs. In particular, there is an undirected edge between two protein sequences when the first one contains motif 1 (resp., 2) and the second one motif A (resp., B).

These methods can be subdivided into two classes: (i) interaction-driven [9], [10], [11]; and, (ii) motif-driven approaches [8], [14]. Interaction-driven methods mine for (quasi-)bicliques, that is, disjoint subsets of vertices for which every vertex from one set is connected to (almost) all vertices of the other set. Such subgraphs exhibit a type of all-versus-all (or most-versus-most) interaction. A motif pair representing the corresponding interaction sites is then derived from the sequence carried by the vertices. The motif-driven approach, in contrast, starts from candidate motif pairs whose support is then evaluated in the network. Although both approaches have shown to produce biologically meaningful results, the second approach has several conceptual advantages over the first: (i) motif pairs are mined directly, not derived; (ii) *all* proteins containing one of the motifs, and not a subset, are taken into account; (iii) if the interactions between two sets of proteins is a consequence of multiple compatible binding sites, such as $\{1, A\}$ and $\{2, B\}$ in Fig. 1, the interaction-driven method necessarily merges them into one motif pair; and, (iv) all interactions of proteins having both binding sites described by the motif pair are taken into account, i.e., the subsets containing each motif do not have to be disjoint.

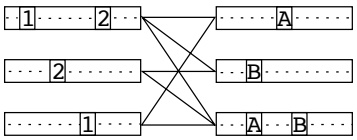


Figure 1: Compatible binding sites $\{1, A\}$ and $\{2, B\}$ as correlated motifs in sequences

In this article, we study different aspects of the motif-driven approach towards CMM for which currently only two techniques have been introduced and implemented. Unfortunately, both methods differ not only in the mining method but also in the used notion of support for correlated motifs. The first method by Tan et al. [14], called D-STAR, uses a χ^2-based scoring function to determine the support but the underlying mining method does not scale to networks containing more than 250 proteins. As contemporary biological networks contain up to thousands proteins (for instance the protein-protein interaction networks of yeast and human [4]), scalability is an increasingly important issue. The second method called MotifHeuristics employs a different, probabilistically motivated notion of support called p-score, is developed by Leung et al. [8] and does

scale to larger networks. Although the authors argue in their paper that MotifHeuristics is superior to D-STAR, it remains unclear if the latter is due to the different support measure or the underlying mining method. Moreover, an in-depth study of support measures *as such* has never been undertaken.

The contributions of this paper are: (i) a thorough empirical study of the effectiveness of various notions of support for correlated motifs; (ii) we have proven that, under reasonable assumptions concerning the used notion of support, the complexity of the correlated motif mining problem is NP-hard and its associated decision problem is in NP. We therefore approach the problem as a combinatorial optimization problem. And (iii) SLIDER, a local search method in which the key component is its neighborhood function which views a motif as a window which slides over the amino acid sequences of the proteins. We validate SLIDER by showing that it outperforms all existing motif-driven approaches on retrieving implanted motif pairs from artificial networks. Furthermore our experiments show that SLIDER is able to tackle CMM on large protein-protein interaction networks.

We assume the reader is familiar with basic concepts of graph theory and computational complexity; any basic textbook in each of these areas, like [6] and [5], will supply the necessary background. We will often use the popular abbreviation *biclique* for a complete bipartite subgraph.

II. Correlated motif mining problem

We model a protein-protein interaction (PPI) network by an undirected labeled graph $G = (V, E, \lambda)$ in which the vertices V correspond to the proteins, the edges E to the interactions and the labels of the vertices to the amino acid sequence of the proteins. Hence, the label function λ maps each vertex $v \in V$ to a string $\lambda(v)$ over the alphabet $\Sigma = \{\mathtt{A}, \ldots, \mathtt{Z}\} \setminus \{\mathtt{B}, \mathtt{J}, \mathtt{O}, \mathtt{U}, \mathtt{X}, \mathtt{Z}\}$.

An (ℓ, d)-*motif* is a string of length ℓ over the alphabet $\Sigma \cup \{\mathtt{x}\}$ containing exactly d x-characters. The character x is interpreted as a wildcard-symbol, i.e., it matches with any character of Σ. For instance, GAQPRNMY matches the $(8, 4)$-motif GxxPxNxY. A protein *contains* an (ℓ, d)-motif X if its amino acid sequence contains a substring of length ℓ that matches X.

Given an (ℓ, d)-motif X and a PPI-network $G = (V, E, \lambda)$, let $V_X = \{v \in V \mid \lambda(v) \text{ contains } X\}$, be the set of proteins in the network containing the motif X, and

$$E_{X,Y} = \{\{u, v\} \in E \mid u \in V_X \wedge v \in V_Y\},$$

be the set of interactions between proteins containing X and proteins containing Y. Hence, the subgraph $G_{X,Y}$ *selected* by a motif pair $\{X, Y\}$ is $G_{X,Y} = (V_X \cup V_Y, E_{X,Y}, \lambda_{|V_X \cup V_Y})$ with $\lambda_{|V_X \cup V_Y}$ the restriction of λ to $V_X \cup V_Y$. Note that V_X and V_Y can share proteins.

A *support measure* f is simply a function mapping a motif pair $\{X, Y\}$ and a graph G to a positive real number $f(\{X, Y\}, G)$. We refer to $f(\{X, Y\}, G)$ as the *support* of $\{X, Y\}$ in G.

We next formulate the correlated motif pair mining problem in a PPI-network (Correlated Motif Mining, CMM):

Input: a PPI-network $G = (V, E, \lambda)$, three natural numbers ℓ, d, k and a support measure f.

Output: the k (ℓ, d)-motif pairs $\{X_1, Y_1\}, \ldots, \{X_k, Y_k\}$ with highest support in G with respect to f.

III. Support measures

Support measures should reflect the power of a motif pair to describe interactions. Several considerations should be taken into account in deciding how to measure the descriptive power of a motif pair for a given PPI-network $G = (V, E, \lambda)$: (i) $E_{X,Y}$ should be significantly larger than expected given G, V_X and V_Y; and, (ii) V_X and V_Y should be large enough in order to minimize the likelihood that the appearance of the motif X, respectively Y, in the sequences of the proteins in V_X, respectively V_Y, is just by chance.

In other words, we want the motifs X and Y to truly represent an overrepresented consensus pattern in the sequences of the proteins in V_X respectively V_Y in order to increase the likelihood that they correspond to, or at least overlap with, a so called *binding site* – a part of the molecule on the surface that makes interactions between proteins from V_X and V_Y possible through a molecular lock-and-key mechanism.

We call a motif pair $\{X, Y\}$ complete if each protein from V_X interacts with each protein from V_Y.

A. A χ^2-based support measure

Tan et al. [14] introduced the χ^2-score for statistical significance as a support measure for CMM:

$$f_{\chi^2}(\{X, Y\}, G) = \begin{cases} \frac{(|E_{X,Y}| - \overline{E_{X,Y}})^2}{\overline{E_{X,Y}}} & \text{if } |E_{X,Y}| > \overline{E_{X,Y}} \\ 0 & \text{if } |E_{X,Y}| \leq \overline{E_{X,Y}} \end{cases}$$

with $\overline{E_{X,Y}}$ the expected number of interactions between V_X and V_Y, which is calculated by assuming a uniform *density* of edges. To that end, let $\mathrm{ed}(G) = |E|/\binom{|V|}{2}$ be the *edge density* of G, i.e., the proportion of edges it has of all its potential edges. Then, $\overline{E_{X,Y}} = \mathrm{ed}(G)M(V_X, V_Y)$, with the maximum amount of edges in the subnetwork

$$M(V_X, V_Y) = \left(|V_X||V_Y| - \binom{|V_X \cap V_Y|}{2} - |V_X \cap V_Y|\right) .$$

If we also use the edge density of the selected subnetwork $\mathrm{ed}(G_{X,Y}) = |E_{X,Y}|/M(V_X, V_Y)$, we can rewrite the χ^2-support of $\{X, Y\}$ for which $|E_{X,Y}| > \overline{E_{X,Y}}$ as

$$f_{\chi^2}(\{X, Y\}, G) = M(V_X, V_Y)\frac{(\mathrm{ed}(G_{X,Y}) - \mathrm{ed}(G))^2}{\mathrm{ed}(G)} .$$

As $\mathrm{ed}(G)$ is a constant for a fixed PPI-network, we clearly see in this form that f_{χ^2} uses two criteria to determine the support of a motif pair $\{X, Y\}$: (i) the difference in edge density of $G_{X,Y}$ and G, which rewards a larger $E_{X,Y}$ than

expected; and, (ii) the (potential) size of $G_{X,Y}$ in terms of the number of edges, which rewards larger V_X and V_Y.

B. p-score: a probabilistic support measure

The p-score is a measure introduced by Leung et al. [8] to evaluate the statistical significance of a motif pair $\{X, Y\}$ in a PPI-network $G = (V, E, \lambda)$ by estimating the conditional probability that there are $|E_{X,Y}|$ or more interactions between V_X and V_Y given the number of interactions involving V_X and assuming a uniform distribution of interactions over all interaction partners. Motif pairs for which this probability is small are considered to be statistically significant.

C. Comparison of f_{χ^2} and f_p

Comparing f_p with f_{χ^2}, a major difference is that f_{χ^2} bases its support on the whole network G, while f_p-support uses only a small environment of the selected subnetwork. Moreover, besides the typical edge distribution assumption, f_p makes implicitly the following additional assumptions: (i) V_X and V_Y are disjoint; and, (ii) every interaction from E_X (E_Y) can be described using X (Y), thus to calculate the support of $\{X, Y\}$ each protein is assumed to have only one binding site. Finally, we stress a design flaw in the definition of f_p: the approximation used becomes less precise when $ed(G_{X,Y})$ becomes larger, i.e., becomes more interesting.

IV. Complexity of CMM

We can prove that CMM is NP-hard for a whole class of support measures and show at the end of the section that f_{χ^2} is a member of that class.

We call a support measures *compliant* if it abides to three reasonable conditions: (i) the support can be computed efficiently; (ii) if the topology of the selected subnetworks of two motif pairs differ only in the number of edges, the one which covers more interactions has higher support; and, (iii) the support of a complete motif pair increases with the size of the selected subnetwork. We call a support measure f *biclique-maximal* if any complete motif $\{X, Y\}$ with $|V_X| = |V_Y|$ scores highest if $G_{X,Y}$ is a biclique and *clique-maximal* if the motif scores highest if it is a clique.

We can show that CMM is NP-hard by proving that even a simplified version of the associated decision (D) problem is already NP-complete. Let D-CMM be the problem to decide whether for a given PPI-network $G = (V, E, \lambda)$, natural numbers ℓ, d, a real number t and a support measure f, there exists an (ℓ, d)-motif pair $\{X, Y\}$ for which $f(\{X, Y\}, G) \geq t$.

Theorem 1: D-CMM is NP-complete for any clique- or biclique-maximal compliant support measure f.

For the proof and a more formal treatment, we refer to [3].

It is easy to see that f_{χ^2} abides the conditions and is biclique-maximal. Indeed, the support for a complete motif pair $\{X, Y\}$ where $|V_X| = |V_Y|$ in any PPI-network G is $M(V_X, V_Y)(1 - \mathrm{ed}(G))^2/\mathrm{ed}(G)$, which is maximal for $|V_X \cap V_Y| = 0$.

Figure 2: Two neighboring $(6, 3)$-motifs as sliding windows on a sequence. Moving from RTxTxx to KxxTxT, shifts the window to the left.

V. SLIDER

Since the decision problem associated with CMM is in NP, CMM can be tackled efficiently as a search problem in the space of all possible (ℓ, d)-motif pairs. If we add the assumption that similar motifs can be expected to get similar support, it has the typical form of a *combinatorial optimization problem*. A number of heuristic algorithms called *meta-heuristics* are known to yield good solutions to a wide variety of combinatorial optimization problems. One such meta-heuristic is *local search* [1]. Local search algorithms move from the current point to a neighboring point in the space of candidate solutions until a locally optimal solution is found, i.e., a solution that maximizes f in its neighborhood. This process is repeated with many initial random starting points and the best results are saved.

Thus, in order to apply local search to CMM, we need to define a neighborhood function N which maps a motif pair $\{X, Y\}$ to its neighbors $N(\{X, Y\})$ in the space of all motif pairs. Consider a motif pair $\{X, Y\}$ and the selected subnetwork $G_{X,Y}$. Ideally, the subnetwork $G_{X',Y'}$ selected by a neighbor $\{X', Y'\} \in N(\{X, Y\})$ should also be "close" to $G_{X,Y}$ in the sense that at least some proteins and interactions should be shared between $G_{X,Y}$ and $G_{X',Y'}$. To that end, we first define a neighbor function N^{slide} on motifs, which will be the basis for a neighbor function on motif pairs. Looking for a match of an (ℓ, d)-motif X in a protein can be seen as sliding a window of length ℓ with $\ell - d$ holes over the sequence until the characters in the holes match the non-wildcard characters of X. Hence, a motif X' obtained by closing a hole on a matching substring and creating a new one while respecting the window size ℓ, guarantees that the same protein will contain X'. In this way, we can slide the motif window to the left or right by punching the new hole before the first or after the last original character, as illustrated in Fig. 2 and formally defined next.

For a motif X, denote by $\mathrm{trim}(X)$, the motif obtained from X by removing leading and trailing wildcards. That is, $\mathrm{trim}(\texttt{xTxTxx}) = \texttt{TxT}$. A motif $X' \in N^{\text{slide}}(X)$ if X and X' have the same length and $\mathrm{trim}(Y) = \mathrm{trim}(Y')$ where Y (resp., Y') is obtained from X (resp., X') by changing one non-wildcard character into a wildcard. In Fig. 2, X equals RTxTxx while X' equals KxxTxT. Now, $X' \in N^{\text{slide}}(X)$ as X (resp., X') can be transformed into $Y =$ xTxTxx (resp., $Y' =$ xxxTxT) by changing one non-wildcard character into a wildcard and Y equals Y' after

stripping leading and trailing wildcards. Next, we define N^{slide} for motif pairs. That is, $\{X', Y'\} \in N^{\text{slide}}(\{X, Y\})$ if $X' \in N^{\text{slide}}(X) \wedge Y' = Y$ or $Y' \in N^{\text{slide}}(Y) \wedge X' = X$. Note that when applying N^{slide} to pairs of motifs, one of the motifs remains fixed. Our experiments, reported in Section 7.3, show that fixing one motif at each step greatly improves the effectiveness.

We define the method SLIDER as local search with: (i) neighbor function N^{slide}; and, (ii) support measure f_{χ^2}.

VI. Datasets

Artificial data. To evaluate the biological relevance of the different notions of support and the power of heuristic methods to retrieve the best motif pairs in terms of describing interactions, we created a number of artificial networks as follows. Each network is composed of 100 proteins which are randomly chosen out of the well-known yeast network [4]. We then generate 50 random $(8, 3)$-motifs [1] and implant 3 to 10 instances of each motif in the sequences of randomly chosen proteins. Then, we implant motif pairs by randomly selecting two implanted motifs X and Y and connecting each protein containing X with each protein containing Y until a chosen minimal edge density is obtained – we used 0.1, 0.2 and 0.3. Consequently, the network obtained is perfect in the sense that each interaction is a direct consequence of an implanted motif pair. Because noise and missing data are an important factor in biological networks, we also evaluate the resistance to noise of both the support measures and heuristic methods. To that end, we create "diluted" versions of each network, by choosing a dilution level a (from 0.05 to 0.3 in 0.05 steps) and flipping the edge relation of each pair of vertices with probability a.

We restrict ourselves to networks of 100 proteins because this is more or less the maximum size for which we are still able to mine the motif pairs with highest support for each support measure by a brute force computation within a reasonable time frame.

Biological data. To assess the effectiveness of SLIDER on larger networks, we ran our method on the high-confidence protein-protein interaction network of yeast [4] consisting of 1620 nodes and 9060 interactions. It is very difficult to measure the biological significance of the found correlated motifs, because only very few of them are actually known. Therefore, we executed a brute force CMM-algorithm over the yeast network on a computer cluster, finding the best 1 000 correlated motifs according to f_{χ^2} and compared these to the results returned by SLIDER. The brute force computation occupied about 100 nodes in the cluster spanning a period of 2 weeks. Its purpose was to create a baseline for motif-driven CMM-algorithms as well as collecting the best correlated yeast motifs for biological analysis (which is still ongoing at this point).

[1] Using entropy analysis, research has shown that the highest amount of structural information per sequence length can be found in subsequences of length 7 to 9 (see Fig. 1 in [12]).

VII. Experiments

With the exception of the brute force run on yeast, all experiments were run on a 3GHz Mac Pro with 4GB of RAM and 8 cores. In the sequel, whenever a timing is mentioned and unless explicitly mentioned otherwise, the experiment was run using only 1 core. Nevertheless, we stress that our SLIDER-prototype, implemented in Java, can use as many processors as are available. In this section, we experimentally assess the effectiveness of (i) support measures to assign a support to a motif pair reflecting its power to describe interactions; and, (ii) neighbor functions to find the motif pairs with highest support by exploring the space of all motif pairs. Furthermore, we compare SLIDER with other motif-driven CMM-methods. To this end, we need a notion of precision that compares an ordered set of motif pairs to a set of motif pairs which is considered to be the "ground truth". Finally, we assess the effectiveness of SLIDER on the yeast network.

A. Precision for motif pairs

We define the similarity between (ℓ, d)-motif pairs $\{X, Y\}$ and $\{X', Y'\}$ in a PPI-network $G = (V, E, \lambda)$ as

$$s(\{X, Y\}, \{X', Y'\}, G) = \frac{|E_{X,Y} \cap_{pos} E_{X',Y'}|}{|E_{X,Y} \cup E_{X',Y'}|},$$

where $\{v, w\} \in E_{X,Y \cap_{pos} X',Y'}$ if there exists substrings s_v and s'_v in $\lambda(v)$ and substrings s_w and s'_w in $\lambda(w)$ such that s_v (resp., s_w) matches with X (resp., Y), s'_v (resp. s'_w) matches with X' (resp., Y'), and, s_v and s'_v as well as s_w and s'_w overlap in at least $\lceil \ell/3 \rceil$ positions in $\lambda(v)$ respectively $\lambda(w)$.

Let $S = \{M_1, \ldots, M_n\}$ be a list of motif pairs, then we reduce S by deleting for every j from 1 to n, every M_i for $i > j$ such that $s(M_i, M_j) \geq 0.9$. We denote the reduced version of S by S^*.

Let T be a set of "ground truth" (ℓ, d)-motif pairs and let $S = \{M_1, \ldots, M_n\}$ be a list of (ℓ, d)-motif pairs to be compared against T. We define the precision of S against T at rank k as the fraction of motif pairs M_i in S^*, $1 \leq i \leq k$ for which there exists a motif pair M_T in T^* such that $s(M_i, M_T) \geq 0.9$. We note that, when $k = |T^*|$, the precision as defined above also corresponds to the usual notion of recall.

B. Evaluation of support measures

Since the most describing motif pairs in real PPI-networks are unknown, we measure the ability of a support measure to assign the highest support to motif pairs on artificial networks with implanted motifs, as described in Section VI. We used a collection of networks G_e^a with edge density $e\%$ and dilution level $a\%$. We compare the support measures by looking at the precision of implanted motif pairs on G_e^a at rank m, where m equals the number of implanted motif

pairs. Remark that, in this setting, recall and precision are the same.

In order to make sure that the f_{χ^2} and f_p assign a meaningful support, we also implemented two naive support measures f_c and f_v. The f_c-support in a PPI-network $G = (V, E)$ is simply the number of interactions covered: $f_c(\{X, Y\}, G) = |E_{X,Y}|$ and

$$f_v(\{X, Y\}, G) = \frac{|E_{X,Y}|}{M(V_X, V_Y) + |V_X \cup V_Y|} \ .$$

Both measures are naive in that they are independent of the interaction distribution in G. It is straightforward to show that both measures are compliant and biclique-maximal.

A visual inspection of the graph in Fig. 3a already indicates that f_{χ^2} globally outperforms the other support measures in selecting motif pairs describing actual interactions. Indeed, at every data point, the precision of f_{χ^2} is the best value or very close to the best value of the four support measures considered. Moreover, comparing precision on diluted networks shows that f_{χ^2} is vastly more robust to noise — a crucial aspect since contemporary PPI-networks still contain large amounts of both noise and missing data [16]. The results on the artificial networks with a greater density showed the same trends.

Thus, we can conclude from this experimental section that f_{χ^2} is superior to all other support measures considered on all merits.

C. Evaluation of neighborhood functions

We also defined several naive neighbor functions on motif pairs, based on simple perturbations of its component motifs. Fig. 3b displays the precision of local search with each of these neighborhood functions on the implanted network of density 10%. The displayed precision is averaged over 5 local search runs. As the speed of local search is highly dependent on the chosen neighbor function, we provided each run the same amount of time (10 minutes).

For the sake of completeness, we also experimented with neighborhood functions on motif pairs where both motifs can be replaced with a neighboring one (in contrast to the previous neighborhood functions where one is kept fixed). Unfortunately, the precision was never larger than 10%, independent of the level of dilution.

D. Comparison with existing methods

D-STAR. Tan et al. introduced the first motif-driven method for CMM: D-STAR [14]. Strictly speaking, D-STAR does not deliver (ℓ, d)-motifs. Instead it returns two strings s_X and s_Y, and two sets of proteins V_X and V_Y together with the indices of the substring of the amino acid sequence of each protein in V_X (respectively V_Y) that differs at most $2d$ characters from s_X (respectively s_Y). As the similarity in Section 7.1 is defined in terms of positions of substrings, we can directly use the returned subsets V_X and V_Y to compare with implanted motifs. Every run of D-STAR on the same network produced the same result, consequently the running time of D-STAR cannot be parameterized. We used the D-STAR implementation freely available on the web.

MotifHeuristics. Another method, called MotifHeuristics, proposed by [8], derives (ℓ, d)-motifs directly within the wildcard model and introduced the probabilistically motivated f_p-support measure. Because we could not obtain an implementation of MotifHeuristics, we implemented our own version based on the algorithmic description in [8].

Comparison. The graph in Fig. 3c depicts the precision of the various methods on the artificial network of density 10%. As a naive baseline, we ran the method Random, evaluating random motif pairs using f_{χ^2}. D-STAR took 5 minutes to finish. We gave Random and SLIDER 10 minutes of computation time. In order to give our unoptimized implementation of MotifHeuristics a fair chance, we allowed it to run 30 times longer than SLIDER (that is 5 hours). The underlying reason why MotifHeuristics takes such a long time is that for every search step a number of supports has to be calculated which approaches the total number of motifs. The graph makes it quite apparent that the success rates of both D-STAR and MotifHeuristics are smaller than or equal to that of SLIDER. Overall, SLIDER is more effective and more robust than its competitors. All algorithms perform significantly better than random search.

When we double the execution time of SLIDER to 20 minutes, the precision increases significantly. The latter execution time is still minor in comparison with the brute force computation which takes about 40 hours.

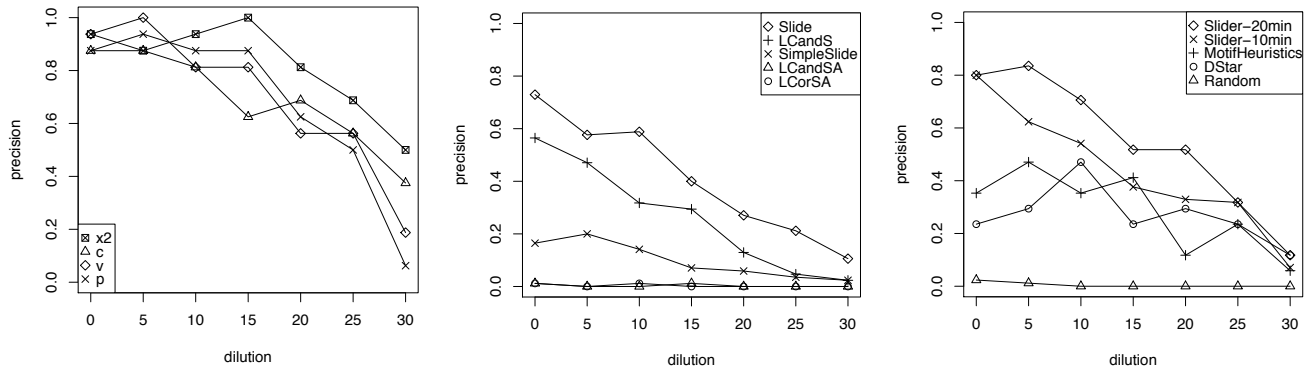
E. Biological validation

Next, we assess the effectiveness of SLIDER on the yeast network. We did not try MotifHeuristics as it already takes a long time on networks of modest size (cf. Section 7.4). Furthermore, although D-STAR terminated on our artificial networks within 5 minutes, the method does not scale to larger networks. In particular, Leung et al. [8] mention an experiment where they executed D-STAR on the yeast network and it did not finish in 5 days, we ourselves have run D-STAR on this network for 48 hours without result.

We ran SLIDER for 20 minutes exploiting all 8 cores of the Mac Pro. The average precision of the 1 000 best results returned by SLIDER over 5 runs, while taking the 1 000 best motifs returned by the brute force computation as a baseline, is 16%. We point out that the name precision is misleading in this context as we do not compare with implanted motifs. The number implies that SLIDER succeeds in recovering no less than 160 of the 1000 best correlated motifs out of a search space of 6×10^{15} (8,3)-motif pairs after only a run of 20 minutes which is quite satisfactory.

VIII. Conclusion

At first sight the present work seems highly related to the mining of frequent patterns in sequences (as for instance in [7]). It is therefore tempting to think about a method which



(a) Support measures (b) Neighborhood functions (c) Existing methods

Figure 3: Precision on artificial networks with implanted motifs of density 10%.

first mines frequent motifs from protein sequences which are then paired together in a second step serving as candidates for high scoring correlated motifs. An examination of the 1 000 top correlated motifs in yeast, however, reveals that each of the participating motifs occur only in 3 to 10 motifs, whereas highly frequent motifs in yeast occur in up to 60 proteins. Therefore, mining correlated motifs is very different from mining frequent motifs.

Van Dijk et al. [15] showed how motifs generated by D-STAR can be used to predict transcription factor interaction on small networks. Using SLIDER rather than D-STAR, the same methodology can be applied to larger networks.

Finally, we mention that we could not confirm the claimed superiority in [8] of MotifHeuristics over D-STAR. In fact, our results clearly show that f_p is inferior to f_{χ^2} in recovering implanted motifs. These tests should be repeated on real world data, but as long as only few biological correlated motifs are known this is not possible.

ACKNOWLEDGMENT

Research funded by a Ph.D grant of the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen). This work was supported by the BioRange programme (SP 2.3.1) of the Netherlands Bioinformatics Centre (NBIC), which is supported through the Netherlands Genomics Initiative (NGI).

REFERENCES

[1] E. Aarts and J. Lenstra, editors. *Local Search in Combinatorial Optimization*. John Wiley & Sons, 1997.

[2] P. Aloy and R. Russell. Ten thousand interactions for the molecular biologist. *Nat Biotechnol.*, 22:1317–1321, 2004.

[3] P. Boyen, F. Neven, D. Van Dyck, A. van Dijk, and R. van Ham SLIDER: Mining correlated motifs in protein-protein interaction networks *Technical report, Database and Theoretical Computer Sciences Research Group, Hasselt University, http://hdl.handle.net/1942/9865*, 2009.

[4] S. Collins et al. Towards a comprehensive atlas of the physical interactome of saccharomyces cerevisiae. *Mol Cell Proteomics.*, 2007.

[5] R. Diestel *Graph Theory, Third edition* Springer-Verlag, 2005.

[6] M. Garey and D. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. 1979.

[7] K. Gouda, M. Hassaan, and M. Zaki. Prism: A primal-encoding approach for frequent sequence mining. In *ICDM*, pages 487–492, 2007.

[8] H. Leung, M. Siu, S. Yiu, F. Chin, and K. Sung Finding linear motif pairs from protein interaction networks: A probabilistic approach. In *Computational Systems Bioinformatics (CSB)*, pg. 111–120, 2006.

[9] H. Li, J. Li, and L. Wong. Discovering motif pairs at interaction sites from protein sequences on a proteome-wide scale. *Bioinformatics*, 22(8):989–996, 2006.

[10] J. Li, G. Liu, H. Li, and L. Wong. Maximal biclique subgraphs and closed pattern pairs of the adjacency matrix: A one-to-one correspondence and mining algorithms. *IEEE Trans. Knowl. Data Eng.*, 19(12):1625–1637, 2007.

[11] J. Li, K. Sim, G. Liu, and L. Wong. Maximal quasi-bicliques with balanced noise tolerance: Concepts and co-clustering applications. In *SDM*, pages 72–83. SIAM, 2008.

[12] M. Šikić, S. Tomić, and K. Vlahoviček. Prediction of protein-protein interaction sites in sequences and 3d structures by random forests. *PLoS Comput Biol*, 5(1):e1000278+, 2009.

[13] M. Stumpf et al. Estimating the size of the human interactome. *Proc Natl Acad Sci U S A*, 105(19):6959–64, 2008.

[14] S. Tan, W. Hugo, W. Sung, and S. Ng A correlated motif approach for finding short linear motifs from protein interaction networks. *BMC Bioinformatics*, 7:502+, November 2006.

[15] A. van Dijk, C. ter Braak, R. Immink, G. Angenent, and R. van Ham Predicting and understanding transcription factor interactions based on sequence level determinants of combinatorial control. *Bioinformatics*, 24(1):26–33, 2008.

[16] C. von Mering et al. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, 417:399–403, 2002.