Finding Motifs in Protein-Protein Interaction Networks

Computational Molecular Biology and Genomics: Project Final Report

Juchang Hua, David Koes, and Zhenzhen Kou
{juchangh,dkoes,zkou}@andrew.cmu.edu
Carnegie Mellon University

December 5, 2003

Abstract

High-throughput methods of protein interaction analysis have resulted in an abundance of data that requires complex computational approaches to interpret. We attempt to extract biological meaning from the topology of a protein-protein interaction graph. We describe an algorithm for finding interesting motifs within the graph and then describe our limited success at deriving biological meaning from the results.

1 Introduction

High-throughput analyses, combined with the exponential growth of genome sequencing, make possible the study of protein-protein interactions at a genomewise scale. When the results of these techniques are combined, we can define the protein networks which operate in living cells. There are two basic types of protein networks.

One is the protein-protein interaction network which shows the direct physical interactions between protein pairs. This network indicates the functional and structural relationship among its nodes. Various post-translational regulations such as phosphorylation and catalysis can be found in it. The high-throughput two-hybrid experiments give systematic information about the specific binding of yeast (Saccharomyces cerevisiae) proteins [3][2]. Also, a recent combined experimental and computational approach defines the protein interaction network for domain recognition[7].

The other basic type of protein network is the genetic regulatory network, which shows how one protein regulates the expression of another one. Most of these regulations occur at the level of transcription. By binding to the gene of the controlled protein, the transcriptional factor can up-regulate or down-regulate the transcription of the RNA. Genetic regulatory networks are usually obtained using gene expression microarray or DNA-chip technologies[4]. Because the number of proteins in a living cell is large, the topologies of the protein networks are usually very complicated. However, based on statistical analysis, researchers have found that protein networks have some properties which make them specific and stable. For example, these networks share the scale free property with the Internet[5].

Protein networks are so large that it is a challenge to extract biological functions or pathways from them even if some global features have been found. The natural thought is to break a network into small elements. “Network motifs”, defined by R. Milo et al in their Science publication [6], are now well-accepted as topological units of protein networks. Network motifs describe the interaction patterns among a few nodes, and these patterns appear a significant number of times all over the protein network. Also, these patterns appear many more times in the real network than in a randomized network, which suggests that they are significant. By comparing the frequency a motif occurrs in a real network to the expected frequency in a randomized network, R. Milo et al found some interesting motifs which are statistically significant and indicative of biological functions. They also found that other complex networks such as the ecological food network and the Internet have their own basic motifs. As the motifs found in distinct networks are different, they can be used to universally classify the networks.

Recently, Johannes Berg and Michael Lässig reported a new algorithm of motif search in protein networks[1]. Inspired by gene sequence alignment, they perform a local graph alignment based on a score function measuring the significance of the subgraphs found. They applied this algorithm to the E. coli regulatory network. Our project partially implements their algorithm and applies it to the protein-protein network from the yeast 2-hybrid experiments.

The 2-hybrid experiment is used to study the physical interaction of two proteins. Two-hybrid refers to the fact that both the proteins studied are hybrids. By hybridizing the pair of proteins to a DNA binding domain (DBD) and activation domain (AD), the expression of the report gene indicates that the pair of proteins interact with each other. The DBD and AD are put together to initialize the transcription of the report gene. Although false positives can be introduced by DBD or AD activity of the subject proteins themselves, and false negatives can be introduced by low binding stability, the two-hybrid experiment is the most applicable approach for protein interaction studies. Ito et al. were the first to established the genome-wide scale of two-hybrid screening and introduced the term "interactome"[3][2]. They cloned all the yeast ORFs (Open Reading Frames) as both DNA-binding domain fusions and activation domain fusions in MATa strains and screened by mating. In addition to the positive reaction in the two-hybrid, they did sequence-tagging of the pair of proteins to obtain the Interaction Sequence Tags (ISTs). Only the interactions with more than 3 IST hits are included in the core data. Their full data contains 4549 interactions and the core data includes 841 interactions among 797 proteins. Because of the increased reliability of the data and the reduced computation complexity, we only tested our program on the core data of Ito et al.

2 Algorithm

In Berg and Lässig's paper[1], they discuss probabilistic motifs derived from families of mutually similar but not necessarily identical patterns. A statistical model for the occurrence of such motifs in a graph is established, from which a scoring function for their statistical significance is derived. Based on this scoring function, a heuristic search algorithm for topological motifs, called graph alignment, is derived.

A key contribution of Berg's paper[1] is the development of the idea of probabilistic motifs. Shen and Milo[6] found motifs which occur more frequently in real networks compared to a suitable null ensemble, but the motifs were exactly identical in shape. Berg generalized the notion of a motif to a stochastic one. The motifs found do not need to be topologically identical. Probabilistic motifs arise as a consensus of a family of sufficiently similar subgraphs in a network. This variation tolerance meets the important characteristic of biological systems, which is similar to sequence analysis where one searches for local sequence similarities blurred by mutations and insertions/deletions rather than for identical subsequences. This was an important reason for us choosing Berg's algorithm for our project.

To distinguish interesting motifs from random subgraphs, Berg characterized the motifs of interest in two ways: they have an enhanced number of internal links, e.g. associated with feedback, and they appear in a significant number of subgraphs. Identifying these local deviations from randomness in networks requires the coherent statistical mechanics of local graph structure, which is established in Berg's paper. Based on the statistical mechanics a scoring scheme is established.

Berg's local graph alignment is conceptually similar to sequence alignment. It is based on a scoring function measuring the statistical significance for families of mutually similar subgraphs. This scoring involves quantifying the significance of the individual subgraphs as well as their mutual similarity, and is thus considerably more complicated than for families of identical motifs.

As a computational problem, graph alignment is more challenging than sequence alignment. Sequences can be aligned in polynomial time using dynamic programming algorithms. For graph alignment, a polynomial-time algorithm does not exist unless P equal NP (graph isomorphism is in NP). Thus, an important issue for graph alignment is the construction of efficient heuristic search algorithms. Berg solved this problem by mapping graph alignment onto a spin model familiar in statistical physics, which can be solved by simulated annealing.

3 Terminology

We adopt the same terminology as Berg and Lässig which, for convenience, we repeat here.

A graph is a set of nodes and links. Labeling the nodes by an index $r = 1, \dots, N$ the network is described by the *adjacency matrix* C, which has entries $C_{rr'} = 1$ if there is a directed link from node r to node r' and $C_{rr'} = 0$ otherwise. The special case of a symmetric adjacency matrix can be used to describe undirected graphs. In a undirected graph, the connectivity of a node, $k_r = \sum_{r'} C_{rr'}$, is defined as the number of links. The total number of links in an undirected graph is denoted by

$$K = \frac{\sum_{r,r'} C_{rr'}}{2}$$

A subgraph of G is given by a subset of k vertices

$\{r_1, ...r_k\}$ and the resulting restriction of the adjacency matrix. More precisely, we define the matrix $c(G, A)$ with the entries

$$c_{i,j} = C_{r_i r_j} (i, j = 1, \dots , k)$$

specifying the internal links of the subgraph for a given ordering A of the nodes. This matrix c is called a motif, which is contained in the graph G. The most important characteristic of motifs for what follows is their number of internal links,

$$L(c) = \sum_{i,j} c_{ij}$$

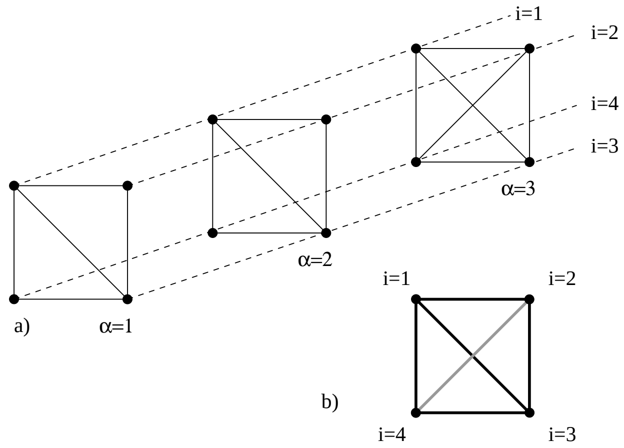


Figure 1: Graph alignment and consensus motif

A graph alignment is defined by a set of several subgraphs $G^\alpha (\alpha = 1, \dots , p)$ and a specific order of the nodes $\{r_1^\alpha, \dots , r_n^\alpha\}$ in each subgraphs; this joint order is again denoted by A. We assume here that the subgraphs are of the same size k. For mutually disjoint subgraphs there are $(k!)^p$ different alignments of the same set of subgraphs. An alignment associates each node in a subgraph with exactly one node in each of the other subgraphs. This can be visualized by k "strings", each connecting nodes with the same index i as shown in Figure 1

A given alignment A specifies a motif in each subgraph $c^\alpha \equiv c(G^\alpha, A)$. The consensus motif of this alignment is given by the matrix

$$\overline{c} = \frac{1}{p} \sum_{\alpha=1}^{p} c^\alpha$$

The consensus motif is a probabilistic motif, the entry c_{ij} denoting the likelihood that a given link is present in the aligned subgraphs. For any two aligned subgraphs G^α and G^β, we can define the pairwise mismatch

$$M(c^\alpha, c^\beta) = \sum_{i,j=1}^{n} [c_{ij}^\alpha (1 - c_{ij}^\beta) + (1 - c_{ij}^\alpha) c_{ij}^\beta]$$

The mismatch is 0 if and only if the matrices c^α and c^β are equal, and is positive otherwise. The average mismatch over all pairs of aligned subgraphs, $\overline{M} \equiv M(\overline{c}, \overline{c})$, is termed the *fuzziness* of the consensus motif $\overline{c}$.

4 Implementation

We have implemented the algorithm of [1] with some simplifications. The main differences are that we use different methods for removing "uninteresting" subgraphs from consideration, we don't use their $\log Z$ normalization factor in the score function, we do not perform parametric optimization, and we use a simple greedy algorithm for multiple alignment construction instead of a second pass of simulated annealing.

We have converted the yeast two-hybrid protein interaction data from Ito into a standard plain text format used by our motif finding application. A sample of this intermediate form is shown in Figure 2. Each protein is assigned a unique integer identifier, is tagged with both its name and a short description, and lists the unique identifiers of the proteins it interacts with. Our application could work with any data source once the data was converted into this format.

The application reads in the specified input file and constructs an undirected graph represented using both an adjacency matrix and adjacency lists. All subgraphs of a specified size k are then found using a recursive enumeration algorithm.

k	**connected subgraphs**	**very connected subgraphs**
3	3947	33
4	46766	199
5	586545	840
6	6709002	3552

Table 1: Number of subgraphs in the graph formed from the Ito data set as the size of the subgraph, k, increases.

The number of subgraphs of size k in a graph with n nodes is $O(\binom{n}{k})$ which grows exponentially in k. Even though we restrict ourselves to connected subgraphs,

```
797
core data: 841 interactions with more than 3 IST hits

ECM11
Protein possibly involved in cell wall structure or biosynthesis
102
399

AMS1
Alpha-mannosidase, hydrolyzes terminal non-reducing alpha-D-mannose residues from alpha-D-mannosides
10
752

DUO1
Protein that interacts with Dam1p and causes cell death upon overproduction
100
84 38 86 416

<continued...>
```

Figure 2: A portion of the Ito yeast two-hybrid data in our intermediate representation.

empirically the number of subgraphs still increases exponentially as shown in Table 1.

Because the number of subgraphs increases so rapidly and the running time and memory consumption of the algorithm is $\Omega(g^2)$ where g is the number of subgraphs, it is necessary to restrict the number of subgraphs considered to just those that are considered to be interesting. We evaluated two methods of reducing the number of subgraphs. Both methods can by applied during subgraph enumeration so that the application doesn't run out of memory in the first phase of the algorithm.

One way to limit ourselves to only interesting graphs would be to only consider those subgraphs where every node of the subgraph has more than one link. We refer to these subgraphs as being "very connected." Very connected subgraph exclude "dangling links" and are also used by [1] in gene regulation networks. It is not clear that there is a reasonable biological justification for imposing this limitation in a protein-protein interaction graph.

Another way we limit the number of subgraphs is by only considering those subgraphs that are unlikely to occur in a random graph with the same connectivities as our input graph. The probability of a subgraph G represented by the adjacency matrix c in this *null ensemble* can be approximated by

$$P_0(G) = \prod_{i,j=1}^{n} (1 - w_{ij})^{1-c_{ij}} w_{ij}^{c_{ij}}$$

where $w_{ij} = degree_i * degree_j / K$. Having computed the probability of each subgraph, we can then consider some threshold-ed number of subgraphs which have a very low probability of appearing by chance alone. Since we are not using the $\log Z$ normalization factor in our scoring function, we expect this technique to be particular effective at removing uninteresting subgraphs.

A third method for reducing the number of subgraphs used by the algorithm would be to compute on generic shapes of subgraphs. There are $O(2^{\frac{k(k-1)}{2}})$ graphs with k nodes. Empirically this is much less than the $O(\binom{n}{k})$ subgraphs we find. This is because many of the subgraphs we find are isomorphic. Instead of computing on the subgraphs themselves, we could compute on a generic *shape* of a subgraph that is weighted by the number of corresponding subgraphs in the actual data. This would require fairly substantial changes to the algorithmic and graph processing code, and due to time constraints we chose not to pursue this method.

Once we have an appropriate set of subgraphs, we then compute the best pairwise alignment for every pair of subgraphs. For each pair of subgraphs, α and β, we enumerate the $n!$ possible alignments and score each alignment using the formula

$$M(c^{\alpha}, c^{\beta}) = \sum_{i,j=1}^{n} [c_{ij}^{\alpha}(1 - c_{ij}^{\beta}) + (1 - c_{ij}^{\alpha})c_{ij}^{\beta}]$$

We define $M^{\alpha\beta}$ to be the minimum score possible between α and β.

The pairwise alignments are used to find a multiple alignment. Instead of finding a multiple alignment directly, we find a set of subgraphs which maximize a score function and then construct a multiple alignment out of this set using a greedy algorithm. The function we use to score set of subgraphs $\mathcal{S}$ is

$$S(\mathcal{S}) = \sigma \sum_{\alpha \in \mathcal{S}} L^{\alpha} - \frac{\mu}{2|\mathcal{S}|} \sum_{\alpha,\beta \in \mathcal{S}} M^{\alpha\beta}$$

The σ and μ parameters are specified by the user. L^{α} is the number of internal links in subgraph α. Thus, we reward sets of subgraphs containing many internal links and penalize sets with many pairs of subgraphs that have poor alignment scores. The σ and μ parameters can be used to balance the effects of these two factors. A significant omission from this formula is a normalization factor that adjusts for the likelihood of subgraphs appearing in the null ensemble. Such a factor ($\log Z$) is very difficult to compute as it requires that we construct the multiple alignment and maximize a complicated function every time we score a subset. Although incremental approximations can be used to improve performance, we omit it for reasons of time (both ours and the running time of the application). It also should be pointed out that even with the normalization factor, the maximization of this score function would not necessarily correspond exactly with the optimal multiple alignment since it is built from the pairwise alignment scores.

We attempt to find a set that maximizes our score function using simulated annealing. We start with a random subset of all subgraphs and randomly pick a subgraph to be excluded or included in this subset. We compare the scores of the original subset and the modified subset. If the score improves, we keep the change. If the score does not improve we will only keep the change with some diminishing probability. This allows us to exit a local minimum. We repeat this process for some fixed number of iterations and hope that the result is a decent approximation of the actual global maximum.

Once we've found a set of subgraphs that does a decent job of maximizing our score function, we construct a multiple alignment from this set using a simple greedy algorithm. We start by fixing the alignment of a single subgraph from the set. Then we inductively pick a subgraph from the set and fix its alignment to be the alignment that aligns best with the most already-aligned subgraphs. Since the subgraphs we find tend to be very similar, this approach works very well in practice.

Having fixed a multiple alignment from the set of subgraphs that maximized our score function, it is simple to build a consensus motif $\overline{c}$. The consensus motif is just a single subgraph which represents the whole multiple alignment by averaging the effects of the individual subgraphs of the alignment.

5 Results

We've applied our implementation of the algorithm to the yeast two-hybrid core data and extracted consensus motifs of sizes 3, 4, and 5. In Figure 3 we show the effect of increasing μ. As this parameter is increased, the fuzziness of the graph (the amount of disagreement in the alignment) decreases and the total number of members in the alignment decreases. There is little biologically relevant information that we could extract from this motif because it was so common. The main cause of the frequency of this motif is the existence of several very "popular" proteins which bind with many other, usually unrelated, proteins. For example, a common protein found at node 3 in the multiple alignment was SRP1, which binds with over 50 proteins. Although identifying such "popular" proteins is useful, they are more easily identified by the simple heuristic of looking at the highest degree nodes in the interaction graph.

In Figure 4 we compare the two different techniques we used to limit the number of subgraphs we considered using a subgraph size of four. In Figure 4(a), all nodes have a degree greater than one, making for a potentially more interesting motif. However, we find that all we're doing is identifying two very "popular" proteins. Over 90% of the subgraphs in this alignment have the proteins SRP1 and TEM1 at positions 3 and 4. Furthermore, it may be that limiting ourselves to only very connected graph is too restrictive. Only 199 of the 46766 possible subgraphs met this criteria. In Figure 4(b), we examine the result of limiting the subgraphs to only 1000 least likely to occur in the null ensemble. The result is a chain of interacting proteins. Interestingly, contained within this motif are four chains which, when merged, form the clatherin-associated protein complex AP-1 as shown in Figure 5. However, this is just four members of a 486 alignment and it is not clear that there is an interesting biological significance to the remaining members of the alignment. Furthermore, it is worth pointing out that the algorithm only succeeded in reducing the initial 1000 subgraphs to 486 interesting subgraphs, which is not a very big reduction.

In Figure 6 we show the motifs we found of size five, limiting the number of subgraphs to the 2000 least likely to occur. In analyzing the data, we found that there were many sub-motifs. That is, many members of the multiple alignment were identical except for a

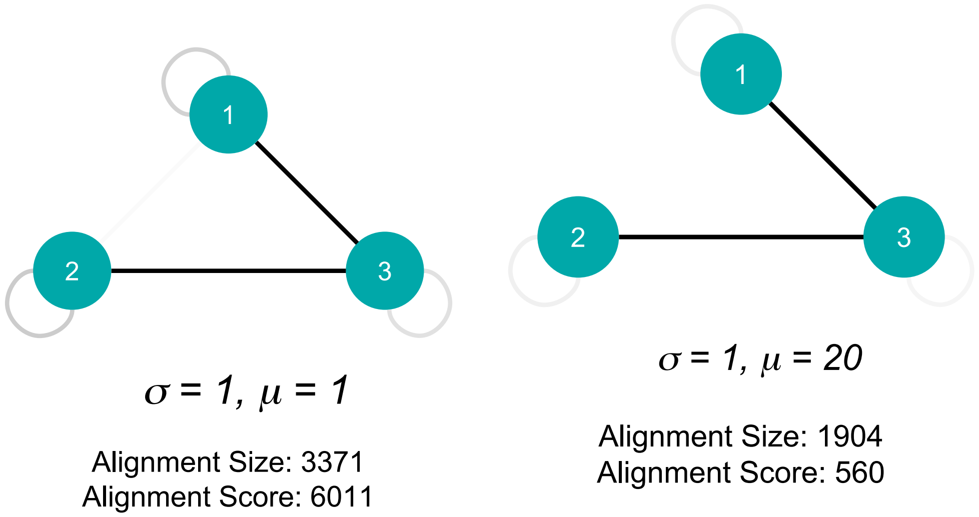


Figure 3: Motifs of size three. The number of initial subgraphs was not restricted. As μ increases the fuzziness of the motif decreases.

1	2	3	4	5	Number matching subgraphs
SNO1	SRP1	**SNZ3**	**SNZ2**	x	13
SNZ1	SRP1	**SNZ3**	**SNZ2**	x	13
SNO2	SRP1	**SNZ3**	**SNZ2**	x	23
SNO1	SRP1	**SNZ3**	**SNZ1**	x	24

Table 2: We can look for overlap within the members of multiple alignment. In the case of the k = 5, threshold-ed data, 8.7% of the members of the multiple alignment are represented by subgraphs containing the four above patterns. The x is a different protein in each subgraph. Protein names in bold are B6 vitamins. The interconnected-ness we found between them is probably a result of their similarities (that is, one can replace another so they interact with the same proteins).

few proteins. Inspired by the clatherin AP-1 complex found in the motif of size four, we found those members of the motif which had the most overlap (differed by just one protein). The top results are shown in Table 2. This technique was successful at discovering a set of vitamin B6 proteins.

6 Conclusion

Our hope in performing these experiments was that we would be able to extract biological meaning from motifs we found in protein-protein interaction networks. Our technique was not very successful at identifying interaction subgraphs from which we could find biological meaning. One reason may be that we omitted the log Z normalization factor resulting in a faster, but possibly less interesting, score function. Another possibility may be that, due to the nature of the interactions, the shape of a protein-protein interaction subgraph may communicate little information that is biologically interesting. However, it is possible that when combined with other analyses, such as finding the maximal overlap of the members of the motif, our technique could yield interesting results.

Our source code is available from `http://www.cs.cmu.edu/ dkoes/research/motifs`.

References

[1] J. Berg and M. Lassig. Local graph alignment and motif search in biological networks. Technical report, Institut fur Theoretische Physik, Universitat zu Koln, 2003.

[2] Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M. and Sakaki, Y. A comprehensive two-

hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci*, 98:4569–4574, 2001.

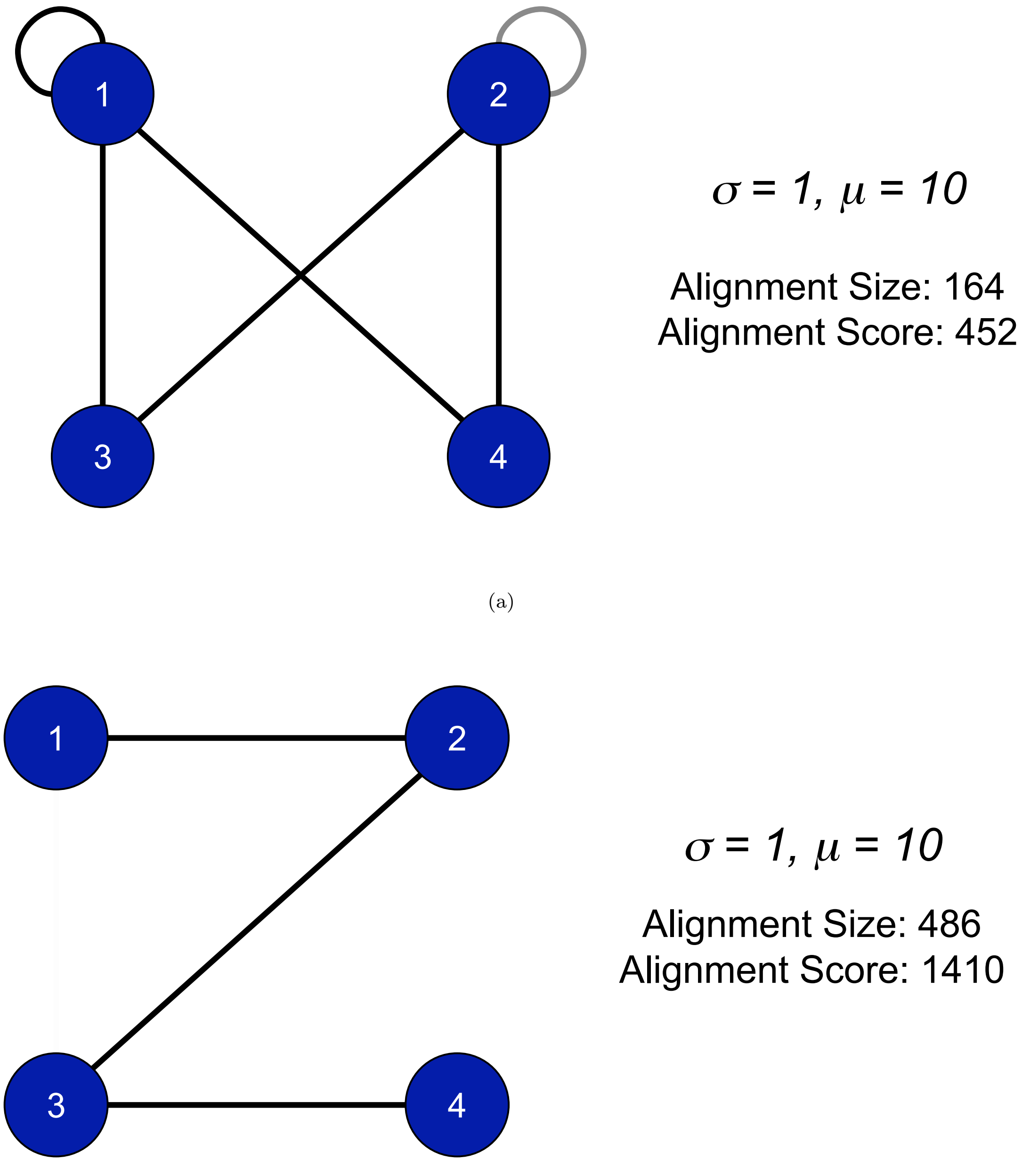
[3] Ito, T., Tashiro, K., Muta, S., Ozawa, R., Chiba, T., Nishizawa, M., Yamamoto, K., Kuhara, S. and Sakaki, Y. Toward a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. *Proc. Natl. Acad. Sci.*, 97:1143–1147, 2000.

[4] D.J. Lockhart and E.A. Winzeler. Genomics, gene expression and dna arrays. *Nature*, 405:827–836, 2000.

[5] Sergei Maslov and Kim Sneppen. Specificity and Stability in Topology of Protein Networks. *Science*, 296(5569):910–913, 2002.

[6] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network Motifs: Simple Building Blocks of Complex Networks. *Science*, 298(5594):824–827, 2002.

[7] Amy Hin Yan Tong, Becky Drees, Giuliano Nardelli, Gary D. Bader, Barbara Brannetti, Luisa Castagnoli, Marie Evangelista, Silvia Ferracuti, Bryce Nelson, Serena Paoluzi, Michele Quondam, Adriana Zucconi, Christopher W. V. Hogue, Stanley Fields, Charles Boone, and Gianni Cesareni. A Combined Experimental and Computational Strategy to Define Protein Interaction Networks for Peptide Recognition Modules. *Science*, 295(5553):321–324, 2002.



(a)

(b)

Figure 4: Motifs of size four. Two different techniques were used to limit the initial number of subgraphs. For Figure 4(a), only those subgraphs with all nodes having degree greater than 1 (very connected subgraphs) were used. For Figure 4(b), just the 1000 subgraphs with the lowest probability relative to the null ensemble.

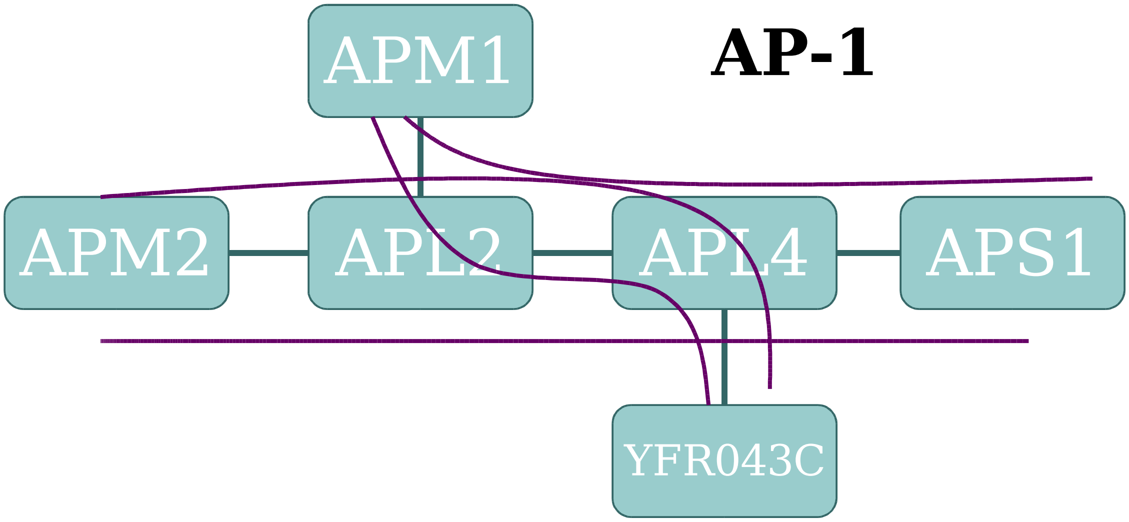


Figure 5: The clatherin-associated protein complex AP-1. The protein YFR043C has an unknown function. The lines represent members of the chain-like motif of size four that, when merged, form AP1.

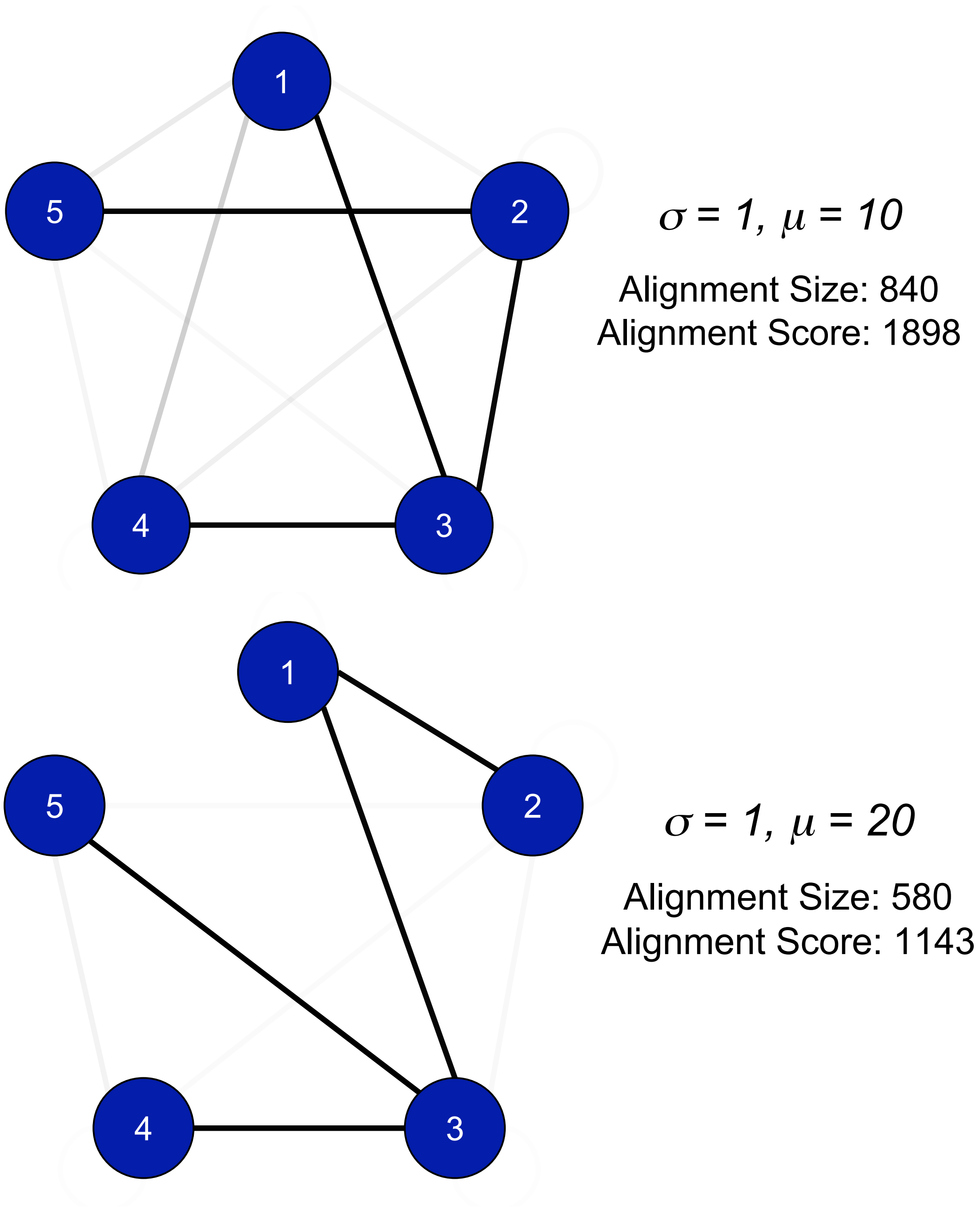


Figure 6: Motif of size five found using two different values of μ. The number of initial subgraphs was restricted to just the 2000 subgraphs with the lowest probability relative to the null ensemble.