

1. Protein sequences are composed of 20 amino acids whereas DNA sequences are comprised of 4 nucleotides. Therefore, it is more computationally involved when we work with protein sequences.
2. For co-regulated DNA sequences, it is assumed that the given DNA sequences contain over-represented subsequences of a similar pattern (motif). However, we can only identify the pair of similar patterns (motif pair) if the set of relevant protein-protein interactions are isolated. This might not be easy as there are many possible subsets of interactions. Even when the set of n interactions can be isolated, there are still 2^{n-1} ways of grouping the protein sequences to identify the motif pair.
3. Usually there is more information available for discovering DNA motif since, for example, the sequences without the binding sites (control set) can provide extra information for solving the problem. However, missing interactions between two protein sequences in the database do not imply the non-existence of motif pairs because the missing protein-protein interaction data might be due to the lack of experiments between these pair of proteins.

A naive approach is to fix a particular protein and the group of proteins that are known to bind to this protein. Then, identify the motif from the group of protein sequences. This can be done using standard motif discovery tools such as MEME [1] or Weeder [8]. And then also find a motif pattern that can uniquely identify the particular protein to form the motif pair. However, this method works only when the number of protein sequences that bind to the same protein is large, say > 4 . In practice, it is usually not the case. In fact, even when a particular protein can bind to a group of many proteins, those bindings might be due to many factors, not just a single motif pair. The problem of finding a motif pair is then reduced to finding a sequence pattern that can uniquely identify that particular protein and also an over-represented sequence pattern in a *subset* (not necessary all) of proteins in the group. Since that particular protein might be uniquely identified by more than one pattern and there can be many subsets of proteins in the group whose sequences have similar patterns, the number of possible motif pairs

can be many. So, it is impossible to determine the correct motif pair, if it exists, which initiates the interactions.

To handle the above problem, [9] proposed to take advantage of prior knowledge of protein groupings according to protein domains. Instead of considering the bindings between a particular sequence and a group of proteins, they consider the bindings between a group of protein sequences and another group of sequences with a particular domain so as to increase the number of sequences in the instance. A modified Gibbs sampling algorithm was developed to identify the motif pair. This method will work if we already know one of the motifs in the motif pair; otherwise, how to isolate two groups of proteins that are related to the same motif pair from the interaction database is non-trivial.

Recently, Tan *et al.* [10] introduced a method to discover motif pairs without knowledge of motifs participating in the motif pair and without any prior knowledge on the protein groupings. The basic idea of their approach is as follows. Based on the input sequences, they generated all possible substring pairs of a certain length from any two interacting sequences. For each possible motif pair, they identified the two groups of proteins that contain an instance of the motifs. They compare the number of observed interactions between these two groups and the expected number of interactions using χ^2 testing. The motif pairs with the highest χ -score (implying the observed number of interactions is much larger than the expected number) were reported. They developed two algorithms, D-MOTIF and D-STAR, for discovering binding motif pairs based on this idea. D-MOTIF can discover the motif pair with the highest χ -score while D-STAR is a heuristics algorithm.

There are several problems with this approach. We found out that the χ -score is not an adequate measure for motif pairs. Since the expected number of interactions decreases with the number of sequences that contain the two motifs, algorithms using χ^2 testing tend to discover binding motif pairs that occur only in a few sequences. For example, when there is only one sequence with motif M_1 and only one sequence with motif M_2 , if there is an interaction between these two sequences, the χ -score will be high. However, M_1 and M_2 are not statistically significant as they occur in one sequence

