Mining Protein Sequences for Motifs

Giri Narasimhan*
Department of Mathematical Sciences,
The University of Memphis,
Memphis, TN 38152.

Changsong Bu
Idax Inc., 5301 Robinhood Rd.,
Norfolk, VA 23669.

Yuan Gao
IBM T. J. Watson Research Center,
Yorktown Heights, NY 10598.

Xuning Wang
Parke-Davis Pharmaceutical Research,
2800 Plymouth Road,
Ann Arbor, MI 48105.

Ning Xu
Department of Mathematical Sciences,
The University of Memphis,
Memphis, TN 38152.

Kalai Mathee
Department of Biological Sciences,
Florida International University,
Miami, FL 33199.

September 12, 2001

Abstract

We use methods from Data Mining and Knowledge Discovery to design an algorithm for detecting motifs in protein sequences. The algorithm assumes that a motif is constituted by the presence of a "good" combination of residues in appropriate locations of the motif. The algorithm attempts to compile such good combinations into a "pattern dictionary" by processing an aligned training set of protein sequences. The dictionary is subsequently used to detect motifs in new protein sequences. Statistical significance of the detection results are ensured by statistically determining the various parameters of the algorithm.

Based on this approach, we have implemented a program called GYM. *The* Helix-Turn-Helix Motif *was used as a model system on which to test our program. The program was also extended to detect* Homeodomain motifs. *The detection results for the two motifs compare favorably with existing programs. In addition, the* GYM *program provides a lot of useful information about a given protein sequence.*

*CORRESPONDING AUTHOR; CURRENT ADDRESS: ECS 389, School of Computer Science, Florida International University, Miami, FL 33199. EMAIL: giri@fiu.edu; TEL: (305) 348-3748; FAX: (305) 348-3549

1 Introduction

A *Motif* is a region or portion of a protein sequence that has a specific structure and is functionally significant. Protein families are often characterized by one or more such motifs. Detection of motifs in proteins is an important problem since motifs carry out and regulate various functions, and the presence of specific motifs may help classify a protein.

We describe a new approach to the problem of automatic motif detection. We use methods from Data Mining and Knowledge Discovery to design an algorithm that displays increased sensitivity as compared to existing algorithms, while maintaining good accuracy and also providing additional information about a given protein sequence. Unlike previous approaches, our algorithm is not based on statistical methods. However, our algorithm does need a "training set" of aligned sample motifs. The basic assumption is that specific combinations (of which there could be many) of residues in specific locations within the motif are responsible for imparting the structure and the functionality to the motif. With this in mind, our algorithm searches for patterns from the sample training set that are present in a new protein sequence. Our approach has similarities to a recent independently developed method (Rigoutsos & Floratos 1998), which does arbitrary motif detection on unaligned protein sequences. This paper describes how our algorithm can be implemented efficiently. The resulting program is called GYM. Finally, we describe our experiments with the *Helix-Turn-Helix Motif*, which was used as a model system on which to test our program. Results of tests on *Homeodomain motifs* are also reported.

2 Motifs in Protein Sequences

As mentioned above, motifs share a common structure and function. We describe the structural and functional properties of the two motifs used in this study.

Helix-turn-helix Motifs The helix-turn-helix motif was the first protein motif to be discovered for site-specific DNA recognition. This motif has been widely investigated and there exists substantial knowledge of the chemical interactions of specific residues. Crystal structures of many of the proteins containing these motifs are also available. Limited information is also available from mutational analysis of the motif and the effect of specific amino-acid substitutions on motif structure. This motif is common to many DNA-binding proteins and plays a crucial role in their binding to DNA. Thus studying these motifs provides an excellent model system for the study of protein motifs that are used in site-specific recognition.

Features of the helix-turn-helix motif have been reviewed extensively (Pabo & Sauer 1992; Brennan & Matthews 1989; Harrison & Aggarwal 1990; Nelson 1995). Briefly, it consists of two α-helical structures separated by a turn. The motif is about 20 to 22 residues in length. The turn consists of three to four amino acids, and the two helices make an angle of approximately $120°$ (Nelson 1995). One of the two helices is responsible for binding to DNA in a sequence-specific manner, and is referred to as the "recognition helix". In most proteins, the second of the two helices is the "recognition helix". Residues of the recognition helix interact directly with bases in the major groove of the DNA. A combination of residues in both the helices are believed to be responsible for maintaining the appropriate angle between the two helices. Proteins with helix-turn-helix motifs share only limited sequence homology in the motif region; the dissimilarity is attributed to the sequence-specific interactions with the bases in the DNA. Most proteins have at most one helix-turn-helix motif; however, we will discuss a family of proteins that have more than one such motif. All the properties mentioned above make automatic recognition of helix-turn-helix motifs a non-trivial algorithmic problem.

Homeodomain Motifs Proteins containing the homeodomain motif play an important role in plant and animal development. These proteins are also DNA binding proteins. The homeodomain (Scott, Tamkun & Hartzell 1989) motif is made up of three α-helices and an extended N-terminal arm. The first and second α-helices pack against each other in an anti-parallel arrangement, while the third α-helix lies perpendicular to them. The third helix is the *recognition helix*; like its counterpart in the helix-turn-helix motif, it interacts with DNA in the major groove and provides the DNA-binding specificity. However, unlike the helix-turn-helix unit, the 60-residue homeodomain forms an independent folded structure and can independently bind to DNA. It is interesting to note that the homeodomain motif contains a canonical helix-turn-helix structure. Mutational and evolutionary analyses and crystal structures of these domains are also available. The homeodomain motifs were chosen for testing because they are almost thrice as long as the helix-turn-helix motifs, giving us the opportunity to test the GYM program on a different motif.

3 Motif Detection

3.1 Existing Methods

To aid in the detection of motifs in protein sequences, classical methods involved carefully crafting a "consensus" sequence to reflect highly conserved residues in the motif. Pabo and Sauer (1992) constructed a consensus sequence for helix-turn-helix motifs based on a multiple alignment of known motif sequences. One simple method to detect helix-turn-helix motifs is to look for the occurrence of such a consensus sequence. More general "consensus" sequences are also possible (as the ones maintained by PROSITE (Bairoch 1992)). Inspired by the concept of regular expressions, such generalized consensus sequences consist of a sequence of sets of amino acids where amino acids within the same set could substitute each other in that position. Nevill-Manning et al. (Nevill-Manning, Wu & Brutlag 1998) presented another method for discovering motifs from families of aligned protein sequences based on automatically constructing such generalized consensus sequences. Earlier, Wu and Brutlag (1996) had showed a way of constructing substitution sets in a statistically significant manner. Another feature of the algorithm due to Nevill-Manning et al. (1998) involved constructing a set of such consensus sequences with the assumption that any one of the sequences in the set could describe the motif. Additionally, the software based on their method (EMOTIF) provided the user with parameters to trade-off sensitivity to specificity. These parameters are claimed to control the number of false positives.

Other sophisticated detection schemes are all statistically motivated. These are typified by the *Profile* method described by Gribskov et al. (1990). The first step, once again, involved making a multiple alignment of known motifs. The next step typically involved computing a probability matrix or a *Score Matrix*, which assigns a different score to each possible residue at each position in the motif. Intuitively, the entries of this matrix represent a measure of the probability that a certain residue occurs in that location, normalized by the background frequencies for that residue. Minor variants exist in the methods employed to compute the position-specific scoring matrix as well as in scoring a match (see, for example, (Gusfield 1997)). Given a score matrix, the detector, when given an input protein sequence, computes a weighted score for every subsequence of the input sequence, and reports the subsequence with the highest score as the detected motif, as long as this score is above a certain threshold.

Dodd and Egan (1990) showed how to compute such a matrix in a simple manner. The probability values they used were simply the frequency of a residue normalized against the background frequencies for that residue. Enhancements on the method were used to detect coiled-coil motifs (Berger, Wilson, Wolf, Tonchev, Milla & Kim 1995; Berger 1995; Wolf, Kim & Berger 1997). Variants of the profile method where a randomized learning algorithm iteratively improves the generated profiles were also used in detecting coiled-coil motifs (Berger & Singh 1997; Singh, Berger & Kim 1998). Other (substantially different) statistically-based methods for detecting motifs include that of using a Hidden Markov Model (Grundy, Bailey, Elkan & Baker 1997) and Gibbs Sampling (Lawrence, Altschul, Boguski, Liu, Neuwald & Wootton 1993).

3.2 Motif Detection – New Method

Our method is not based on statistical methods. It discovers patterns in known motifs to compute a pattern dictionary. Detection of a motif in a new protein sequence is then a function of which patterns from the dictionary are present in the new protein sequence.

The first assumption that our algorithm makes is that an appropriate length of the motif is known beforehand. This is a reasonable assumption to make since this is true for most of the known motifs (as in the case of helix-turn-helix motifs and homeodomain motifs). For example, there is ample evidence to show that a helix-turn-helix motif lies within a window of size at most 22. The second assumption is that a reasonably large number of motifs are known and have been detected and verified by experiment in the standard way. The training set can be chosen from these known motifs. The third assumption is that a **combination** of key residues are sufficient to constitute the necessary physical structure and to give it the functionality; the rest of the parts of the motif may serve other purposes. Note that this is where we differ from the assumptions made for the other methods. While many of the methods attach separate significance to the occurrence of specific residues in specific locations in the motif, they do not account for the reinforcing effect of a combination of specific residues. For instance, residue x in location l_x may be very significant only if residue y is in location l_y and residue z in location l_z simultaneously. Residue x in location l_x may not be very significant otherwise and may not occur frequently in that location in known motifs and thus may not have a high score in the score matrix. It is likely that the patterns in the pattern dictionary discovered by our method represent such "reinforcing" combinations, helping in the detection of new motifs. Finally, we assume that a "good" combination of residues must occur "frequently enough" to be called a valid pattern for the motif. To account for relatively rare reinforcing combinations, we opted for setting an absolute threshold value to decide whether a combination occurs "frequently enough", as opposed to a requirement that a combination occurs in a specified percentage of the sequences in the training set.

Rigoutsos and Floratos (1998) independently devised a method to discover unknown motifs without doing alignment, i.e., the training set for their program is a set of unaligned protein sequences. Their method is based on similar ideas of generating patterns, which in turn could be used to perform detection. Other related methods are reviewed by Brāzma et al. (1995). While the overall philosophy of these methods coincide with ours, our algorithm differs from them in that it detects known motifs after being trained on a set of aligned sequences for the same motif, thus making use of all available knowledge about the motif. Our methods also share some overlap with that of Nevill-Manning et al. (1998). The fundamental differences lies in the way the threshold (this concept is explained below in Section 4.1) is used; they require that their motifs "cover" some percentage of the sequences in the training set. The idea of correlations between residues in specific locations were also explored by Berger et al. (1995); while significant correlations were observed, only pairwise correlations were considered.

4 The New Algorithm

Here we present our new algorithm for detecting known motifs in protein sequences. We will refer to this algorithm as the "Pattern Dictionary" method. Based on this new algorithm, we have implemented a program (called "GYM") that detects helix-turn-helix motifs. A second version of GYM (Version 2.0), incorporating various improvements was also implemented. Unless explicitly specified, in the rest of the paper when we refer to the GYM program, we will be referring to both the versions. GYM was also modified and retrained to detect homeodomain motifs in protein sequences. Note that the program can be modified to detect other motifs.

The algorithm requires that an approximate length of the motif be known beforehand and that a reasonably large number of motifs are known and have been detected and verified by experiment in the standard way. The training set can be chosen from these known motifs. The algorithm consists of two parts. The first part is a preprocessing step that needs to be performed only once. The second part is where the actual motif detection takes place.

The preprocessing phase can be called the **Pattern Mining** phase. The input to this phase is the set of known and aligned motifs, or the *Master Set*. The output is a *Pattern Dictionary* consisting of frequently occurring *Patterns* within the Master Set. The preprocessing phase is described in detail in Section 4.1. The input to the second part, or the **Detection** phase, consists of the pattern dictionary output from the preprocessing step and the input protein for which the motif detection needs to be performed. The detection phase is described in Section 4.2.

The output of the detection algorithm will indicate whether or not the protein sequence contains a motif, the location of this motif, a score indicating the confidence of the prediction, along with a list of proteins from the master set that share high sequence homology with the detected motif as inferred from matching patterns from the dictionary.

4.1 Preprocessing: "Pattern Mining"

The input to this phase is a master set of aligned motifs without spaces. Thus when two motifs are aligned, either the amino acids in a certain location in the motif match or have a mismatch. Figure 1 shows an example of a set of aligned motif sequences, where each motif is of length 7; this is a hypothetical motif that is simply used to illustrate the method. Note that each of these motif sequences occur at different locations in different proteins as indicated in Figure 1.

Location in Seq.	Sequence 1 2 3 4 5 6 7	Protein Name
14	`G V S A S A V`	*Ka* RbtR
32	`G V S E M T I`	*Ec* DeoR
33	`G V S P G T I`	*Ec* RpoD
76	`G A G I A T I`	*Ec* TrpR
178	`G C S R E T V`	*Ec* CAP
205	`C L S P S R L`	*Ec* AraC
210	`C L S P S R L`	*St* AraC

Figure 1: **Aligned Motifs – An example**

Every amino acid in each of the hypothetical motif sequences in Figure 1 is associated with a position in the motif. Thus protein *Ka* RbtR has amino acid `G` in location 1, `V` in location 2, `S` in location 3, and so on. We thus represent the motif by a sequence of pairs, where each pair $<aa, pos>$ consists of an amino acid and its position in the motif. We simplify the notation and denote these amino acids by pairs of symbols such as `G1`, `V2`, and `S3`, respectively. *Ka* RbtR would thus be denoted by the set of pairs {`G1, V2, S3, A4, S5, A6, V7`}.

A **Pattern** is simply a set of pairs. Thus, {`G1, S3, T6`} and {`C1, P4, S5, L7`} are two examples of patterns. Protein *Ec* DeoR contains the pattern {`G1, S3, T6`}, but does not contain the pattern {`C1, P4, S5, L7`}. The *length* of a pattern is defined as the number of pairs in it. Thus, {`G1, S3, T6`} is a pattern of length 3; this pattern is also shared by protein *Ec* CAP and *Ec* RpoD. The *support* of a pattern is the number of proteins in which it appears. For the 7 motif sequences in Figure 1, the patterns {`G1, S3, T6`} and {`G1, S3, V2`} have a support of 3, while pattern {`G1, S3`} has a support of 4. A pattern is called a *significant* pattern (or a frequent pattern), if its support is no less than a certain *threshold*. A significant pattern is called *maximal* if it is not contained in any other significant pattern. For a threshold value of 3, the pattern {`G1, S3, T6`} is significant, but not maximal, since the maximal significant pattern {`G1, I7, S3, T6, V2`} contains it. The pattern mining phase outputs a list of all maximal frequent patterns. This list will henceforth be referred to as the *pattern dictionary*. For a threshold value of 3, the dictionary that will be output for the example in Figure 1 would be that in Figure 2(a). If we lower the threshold to 2, the dictionary that will be output would be as shown in Figure 2(b).

Pattern length	Maximal Patterns	Support
2	{ `S3, P4`}	3
2	{ `S3, S5`}	3
3	{ `G1, T6, I7`}	3
3	{ `G1, S3, T6`}	3
3	{ `G1, V2, S3`}	3

(a) THRESHOLD = 3

Pattern length	Maximal Patterns	Support
2	{ `S3, P4` }	3
3	{ `G1, S3, V7`}	2
5	{ `G1, V2, S3, T6, I7` }	2
7	{ `C1, L2, S3, P4, S5, R6, L7` }	2

(b) THRESHOLD = 2

Figure 2: **Frequent patterns for aligned motifs from Figure 1 with different THRESHOLD values.**

Algorithm *Pattern-Mining*
Input : Motif length m, support threshold T, and list of aligned motifs.
Output : Dictionary L of frequent patterns.
1. Generate all frequent patterns of length 1 and insert into list L_1.
2. **for** $i = 2$ to m **do**
3.   **for** every pair of patterns $p, q \in L_{i-1}$ such that $|p \cap q| = i - 1$ **do**
4.   Insert pattern $p \cup q$ into list E_i
5.   **for** every pattern $p \in E_i$ **do**
6.   **if** $(support(p) > T)$ **then**
7.   Insert p into L_i.
8.   Remove all subsets of p from L_{i-1}.
9.   **if** $(|L_i| \leq 1)$ **then**
10.   **return** $L = \cup_i L_i$

Figure 3: **Pattern Mining Algorithm**

The algorithm goes through at most m (length of the motif) iterations. In the i-th iteration, it generates all frequent patterns of length i. In the i-th iteration, the algorithm first generates a collection of potentially frequent patterns and then their supports are computed to verify if they are frequent enough. The collection of potential patterns generated in the i-th iteration consists only of those patterns that are obtained by the set union of two frequent patterns of length $i - 1$ that differ in exactly one item. Even this observation is not enough to efficiently generate all potentially frequent patterns. We first present the algorithm in Figure 3 followed by a discussion of the implementation details that make it efficient.

Implementation Details Enumerative algorithms that generate such a dictionary of frequent patterns are likely to be inefficient because of the combinatorial explosion in the number of possible patterns. A simple counting argument shows that if the *motif length* is denoted by m, there are $20m$ possible patterns of length one (since there are only 20 amino acids), and $O((20m)^k)$ possible patterns of length k. However, it was recently shown (Parida, Rigoutsos, Floratos, Platt & Gao 2000) that the number of frequent or significant patterns is also bounded by $O(mn)$, where n is the number of motif sequences in the input.

The naive algorithm of generating all possible patterns of length i and checking whether it is significant or not will clearly be very inefficient. Our algorithm is based on an algorithm from the Data Mining by Agrawal et al. (1996), and is able to avoid the generation of most infrequent patterns by using an efficient screening process to be described below. The basic idea behind the algorithm by Agrawal et al. is that if a pattern occurs frequently, then every subset of this pattern must necessarily be frequent. This also implies that if a pattern does not occur frequently enough, then all supersets of this pattern may be immediately discarded.

In order to make this process efficient, all patterns are stored in a canonical form with their items in sorted order. For example, the canonical form for a pattern {`G1, A2, I4, A5`} would be the sequence {`A2, A5, G1, I4`}. The sorting is done in a simple lexicographic manner. Furthermore, the list of patterns are also sorted (again, in a lexicographic order). Once the dictionary is put in this form, a potentially frequent pattern is generated if two frequent patterns of length $i-1$ share the first $i-2$ items in common. But then, since the list is in sorted order, such patterns are going to be next (or at least close) to each other in the list. To be more precise, a block of k patterns that share the first $i-2$ items are going to contribute $O(k^2)$ potentially frequent patterns of length i. Another important observation that is not obvious is that the list does not need to be sorted if things are processed in a proper order. When a new pattern is generated, it is automatically put at the end of the list. Because of the order in which things are considered, the list remains sorted. For the first iteration, the input sequences are scanned and for every amino acid in the input sequences, either a new pattern of length one is created or the support of an existing pattern is incremented. The implementation resulting from the algorithm is very fast in practice.

4.2 The Detection Algorithm

The detection algorithm is now quite straightforward and is described in Figure 4. It takes as input a motif length m, the dictionary of significant patterns L output by the *Pattern-Mining* algorithm (Figure 3), an integer k representing the number of best matches required as output, and the given protein sequence P to be examined for the motif. We slide a window of length m across the input sequence P. The subsequence of P that lies in the window is then matched against every significant pattern in L. This is performed in a subroutine called Match. Match returns a *Match-Score* that quantifies how well the window matched against the patterns in the dictionary

L. While it is convenient to think of Match-Score as a number, as explained later, it is in reality a collection of measures that describe the quality of the match. The k best Match-Scores along with the corresponding window locations is maintained. Finally, all matches from the list of the k best matches whose quality exceeds a pre-specified threshold are reported as possible motif locations.

Algorithm *Motif-Detection*
Input : Motif length m, threshold score T,
dictionary L of patterns,
number of best matches k,
and input protein sequence $P[1..n]$.
Output : Information about motif(s) detected.
1. Best-Match-Score = Match($P[1..m], L$)
2. **for** $i = 2$ to $n - m + 1$ **do**
3. Match-Score = Match($P[i..i + m - 1], L$)
4. Update list of k best matches found so far.
5. **for** $i = 1$ to k **do**
6. **if** i-th best match-score exceeds T **then**
7. Report it as possible motif location

Figure 4: **Motif Detection Algorithm**

Match Parameters In our first implementation of GYM (Version 1.0) (Gao, Mathee, Narasimhan & Wang 1999) the following were the parameters that defined the quality of a match of a window of size m with a dictionary of proteins: **LPM** – the length of the longest pattern matched, **NDP** – the number of distinct positions from the window that matched some pattern, and **NPM** – the number of distinct maximal patterns matched.

In the second implementation of GYM (Version 2.0), the parameter **NDP** was replaced by a weighted version called **WNDP**. The weighted score uses the widely used BLOSUM62 matrix (Henikoff & Henikoff 1992). WNDP is computed by taking the BLOSUM62 matrix score for substituting each amino acid in the pattern by itself. The use of a weighted score is a natural way to weight patterns in the dictionary, and helped to increase the sensitivity of the scoring scheme used.

Comparing two matches The selection of the list of parameters and the process of comparing two matches has been developed and fine-tuned after a close study of experimental data. A significant pattern from the dictionary represents a combination of amino acids in specific locations that (potentially) positively reinforce the motif structure. Thus, the longer the pattern, the greater the number of positive reinforcements to the structure, and consequently, the better the quality of the match. By a similar argument, NDP and NPM are also significant.

The Match Procedure In GYM 1.0, a match is said to exceed the threshold (i.e., a HTH motif is detected) if one of the following is true: (a) LPM is at least 5, OR (b) LPM equals 4, and NPM is at least 2, OR (c) LPM equals 4, NP equals 2, and NPM is at least 6. In GYM 2.0, a match is said to exceed the threshold if both the following are true: (a) LPM is at least 4, AND (b) WNDP is at least 29. The choice of the threshold values in GYM 2.0 is discussed later in Section 5.

Output of Algorithm The algorithm is designed to output the k locations with the highest match score, as long as the scores are above a pre-specified threshold. The algorithm will output the location of the motif as well as the residues in the motif. It also prints the match parameters, i.e., LPM, NDP, and NPM, for the k best matches. The output includes the set of patterns that are matched at the predicted motif location. It also indicates which particular residues in the motif are present in the patterns matched.

Lastly, the output gives a list of proteins from the training set that exhibit the same patterns found in the motif of the input protein sequence. For new protein sequences, this could provide clues to the family to which this protein may belong in terms of its function. This information would become more valuable when it is combined with similar information from other motifs found in the same protein sequence. We also conjecture that a more careful study of the list of patterns matched in a protein could help in determining evolutionary relationships between proteins.

Here we provide some justification for the information output by the algorithm. Clearly, the algorithm is based on detecting the patterns from the dictionary that are present in the given protein sequence. It thus makes sense to output the set of patterns present. However, also printing out the list of proteins from the training set that exhibit the same pattern needs to be justified. We note that a pattern that is generated by GYM during the pattern mining phase is likely to have structural significance. To support this claim, we point to the recent work of Rigoutsos et al. (1999). They noted that if a pattern generated by a data mining algorithm appears in the motif region of several protein sequences, it is often true that the corresponding three-dimensional substructures defined by the amino acids of that pattern in each of the proteins are also identical, i.e., the substructures corresponding to the amino acids in the pattern tended to align almost precisely in three-dimensions. Three-dimensional alignment was demonstrated in (Rigoutsos, Gao, Floratos & Parida 1999) by showing that the root mean square (RMS) error between substructures involving elements of the pattern is small (i.e., less than 1.5 Angstroms). They called these substructures as "3-D motifs".

5 Testing the Implementation

The pattern dictionary algorithm described above was implemented, trained, and tested. We ran two independent sets of experiments on the protein sequences. One set was run with the assumption that the helix-turn-helix motif was located within a window of 20 residues; the other set was run with a window size of 22. Since the differences were minor, we only present the results for a window size of 22. On the whole, the results for a window size of 20 were subsumed by the results for a window size of 22, i.e., sometimes the results for a window size of 22 may contain a prediction that is missed by the results for a window size of 20.

Choice of Master Set We first discuss the choice of the training set, also referred to as the "Master Set". In general, the choice of a training set is a non-trivial problem, and could determine the success or failure of a motif detection method. Also, automatic generation of a training set is a difficult problem. We initially used the same training set (91 proteins) as Dodd and Egan (1990), but subsequently eliminated three proteins. For these three proteins the GYM and DE programs reported different motif locations, and experimental evidence defining the precise locations was not available in the literature. We deleted these proteins so that their motifs could not bias the training set. The three proteins deleted were: (a) SpoOA *Bacillus subtillis* (Assn. No: 134739): GYM predicted positions 5 and 219, DE predicted 198; (b) XylR *Bacillus subtillis* (Assn. No: 98448): GYM predicted position 361, DE predicted 29; (c) pSC101 rep (Assn. No: 281929): GYM predicted a marginal motif at location 103, while DE failed to predict any motif.

Choice of Threshold Values Next we discuss the choice of various threshold values used by GYM. The support threshold T in the *Pattern-Mining* algorithm represents a tradeoff between the sensitivity and the number of false positives that can be generated by the detection algorithm. When the threshold is higher, while fewer patterns are generated, the resulting patterns are likely to have higher statistical significance. On the other hand, if the threshold is lower, sensitivity is higher since good patterns with lower statistical significance can be detected. The threshold value used for our experiments was equal to the maximum value that optimized the detection of motifs from the training set itself, i.e., if the threshold is any higher, then the detection algorithm failed on some instances from the Master Set itself. For the given training set, we chose a threshold of 4 (in both GYM 1.0 and 2.0).

For the pattern detection algorithm, the minimum threshold for LPM (the length of the longest pattern) was chosen to be 4. This choice was motivated by the result of the following experiment: the LPM value for a randomly permuted protein sequence or a randomly permuted motif sequence was never larger than 3. Furthermore, all sequences from the master set had patterns of length at least 4. This suggests that LPM of 4 is natural choice for this threshold and is also statistically significant.

Also, in GYM 2.0, the threshold on WNPD (the weighted score) was chosen to be 29. The minimum WNPD score for all the (positive) sequences tested was 19. In direct contrast, the maximum WNDP score from the Negates set and from a large number of randomly permuted motif sequences (permuted from the positive sequences) turned out to be 37. Thus a clear choice of the threshold was not forthcoming from such an experiment. The following procedure was then used to decide on a choice of 29 for the threshold: for every choice of threshold value in the range 19 through 37, we plotted the number of false positives and (overall) false negatives. As was expected, as the threshold value was increased, the number of false positives increased and the number of false negatives decreased. A threshold value of 29 represents the best balance of the conflicting interests.

Finally, we chose to output only the two (i.e., $k = 2$) best locations detected by GYM; choosing $k = 1$ was inadequate because there are a number of proteins with more than one helix-turn-helix motif.

Choice of Test Set Since we wanted to test the performance of the GYM program on a diverse set of proteins, the GYM program was then tested on several families of proteins. Some of the sub-families (such as the SigE sub-family) are not represented in the training set. The sequences were downloaded from GenBank Protein Sequence Database maintained by the National Center for Biotechnology Information (NCBI). For the helix-turn-helix motif, we ran the GYM 1.0 program on 675 protein sequences and GYM 2.0 on 721 sequences.

Of the input sequences, GenBank had the motif location annotated in only about 36% of the cases. However, the database had no information on how these locations were determined (What program was used to determine the motif location? Were any laboratory experiments performed to verify the claim?). Thus we consider these annotations unverified, as are the predictions of our program. In order to confirm our results with an independent program, we also implemented the score matrix method described by Dodd and Egan (1990). For the sake of convenience, we refer to this program as DE, and we refer to our program as GYM. Among the proteins selected, 93 are proteins involved in metabolic pathways and other enzymatic reactions. These are presumed not have a helix-turn-helix motif, since they are unlikely to bind to DNA. We refer to this family as the "Negates" family (see section 6.3).

Overview of Experimental Results The programs GYM 1.0 and DE disagreed on 68 (approximately 10%) of the 675 sequences. Of these disagreements 23 were from the negates set, indicating a large number of false positives for our program. The programs GYM 2.0 and DE disagreed on 61 (approximately 8%) of the 721 sequences. However, the number of disagreements in the negates set were significantly reduced to 7. Among the disagreements in both sets of experiments, 23 of them were from the "Sigma" family (sigma factor proteins). As discussed in section 6.2, there is evidence to support the helix-turn-helix predictions made by our program for these 23 sigma factor sequences.

It is interesting to note that the DE program, by virtue of its design, makes a sharper distinction between its first choice and its second choice in terms of the weighted scores. This is not true of the GYM program. While this was an asset in dealing with proteins that had two helix-turn-helix motifs, it could be considered a drawback in other cases.

The results are discussed in detail in Sections 6.2- 6.4. The scores are first summarized in Figure 5 (experiments with GYM 1.0) and in Figure 6 (experiments with GYM 2.0). The first column specifies the family of proteins tested. The next two columns state the number of sequences tested and the number on which the two programs agreed on a motif location. The two programs are said to agree if they have a common location within their top two choices and if the corresponding scores are above the threshold. The next two columns indicate how many of the sequences tested had published annotations for the motif location and how many of these matched with GYM's predictions. Once again, an annotation is matched if one of the top two locations (if above the threshold) are the same. While this was not known before testing, we noticed that of the 675 proteins analyzed in the first set of experiments, only 459 of the detected motifs were unique, i.e., 216 of the sequences had motifs that were identical to the motifs in other sequences. Even if these are discounted, the overall results show that out of 459 ($= 675 - 216$) sequences tested, the two programs agreed on $630 - 216 = 414$ (about 90%) sequences.

We also modified the GYM 1.0 program to detect homeodomain motifs. An appropriately modified version of the DE program was also used. After training the two programs with 121 sequences, we ran the two programs on 524 protein sequences. There was overwhelming agreement between the two programs. The results are summarized in Figures 5 and 6.

Motif	Protein Family	How Many Tested	GYM = DE Agree	How Many Annotated	GYM = Annotated
Helix-Turn-Helix Motif (Window Size = 22)	Master	88	88 (100%)	13	13
	Sigma	304	270+23 (96%)	96	82
	Negates	93	70 (75%)	0	0
	LysR	127	125 (98%)	95	93
	AraC	63	54 (86%)	37	29
	Total	675	607+23 (93%)	241	217 (90%)
Homeodomain Motif (Window Size = 60)	Master	121	121 (100%)	121	121
	Rest	403	390 (97%)	385	370
	Total	524	511 (98%)	506	491 (97%)

Figure 5: **Summary of Motif Detection Results (GYM 1.0)**

Motif	Protein Family	How Many Tested	GYM = DE Agree	How Many Annotated	GYM = Annotated
Helix-Turn-Helix Motif (Window Size = 22)	Master	88	88 (100%)	13	13
	Sigma	314	284+23 (98%)	96	82
	Negates	93	86 (92%)	0	0
	LysR	130	127 (98%)	95	93
	AraC	68	57 (84%)	41	34
	RReg	116	99 (85%)	57	46
	Total	721	653+23 (94%)	289	255 (88%)

Figure 6: **Summary of Motif Detection Results (GYM 2.0)**

6 Results and Discussion

Now we discuss in detail the results on each individual family of proteins from the test set.

6.1 Master Set

The GYM and DE programs were first tested on the 88 sequences from the Master Set (the same set they were trained with). The two programs agreed on the locations of the motif in all of them.

6.2 Sigma Family

The proteins selected for this set are all sigma factors, which are known to be DNA-binding proteins. The sigma subunit of eubacterial RNA polymerase is required for recognition of promoter sequences and initiation of transcription from those sites (Lonetto, Gribskov & Gross 1992). Two major subfamilies of the sigma family of proteins have been identified: (i) the σ^{70} or RpoD subfamily, which is used by most of the "housekeeping" genes expressed during exponential growth, and (ii) the alternative sigma factor subfamily, including RpoS, RpoE, FliA, etc., which are involved in coordinated expression of sets of genes during a change in metabolic or developmental state.

The results for this set of proteins were quite interesting. Both GYM 1.0 and 2.0 disagreed with the predictions of DE for about 10% of the sequences tested. On closer inspection we found that the predicted locations for 23 of them were about 90-93 residues apart, and that most of them were from the RpoS subfamily. Members of the sigma family are known to have two helix-turn-helix motifs (Lonetto, Gribskov & Gross 1992; Gruber & Bryant 1997) in regions 3.1 and 4.2, which are about 90-93 residues apart. For the RpoS subfamily, the GYM program detected the motif in region 3.1, while DE picked the one in region 4.2, thereby accounting for the disagreements. It is possible that there is difference in the strength of the motifs in the two regions. It is also interesting to note that the motifs in region 3.1 were not represented in the master set but GYM was still able to detect them.

6.3 Negates Family

Among the proteins analyzed in the two sets of experiments, 93 were specifically chosen as proteins involved in metabolic pathways and other enzymatic reactions. We presumed that these proteins are unlikely to have a DNA-binding function, and consequently are unlikely to have a helix-turn-helix motif. We refer to this set of proteins as the "negates" family.

Of these 93 proteins, GYM 1.0 predicted a potential helix-turn-helix motif for 23. We initially interpreted these results as "False Positives"; however, inspection of the crystal structure of one of these proteins (Adenylosuccinate Synthetase from *Escherichia coli*, Assn. No: 1942847) revealed three α-helices at locations 183-191, 193-201, and 204-214. The GYM program predicted a motif at locations 188-219, which includes the last two α-helices. This demonstrates that the GYM program is able to detect structural motifs with high sensitivity, and that a helix-turn-helix structure can occur without guaranteeing DNA-binding function for the motif region.

It is significant that the use of the weighted scores in GYM 2.0 and a careful choice of the corresponding weighted score threshold reduced the "False Positives" to a mere 7%.

6.4 LysR, AraC and RReg families

The proteins from the LysR family (Schell 1993) are predominantly similar-sized, autoregulatory transcriptional regulators. In response to different inducers, members of this family of proteins activate divergent transcription of linked target genes or unlinked regulons encoding extremely diverse functions. Mutational studies and amino acid sequence similarities have identified a DNA-binding domain employing a helix-turn-helix motif (residues 1-65).

The AraC/XylS (Gallegos, Schleif, Bairoch, Hofmann & Ramos 1997) family of transcriptional regulators includes proteins and predicted polypeptides derived from translation of DNA sequences. Members of this family are about 300 amino acids long and have three main regulatory functions: carbon metabolism, stress response, and pathogenesis.The conserved region contains all the elements required to bind DNA target sequences and to activate transcription from cognate promoters.

The RReg family consists of **R**esponse **Reg**ulators of the two-component regulatory system superfamily involved in sensory transduction, with NtrC being the prototypic member. These response regulators are characterized by a conserved N-terminal domain of approximately 125 amino acids, central domain that is involved in interaction with sigma factors and a C-terminal with helix-turn-helix motif (Hoch & Silhavy 1995).

GYM 1.0 was tested only on the AraC and LysR families. GYM 2.0 was tested on all the three families – LysR, AraC and RReg.

The predictions of GYM 1.0 agreed with those from DE much more for the LysR family than the AraC family. Among the disagreements in the LysR family was one (*Sc* BlaA; Assn. No.: 461627) for which the published location matches GYM's strong prediction for a motif at location 17; whereas, neither of the top two DE predictions included this location. The other disagreement in the LysR family was one (*Bs* YofA; Assn. No.: 2634236) for which both programs had marginal scores at different locations.

Among the disagreements in the AraC family, there were several that were displaced by about 60 residues. This raises the question whether there is a second helix-turn-helix structure in that location (as observed in the Sigma family). There is evidence to suggest that this region containing the second motif may have biological significance (Gallegos, Schleif, Bairoch, Hofmann & Ramos 1997). It is also interesting to note that the average "score" given by both programs was relatively low for members of the AraC family, suggesting somewhat different characteristics for the helix-turn-helix motifs in this family.

The performance of GYM 2.0 on RReg was very similar to that on AraC in the sense that its predictions agreed to about 85%. Just as with GYM 1.0, GYM 2.0 agreed with the predictions of DE much more for the LysR family.

6.5 Homeodomain Motif

Both GYM 1.0 and DE programs were retrained with homeodomain motifs (60 residues in length) from 121 proteins to generate a corresponding pattern dictionary. The two programs agreed with each other's predictions and with the database annotations approximately 97% of the time. Such a high percentage of agreement may arise because homeodomain motifs are found in closely related proteins, or because the proportion of amino acids conferring DNA-binding specificity (amino acids that are different) is much smaller than the proportion conferring α-helical structure (amino acids that are similar). We are unable to offer any explanations for the cases where the predictions of the two programs differed. The choice of the threshold for LPM was 8.

7 Further Refinements to GYM

7.1 Refining the Pattern Dictionary

The occurrence of false positives in GYM's results are likely to be the result of the generation of "spurious" patterns in the pattern dictionary. It is possible that a pattern is detected because the proteins are related and not because this pattern reinforces a motif structure. With the goal of reducing the number of false positives even further, the pattern dictionary was carefully inspected. Any pattern that is seen in a protein from the negates set is clearly a spurious pattern. However, it is not possible to simply eliminate such patterns since a subpattern of this pattern may be significant and eliminating it could increase the number of false negatives. In order to address this issue, we replaced every spurious pattern in the pattern dictionary by every subpattern of length 1 less than that of the spurious pattern. Then we incrementally eliminated each spurious pattern to see if it affected the number of false positives and false negatives. After pruning the dictionary using the above procedure, we were able to reduce the number of false positives from 7% to under 4% (without changing the number of false negatives). These changes will be incorporated in GYM Version 3.0 to be released in the near future.

7.2 Mutational Data

Substitutions of amino acids within the motif can diminish or enhance the strength of the motif. The pattern dictionary method gives us the unique ability to predict the effect of such substitutions. By observing the changes in the set of patterns matched (new patterns may be matched, while old matches may be lost), the effect of the substitution can be quantified. There is data available in the literature describing the effect of single amino acid changes within the helix-turn-helix motifs for specific proteins (Pfau, Arvidson & Youderian 1994; Kim, Makino, Amemura, Nakata & Shinagawa

1995; Siegele, Hu, Walter & Gross 1989; Bushman, Shang & Ptashne 1989). Work is underway to analyze such data using the strategy described above.

8 Conclusions

The GYM program has excellent ability to predict helix-turn-helix motifs. It appears to have increased sensitivity over the DE program and can detect motifs with greater differences from the training set. On the negative side, the GYM 1.0 program appears to have a higher number of false positives; this may result from detection of closely placed α-helices that are not directly involved in DNA binding. The improved GYM 2.0 showed significantly lower number of false positives as compared to GYM 1.0. This was achieved by using a score weighted by the use of BLOSUM62 substitution matrix, and by careful choices of threshold values resulting from sound statistical experiments. Modification of the GYM program to detect the longer homeodomain motifs was also successful and resulted in very high agreement with DE and the database sequence annotations.

The previous statistical methods for motif detection are limited in that the sample training set must contain sufficient representation of amino acid substitutions that preserve or reinforce the particular structure and function of the motif. Some combinations of residues in specific locations can reproduce that structure, whereas others cannot. Until we can model these complex molecular interactions, the next best thing will be to detect and enumerate successful combinations of residues that form such motifs.

When dealing with specific motifs, mining for patterns from aligned sequences is likely to perform better than mining from unaligned sequences since all available knowledge about the motif is being utilized.

Using GYM via the Internet GYM 2.0 is now available for use at

`http://www.msci.memphis.edu/~giri/GYM2/welcome.html`

Acknowledgments We thank Firasath Ali, Mu Yang, and Xingqiang Wang for help with earlier implementations. We thank Yi De for helpful discussions. We are very grateful to Dr. Martha Howe for a careful reading of the paper, for many critical comments, and for help with interpreting our results. The work of GN was partially funded by NSF grant CCR-940-9752 and by a grant from Cadence Design Systems, Inc.

References

Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., & Verkamo, A. I. (1996). Fast discovery of association rules. In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, & R. Uthurusamy (Eds.), *Advances in Knowledge Discovery and Data Mining* chapter 12, (pp. 307–328). MIT Press.

Bairoch, A. (1992). PROSITE: a dictionary of sites and patterns in proteins. *Nucleic Acids Research*, *20*, 2013–18.

Berger, B. (1995). Algorithms for protein structural motif recognition. *Journal of Computational Biology*, *2*, 125–138.

Berger, B. & Singh, M. (1997). An iterative method for improved protein structural motif recognition. *Journal of Computational Biology*, *4*(3), 261–273.

Berger, B., Wilson, D. B., Wolf, E., Tonchev, T., Milla, M., & Kim, P. S. (1995). Predicting coiled coils by use of pairwise residue correlations. *Proceedings of the National Academy of Science USA*, *92*, 8259–8263.

Brāzma, A., Jonassen, I., Eidhammer, I., & Gilbert, D. (1995). Approaches to the automatic discovery of patterns in biosequences. Technical Report 113, Department of Informatics, University of Bergen, Norway.

Brennan, R. G. & Matthews, B. W. (1989). The helix-turn-helix DNA binding motif (minireview). *J. Biol. Chem.*, *263*(4), 1903–6.

Bushman, F. D., Shang, C., & Ptashne, M. (1989). A single glutamic acid residue plays a key role in the transcriptional activation function of lambda repressor. *Cell*, *58*(6), 1163–1171.

Dodd, I. B. & Egan, J. B. (1990). Improved detection of helix-turn-helix DNA-binding motifs in protein sequences. *Nucleic Acids Research*, *18*(17), 5019–26.

Gallegos, M., Schleif, R., Bairoch, A., Hofmann, K., & Ramos, J. (1997). AraC/XylS family of transcriptional regulators. *Microbiol. Mol. Biol. Rev.*, *61*(4), 393–410.

Gao, Y., Mathee, K., Narasimhan, G., & Wang, X. (1999). Motif detection in protein sequences. In *Proc. of the Sixth Intl. Symp. on String Processing and Information Retreiaval*, (pp. 63–72).

Gribskov, M., Lüthy, R., & Eisenberg, D. (1990). Profile anlysis. In R. F. Doolittle (Ed.), *Molecular Evolution: Computer Analysis of Protein and Nucleic Acid Sequences*, volume 183 of *Methods in Enzymology* (pp. 146–159). Academic Press.

Gruber, T. M. & Bryant, D. A. (1997). Molecular systematic studies of eubacteria, using σ^{70}-type sigma factors of group 1 and group 2. *J. Bacteriol.*, *179*(5), 1734–1747.

Grundy, W. N., Bailey, T. L., Elkan, C. P., & Baker, M. E. (1997). Hidden markov model analysis of motifs in steroid dehydrogenases and their homologs. *Biochem. Biophys. Res. Comm.*, *231*, 760–766.

Gusfield, D. (1997). *Algorithm on Strings, Trees, and Sequences.* Cambridge University Press.

Harrison, S. C. & Aggarwal, A. K. (1990). DNA recognition by proteins with the helix-turn-helix motif. *Ann. Review of Biochem.*, *59*, 933–969.

Henikoff, S. & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA*, *89*, 10915–10919.

Hoch, J. A. & Silhavy, T. J. (1995). *Two-component signal transduction.* Amer. Soc. Microbiol. Press, Washington D.C.

Kim, S. K., Makino, K., Amemura, M., Nakata, A., & Shinagawa, H. (1995). Mutational analysis of the role of the first helix of region 4.2 of the sigma 70 subunit of escherichia coli RNA polymerase in transcriptional activation by activator protein PhoB. *Mol Gen Genet*, *248*(1), 1–8.

Lawrence, C. E., Altschul, S. F., Boguski, M. S., Liu, J. S., Neuwald, A. F., & Wootton, J. C. (1993). Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignment. *Science*, *262*, 208–214.

Lonetto, M., Gribskov, M., & Gross, C. A. (1992). The σ^{70} family: Sequence conservation and evolutionary relationships. *J. Bacteriol.*, *174*(12), 3843–3849.

Nelson, H. C. M. (1995). Structure and function of DNA-binding proteins. *Current Opinion in Genetics and Development*, *5*, 180–189.

Nevill-Manning, C. G., Wu, T. D., & Brutlag, D. L. (1998). Highly-specific protein sequence motifs for genome analysis. *Proc. Natl. Acad. Sci. USA*, *95*, 5865–5871.

Pabo, C. O. & Sauer, R. T. (1992). Transcriptional factors: Structural families and principle of DNA recognition. *Ann. Rev. Biochem.*, *61*, 1053–95.

Parida, L., Rigoutsos, I., Floratos, A., Platt, D., & Gao, Y. (2000). Pattern discovery on character sets and real valued data: Linear bound on irredundant motifs and an efficient polynomial time algorithm. In *Proc. of the 11th Annual ACM/SIAM Symposium on Discrete Algorithms (SODA '00)*, (pp. 297–308).

Pfau, J., Arvidson, D. N., & Youderian, P. (1994). Mutants of escherichia coli trp repressor with changes of conserved, helix-turn-helix residue threonine 81 have altered DNA-binding specificities. *Mol. Microbiol.*, *13*(6), 1001–1012.

Rigoutsos, I. & Floratos, A. (1998). Motif discovery in biological sequences without alignment or enumeration. In *Proc. of the Second Annual International Conference on Computational*

Molecular Biology, RECOMB 98, (pp. 221–227).

Rigoutsos, I., Gao, Y., Floratos, A., & Parida, L. (1999). Building dictionaries of 1d and 3d motifs by mining the unaligned 1d sequences of 17 archaeal and bacterial genomes. In *Proc. of the 7th Intl. Conference on Intelligent Systems for Molecular Biology (ISMB '99)*, (pp. 223–233).

Schell, M. A. (1993). Molecular biology of the LysR family of transcriptional regulators. *Annu Rev Microbiol*, *47*, 597–626.

Scott, M. P., Tamkun, J. W., & Hartzell, G. W. (1989). The structure and function of the homeodomain. *Biochim Biophys Acta*, *989*(1), 25–48.

Siegele, D. A., Hu, J. C., Walter, W. A., & Gross, C. A. (1989). Altered promoter recognition by mutant forms of the σ^{70} subunit of escherichia coli RNA polymerase. *J Mol Biol.*, *206*(4), 591–603.

Singh, M., Berger, B., & Kim, P. S. (1998). Computational learning reveals coiled coil-like motifs in histidine kinase linker domains. *Proceedings of the National Academy of Science USA*, *95*(6), 2738–2743.

Wolf, E., Kim, P. S., & Berger, B. (1997). Multicoil: A program for predicting two- and three-stranded coiled coils. *Protein Science*, *6*, 1179–1189.

Wu, T. D. & Brutlag, D. L. (1996). Discovering empirically conserved amino acid substitution groups in databases of protein families. In *Porceedings of the Fourth International Conference on Computational Biology (ISMB-96)*, (pp. 230–240).