COT 6936

Term Project Report

HTH Detection Training Set Selection Based on Phylogenetic Trees

by

Yanli Sun (suny@cs.fiu.edu)
Zhenyue Deng (zdeng01@cs.fiu.edu)

1. Introduction

A motif is a portion of a protein sequence that has a specific structure and is functionally significant. The presence of motifs in a protein is very useful for characterizing and classifying that protein.

GYM is a motif prediction algorithm based on data mining technique. It was developed by Dr. Giri and his research group. GYM includes two phases: pattern mining and motif detection. In pattern mining phase, GYM examines the training set of a specific type of motif, progressively discovers significant patterns and stores the maximal ones into pattern dictionary. In GYM, a larger pattern is obtained by merging two significant sub-patterns of one amino acid less in length and one amino acid in difference. In the detection phase, GYM examines how well a given protein sequence matches the patterns in the dictionary so as to predict the presence and location of the motif in that protein.

GYM uses a threshold **Support** to identify the significance of pattern occurrence among the training set in the pattern mining phase. This threshold represents the trade-off between the prediction sensitivity and false positive rate in detection phase. A pattern is said to be significant only if the number of its occurrences exceeds the threshold.

GYM uses 88 known HTH motif sequences as its training set. From the experiment results, it exhibits excellent ability in predicting HTH motifs in some protein families such as Sigma and LysR. Also a small number of false positives (7%) were found in experiments in Negetes family in which HTH motif is unlikely to exist.

This project explores a training set selection (or refinement) algorithm by means of reducing biased sequences in training set to improve the overall performance of GYM. In our algorithm,

Phylogenetic Tree is used as an effective tool to figure out similarity factors among sequences in order to identify biased sequences.

2. Theoretical basis

Pattern mining is actually a learning process and training set is the very source where knowledge can be obtained. From this point we can see that the selection of the training set is vital to the success of the mining phase, and in turn, the whole algorithm. There are several difficulties in choosing a “good” training set:

- Some errors or inaccuracy may exist.
- Some sequences might be redundant because they might be just slightly different in composition due to works conducted and reported at independent labs or by mutation.
- There might be an excessive number of motifs of a specific structure in the set, therefore biasing the result towards this specific structure.

It is obvious that redundant and biased sequences exhibit significant similarity to each other. Those compositionally very similar sequences in the training set may result in false (spurious) patterns with respect to following two cases:

- Some non-HTH motif enforcing amino acid residues could be picked up as enforcing pattern due to their high occurrences among those similar sequences. In this case, the patterns contain purely non-motif-enforcing residues. We refer to it as a "pure spurious pattern"

- Sometimes non-motif-enforcing residues adjacent to motif-enforcing residues could be chosen as part of the pattern. We refer to it as a "partial spurious pattern".

Obviously, pure spurious patterns are the major cause of false positives. Partial spurious patterns can cause false negative because some actual motif sequences cannot match the entire pattern very well due to the extra non-enforcing sub-pattern inside the entire pattern.

For pattern mining process, it is necessary that actual motif patterns must present enough frequency (higher than the *threshold* value) so as to be successfully discovered. The higher the frequency of a pattern, the more likely it is to be discovered. On the other hand, if the number of similar sequences that appear in the training set exceeds a reasonable value, some non-motif-enforcing residues could be mistakenly chosen as part of motif-enforcing patterns due to their high occurrences.

Ideally, choosing the sequences in a training set should consider the following two criteria:

- All sequences evenly fall into some logical subgroups according to their alignment similarity or biological similarity.
- In each subgroup, the number and permutation of contained sequences exhibit high occurrences of good patterns and non-enforcing "noise" scattered randomly.

Therefore, by carefully controlling the similarity present among training sequences, it is possible to effectively reduce the probability of false patterns in GYM's data mining and improve its overall performance.

3. Training Set Selection Algorithm

Similarity among motif sequences can be figured out by either pairwise or multiple alignments. Their results are given in the form of scores. Although alignment scores are sufficient to indicate how well the sequences match one another, they are unable to give further information among those sequences, such as evolutionary and classification information.

Phylogenetic Tree is a useful tool to identify similarities among protein sequences as well as their evolutionary and classification information.

Topologically, such trees classify different proteins into different families (sub-trees) and the value bound to each branch gives evolutionary distance for the underlying family of proteins. Sequences in one family are closer than those from different families. For an internal node, the farther it is away from the root, the more likely that all its descents are closer.

In order to precisely measure the topological and evolutionary factors in a phylogenetic tree, we developed a comprehensive value (score) for each node. This value is recursively defined as:

- The root's value is 0
- For a non-root node, the value is defined as:

$$(a \times \text{Degree})^{(\text{TD-depth})} * (\text{Index} + b \times (1 - \text{DistanceP}) + g \times (1 - \text{Distance})$$

Where

Options	Description
Degree	The maximum branching degree present in the tree.
TD	Total depth of the tree
Depth	The depth of current node
Index	The sequential index among siblings
DistanceP	The total distance between root and its parent node.
Distance	The distance from the underlying node to its parent
a	Constant factor, default is 1.5
b	Constant factor, default is 1
g	Constant factor, default is 1

The difference between comprehensive values of two nodes reflects the evolutionary similarity between them. Once the comprehensive values have been obtained for all sequences, they are put into a list in order by traversing the tree in a depth-first manner. The pseudo code to evenly pick up training set sequences based on their comprehensive values is described as below:

```
Num := N                    #sizeof(CandidateSet)
While (Num > DesiredSize)
Do
   Find pair (n_i,n_i+1) of minimum distance
   If I==1 then remove n_i           #first
   else If I==N then remove n_i+1     #last
   else
      Middle := ( value[n_i-1] + value[n_i+2] ) /2
      If value[n_i] is close to Middle
      then remove n_i    else remove n_i+1
```

4. Implementation

The above algorithm has been implemented in a C++ program called *Select*. The candidate set is the original Master Set of GYM, containing 88 HTH motif sequences with length 22. The phylogenetic trees were generated by feeding the candidate set into CLUSTAL W with different scoring matrices and tree types. The phylogenetic tree is given in a textual form called tree script. A simple example of tree script and its phylogenetic tree is shown in Figure 1.

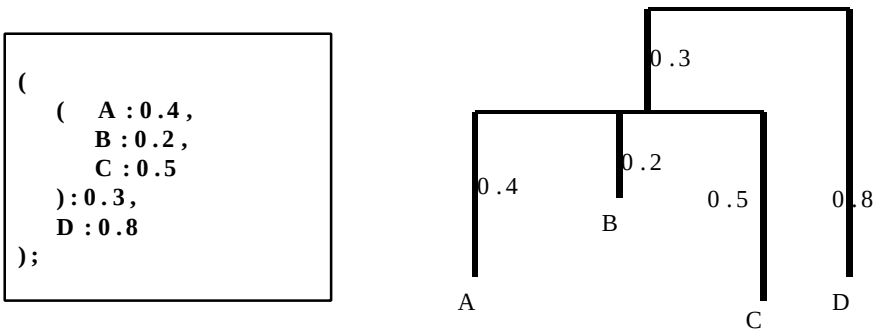

Figure.1 Illustration of phylogenetic tree&Script

Select takes the tree script and other parameters (number of output sequences out of the original 88 sequences as well as a, b and g) as input. Based on the input, the program automatically calculates comprehensive values for each sequence. The final~~ly~~ output training set contains desired number of

sequences whose comprehensive values are evenly distributed. The data flow of Select program is shown in Figure-2.

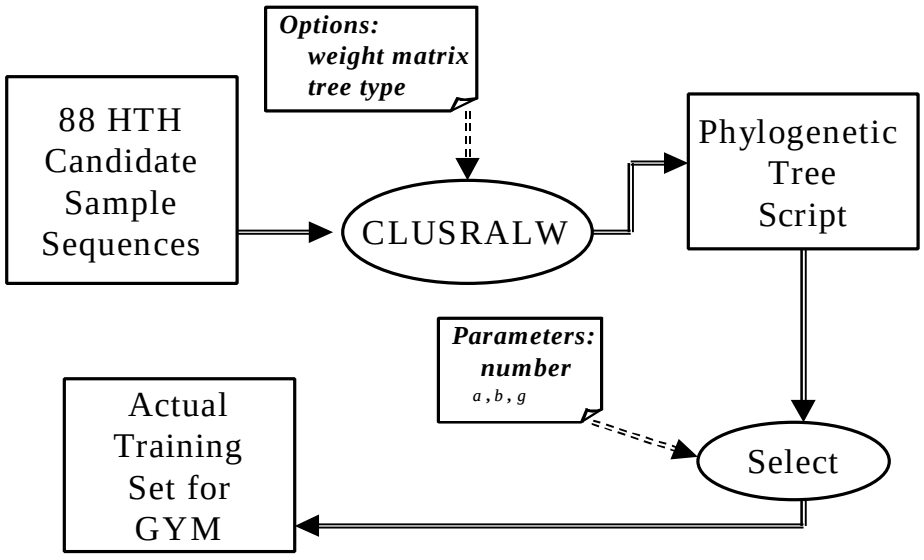

Figure.2 Illustration of Data Flow

Select is implemented as a command line application that can be run on UNIX, LINUX and Windows systems. The usage of Select is listed as:

Select -iTree -oOutput -nNum -aA -bB -gG

Where

Tree	The tree script file
Output	Output file name--Training set
Num	Number to be chosen from the set of candidates

A	a ---default is 1.5
B	b ---default is 1
G	g---default is 1

5. Results and Analysis

We used two different phylogenetic trees generated from CLUSTAL W based on GYM's Master Set in our testing. One was generated with the default setup and the other one was generated with BLOSUM matrix and PHILIP tree type. The two trees exhibit some structural differences as below:

	CLUSTAL W Option		**Tree Depth**	**Maximum Degree**
	Matrix	**Tree Type**		
Tree-1	Default	Default	14	3
Tree-2	BLOSUM	Philip	9	3

Based on the generated scripts Tree-1 and Tree-2, we used ***Select*** to generate a series of training sets whose lengths were ranging from 20 to 85 out of the total 88 sequences with different combinations of a, b and g. Those training sets were then tested on GYM and their results were carefully analyzed.

In experiments, we found that the results of ***Select*** are not sensitive to parameters a, b and g as long as they fall in a reasonable range ($a>1$, $b \in (0,1)$ and $g \in (0,1)$). Beyond the reasonable range, ***Select*** can yield quite different results, which usually lead to poor results on GYM. The default setup for a, b and g ($a=1.5$; $b=1$; $g=1$) is good enough in practice.

GYM behaved differently on different training sets varying in length and tree type (Tree-1 and Terr-2). We will discuss them later in this section. For the sake of comparison, the original GYM 2.0 testing result is also given in Table-1.

Protein Family	Number of Sequences Tested	GYM = DE Agree	How many Annotated	GYM= Annotated	False Positive
Master	88	88(100%)	13	13	N/A
Sigma	314	284+23(98%)	96	82	N/A
Negate	93	86(92%)	0	0	7
LysRe	130	127(98%)	95	93	N/A
Arace	68	57(84%)	41	34	N/A
Rreg	116	99(85%)	57	46	N/A
Total	809	764(94%)	302	268(89%)	

Table-1. GYM Original Testing Result

5.1. Result of first group based on Tree-1

Testing on Tree-1 derived testing cases did not show significant improvement. Some testing results are listed in table-2, 3 and 4.

Protein Family	Number of Sequences Tested	GYM = DE Agree	How many Annotated	GYM= Annotated	False Positive
Master	88	79(90%)	13	12	N/A
Sigma	314	285+23(98%)	96	89	N/A
Negate	93	88(95%)	0	0	5
LysRe	130	130(100%)	95	91	N/A
Arace	68	59(87%)	41	31	N/A

Rreg	116	101(87%)	57	55	N/A
Total	809	765(95%)	302	278(92%)	

Table-2. GYM Result based on [DEFAULT, DEFAULT, 75]

Protein Family	Number of Sequences Tested	GYM = DE Agree	How many Annotated	GYM= Annotated	False Positive
Master	88	83(94%)	13	13	N/A
Sigma	314	288+23(99%)	96	89	N/A
Negate	93	87(94%)	0	0	6
LysRe	130	127(98%)	95	93	N/A
Arace	68	59(87%)	41	31	N/A
Rreg	116	100(86%)	57	56	N/A
Total	809	759(94%)	302	282(93%)	

Table-3. GYM Result based on [DEFAULT, DEFAULT, 80]

Protein Family	Number of Sequences Tested	GYM = DE Agree	How many Annotated	GYM= Annotated	False Positive
Master	88	83(94%)	13	13	N/A
Sigma	314	287+23(99%)	96	90	N/A
Negate	93	86(92%)	0	0	7
LysRe	130	128(98%)	94	93	N/A
Arace	68	59(87%)	35	32	N/A
Rreg	116	98(84%)	45	56	N/A
Total	809	764(94%)	302	284(94%)	

Table-5. GYM Result based on [DEFAULT, DEFAULT, 81]

Although in some cases, GYM can give better prediction rate on some special protein families and lower false positive rate for Negates family, the ability to predict motif presence in Master Set is very poor. Since different score matrices were used in training set selection and GYM, they interpret similarity in different ways. Select did not remove the most similar sequences from the point of GYM. We believe that is the main reason to explain above scenario.

5.2. Result of second group based on Tree-2

Several testing cases showed significant improvements in GYM, with increased detection rate on some protein families (e.g. Sigma, Rege and Lysr) where HTH motif existences are verified and decreased false positive rate on Negates family where HTH motif is unlikely to exist.

The GYM testing results upon those Tree-2 derived training sets are listed in Table-5, 6 and 7.

Protein Family	Number of Sequences Tested	GYM = DE Agree	How many Annotated	GYM= Annotated	False Positive
Master	88	88(100%)	13	11	N/A
Sigma	314	283+23 (98%)	96	89	N/A
Negate	93	89 (96%)	0	0	4
LysRe	130	127(98%)	95	89	N/A
Arace	68	55(81%)	41	26	N/A
Rreg	116	98(85%)	57	55	N/A
Total	801	763(94%)	302	270(89%)	

Table-5. GYM Result based on [BLOSUM, PHILIP, 80]

Protein Family	Number of Sequences Tested	GYM = DE Agree	How many Annotated	GYM= Annotated	False Positive
Master	88	88(100%)	13	12	N/A
Sigma	314	283+23(97%)	96	89	N/A
Negate	93	88(95%)	0	0	5
LysRe	130	127(98%)	95	89	N/A
Arace	68	58(85%)	41	31	N/A
Rreg	116	98(84%)	57	56	N/A
Total	809	765(95%)	302	277(92%)	

Table-6. GYM Result based on [BLOSUM, PHILIP, 82]

Protein Family	Number of Sequences Tested	GYM = DE Agree	How many Annotated	GYM= Annotated	False Positive
Master	88	88(100%)	13	13	N/A
Sigma	314	283+23(97%)	96	89	N/A
Negate	93	87(94%)	0	0	6
LysRe	130	127(98%)	95	89	N/A
Arace	68	58(85%)	41	33	N/A
Rreg	116	96(83%)	57	56	N/A
Total	809	762(94%)	302	280(93%)	

Table-7. GYM Result based on [BLOSUM, PHILIP, 84]

From the above exciting results, it's not hard to conclude that improvement mainly came from the reduction of two types of false patterns in the pattern dictionary. That is, out algorithm can effectively filter out redundant and biased sequences in the training set.

6. Conclusion

This project presents a promising approach for training set refinement in pattern mining by means of similarity control among sequences. For pattern mining technique, like the two sides of a coin, similarity represents the trade-off between the sensitivity of both true positives and false positives. In practice, the optimal similarity control can only be achieved by experiments. Theoretically, there is no algorithm that can automatically figure out the optimal similarity threshold without further biological knowledge.

7. Acknowledgement

Many thanks go to Dr. Giri, for his invaluable advice and extensive help, as well as the original idea on the training set selection algorithm.

8. Reference List

[1] Narasimhan, G., Bu, C., Gao, Y., Wang, X., Xu, N., Mathee, K. (2001). Mining Protein Sequences For Motifs. Journal of Computational biology.

[2] ClustalW manual.

[3] Gao, Y., Yang, M., Wang, X., Mathee, K., Narashimhan, G. (1997) Detection of HTH motifs via Data Mining. Int’l Conference on Bioinformatics.

[4] Dr. Narasimhan’s lecture notes. (2002).