Motif Discovery from Large Number of Sequences: a Case Study with Disease Resistance Genes in *Arabidopsis thaliana*

Irfan Gunduz, Sihui Zhao, Mehmet Dalkilic and Sun Kim
Indiana University, School of Informatics
Informatics Building 901 East 10th Street Bloomington, IN 47408
{igunduz, sizhao, dalkilic, sunkim2}@indiana.edu

Abstract

Motif discovery from a set of sequences is a very important problem in biology. Although a lot of research has been done on computational techniques for (sequence) motif discovery, discovering motifs in a large number of sequences still remains challenging. We propose a novel computational framework that combines multiple computational techniques such as pairwise sequence comparison, clustering, HMM based sequence search, motif finding, and block comparisons. We tested this computational framework in its ability to extract motifs from disease resistance genes and candidates in Arabidopsis thaliana genome and discovered all known motifs relating to disease resistance. When the same set of sequences was submitted to MEME and Pratt (motif discovery tools) as a whole without clustering, they failed to detect disease resistance gene motifs. The crucial component in this framework is clustering. Among the benefits of clustering is computational efficiency since the set of sequences are divided into smaller groups using a clustering algorithm.

1. Introduction

Motif is a highly conserved region across a subset of proteins that share the same function. Such conservation is mainly due to the relatively higher selection pressure through the evolution while the nonessential region will diverge from each other [9].

Motif detection is one of the central problems in biology; consequently, a lot of research has focused on developing computational techniques for their discovery. In general, motif discovery relies upon either statistical or combinatorial pattern search techniques [3,5,6]. Both of these approaches become less effective, however, as the number of sequences increases. This is true especially when a set of sequences include multiple families. Obviously, motif discovery will be more effective when the input set of sequences is divided (clustered) into multiple groups, each of which corresponds to a set of functionally similar sequences. However, clustering sequences is another challenging problem in which the current computational techniques have had limited success. Thus, the combination of clustering and motif discovery is not trivial.

We propose a novel computational framework that combines multiple computational techniques such as pairwise sequence comparison, clustering, hidden Markov model (HMM) based sequence search, motif finding, and block comparisons. By using this framework, we extracted and clustered a set of disease resistance genes and candidates in *Arabidopsis thaliana* and discovered all known motifs relating to disease resistance. When the same set of sequences submitted to MEME and Pratt (motif discovery tools) as a whole without clustering, they fail to detect disease resistance gene motifs. In addition, our procedure is computational efficient since the set of sequences are divided into smaller groups using a clustering algorithm.

2. Computational Framework

The basic intuition underlying our framework is letting the clustering guide motif discovery. However, the main concern is that there is no guarantee that the clustering result is correct. To circumvent this hurdle, we first cluster sequences with very strict criteria, thus each group will contain sequences from the same family with high confidence. There may be different groups belonging, however, to the same family; and so, they need to be merged. Each of the resulting groups of sequences is submitted to motif discovery programs. The procedure is illustrated in

1. The initial set of well annotated sequences of interest is extracted from genomes and all pairwise sequence comparisons are done using BLAST [8]. Using a score from pairwise comparison such as E-value or Smith-Waterman scores and percentage of overlap, the set of

sequences are clustered into multiple groups. We used BAG [1] for the clustering step.

2. A hidden Markov model (HMM) is built for each cluster. Then each cluster is expanded by an iterated HMM search. We used HMMER [2] for this step. On each HMM iteration, clusters need to be refined by discarding clusters which are the subset of other clusters.
3. If the number of sequences is too large in one cluster, sequences in the cluster need to be clustered again into smaller groups to reduce background noise for motif detection. This can be easily done by utilizing the pairwise comparisons computed during the first stage of the procedure.

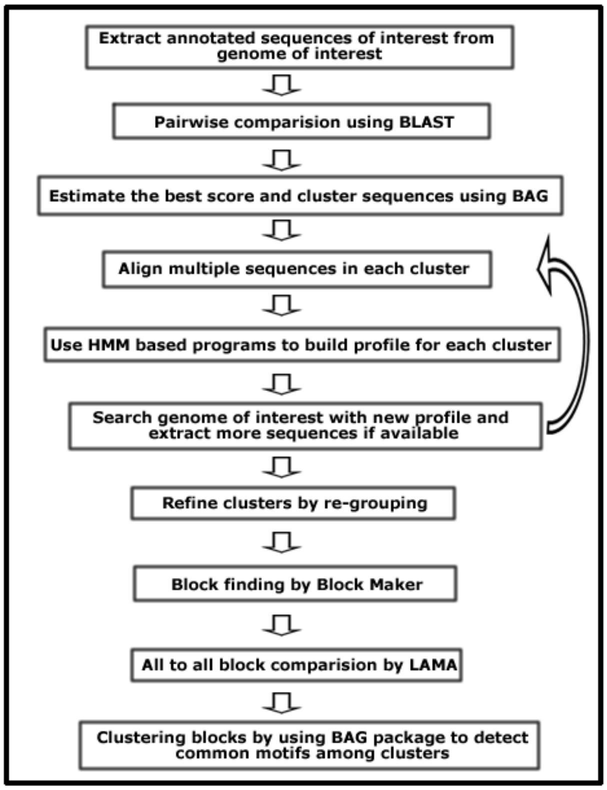
Extract annotated sequences of interest from genome of interest
Pairwise comparision using BLAST
Estimate the best score and cluster sequences using BAG
Align multiple sequences in each cluster
Use HMM based programs to build profile for each cluster
Search genome of interest with new profile and extract more sequences if available
Refine clusters by re-grouping
Block finding by Block Maker
All to all block comparision by LAMA
Clustering blocks by using BAG package to detect common motifs among clusters

Figure1. Computational framework to find motifs for large number of sequences

4. To detect discrete and common motifs for each cluster, sequences in each cluster are submitted to BlockMaker [3]. Blocks detected by both Gibbs and Motifs are considered for further analysis. Alternatively, instead of BlockMaker,
5. sequences in each cluster can be submitted directly to MEME [5] or Pratt[6] to find motifs.
6. Blocks then can be compared to each other by using LAMA [4], block comparison software. Blocks from different clusters are matched (merged) by clustering with BAG using all pairwise block comparisons using LAMA. Common motifs in different clusters are identified in multiple clusters to make a final set of clusters.

3. Validation Experiment

To test our computational framework, 116 well-annotated disease resistance genes and candidates were identified from *Arabidopsis thaliana* genome and clustered in 32 distinct groups. Each group consists of 2 to 17 sequences. HMMER [2] was run to pick similar sequences from *Arabidopsis thaliana* genome database. After several HMM iterations, clusters were refined. A cluster was discarded if it was subsumed by another larger cluster. As a result 28 clusters were discarded. Remaining four clusters contained 792 unique sequences. Cluster-1 possesses 96, Cluster-2 44 Cluster-3 641 and Cluster-4 11 sequences. To ensure that all sequences in each cluster are related to disease resistance genes or candidates, sequences in each cluster were searched against the PFAM database. Sequences in Cluster 1 and 2 have domains that exhibit high similarity to LRR (leucine rich repeat regions), Kinase, and NB-ARC. Most of the cloned disease resistance genes possess these domains [7] indicating that our HMMER search detected disease resistance related sequences in Cluster 1 and 2. On the other hand sequences in cluster 3 and 4 exhibited similarity with serine/throsine kinase domains.

Sequences from each cluster were than submitted to Block Maker to detect conserved regions as blocks. In each cluster 2 to 12 blocks were obtained. Block Maker utilizes two different techniques: GIBBS and MOTIFS. Motifs that were identified by both were then accepted for the next step. LAMA was then used for block pair-wise comparison and a Z-score value of five was set as the cutoff value for BAG to cluster LAMA results. As a result, three common blocks in cluster 1 and cluster 2 are identified (see Fig. 2).

For the comparison with motif discovery without clustering, all four clusters were merged, redundant sequences were eliminated, and the resulting 792 unique sequences were submitted to MEME and Pratt to find motifs.

With the computational framework described above, 5 motifs, including (Toll-IL receptor) TIRI, TIRII, Kin1, Kin2, (Nucleotide Binding Site) NBS-B and LRR (Leucine reach repeats) were detected in 75 of 96 sequences of Cluster-1. Sequences in cluster-1 include RPS4 and putative RPP1 and RPP5 disease resistance genes. Six motifs including RNBS-A, RNBS-B, RNBS-D, Kin1, Kin2 and GLPLA motifs were detected in 39 of

44 sequences in cluster 2. Resistance genes have been recognized by the presence of these motifs [7]. Disease resistance genes RPM1, putative RPP8 and RPP13 were detected in cluster 2. Previous studies indicate that there are two major classes of disease resistance genes in plants. Genes in of the group conserve motifs comprising N-terminal domain with Toll/Interleukin-1 receptor homology (TIR), Kinase, NBS and LRR domains. Resistance genes RPP1 and RPP5 are in the first group. Genes in second group conserve motifs comprising Kinase, NBS and LRR domains. Main difference between these two classes of resistance genes are the absence or presence of TIR domain [7]. Our computational procedure above clearly grouped disease resistance genes in two groups one with TIR and other without TIR (Figure 2). Furthermore in each group functional motifs were successfully determined.

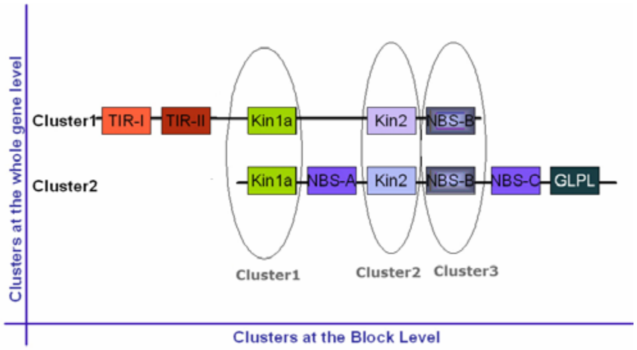

Figure 2. Disease resistance gene clusters. Disease resistance related sequences were clustered shown at the horizontal level. Blocks from Cluster-1 and Cluster-2 were clustered shown at the horizontal level.

VFXSFRGXDVRXXFLSH determined as consensus sequence for TIRI domain in cluster-1 based on entropy. X indicates that in a given position none of the amino acid is more then 50% or frequency of two amino acids together not more than 0.8. Relative entropy of this motif shown as logos [10] (Figure 3).

Same as YASSSWCLDEL were determined as consensus sequence for TIRII (Figure 4), VGIXGXXGIGKTTI were determined as consensus sequence for Kin1a (Figure 5), consensus sequence VLDDVD for Kin2 and consensus sequence GSRIIVTTXD for NBS-b domains (Figure 6).

In cluster-2 consensus sequence GMGGXGKTTL for Kin1a with pre-p-loop (Figure 7), consensus sequence FDXXIWVXVS for NBS-a (Figure 8), consensus sequence LLXLDDXW for Kin2 (Figure 9), consensus sequences GXKVXXTTR for NBS-b (Figure 10), consensus sequence CLXXXEAWXLF (Figure 11) and consensus sequences CXGLPLA for GLPL motifs (Figure 12).

Relative entropy of LRR domain in Cluster-1 and Cluester-2 shown as logos in figure 13 and 14 respectively.

In cluster-1 three new sequences which were not annotated as disease resistance like protein were identified those are GI 15236505, GI 15242136 and GI 15233862. In cluster-2 sequences GI 15221277, GI 15221280, GI 15217940 and GI 15221744 were detected to share common motifs with disease resistance genes however were not annotated as disease resistance like proteins.

Only Serine/Throsine Kinase domain were detected in cluster 3 and cluster 4. Furthermore none of the previously cloned disease resistance genes were detected in these groups. Therefore sequences in clusters three and four are not considered as disease resistance related.

BAG successfully clustered common blocks from both groups. As a result Kin1a, Kin2 and RNBS-B from both groups were clustered in same groups (Figure2). Protein kinase domains were detected in sequence cluster 3, but no motifs related to disease resistance genes have been detected in sequence clusters 3 and 4. Clustering LAMA results is useful to detect common motifs among different classes of sequences.

When sequences in each cluster were submitted to MEME and Pratt, they detected most of those domains. When sequences without clustering were submitted to MEME, it took more than 9,000 minutes on a Pentium IV 1.7 GHz machine running Linux to find motifs for both MEME and Pratt. Furthermore, none of the disease resistance related motifs were detected.

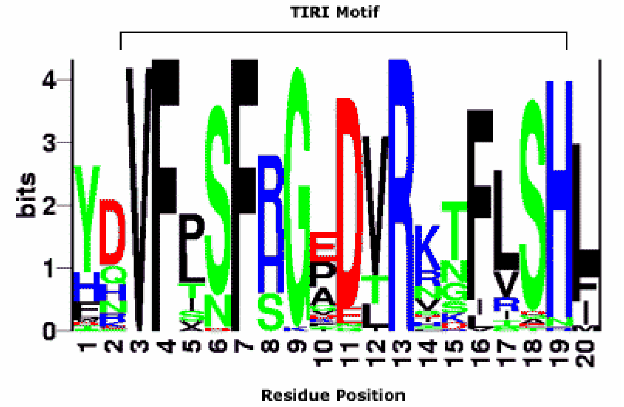

Figure 3. Logos representing relative entropy of Block1 in cluster-1. TIR motifs start from 3th position and end up 19.

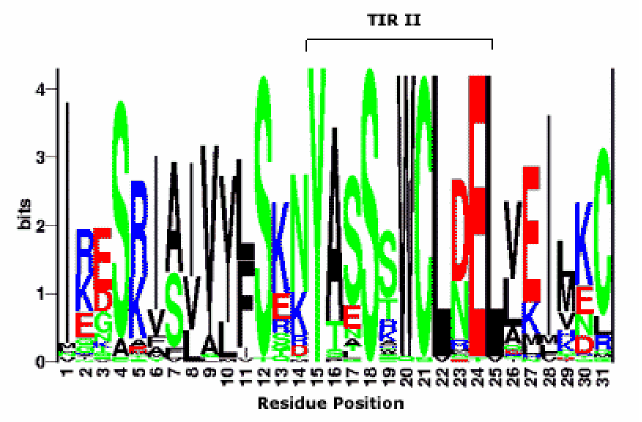

Figure 4. Logos representing relative entropy of Block2 in cluster-1. TIRII motifs start from 15th position and end up 25.

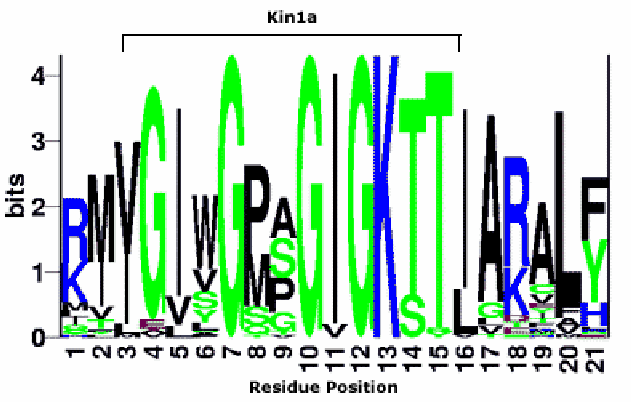

Figure 5. Logos representing relative entropy of Block3 in cluster-1. Kin1a with pre-p-loop start 3th position and en up 15.

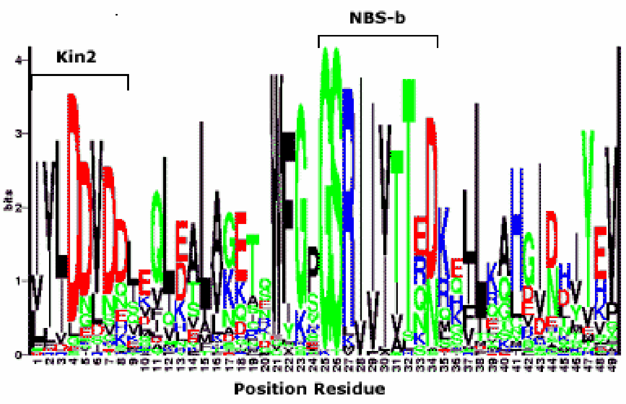

Figure 6. Logos representing relative entropy of Block4 in cluster-1. Positions 1 to 7 are Kin2 motif and Position 25 to 34 are NBS-b motif.

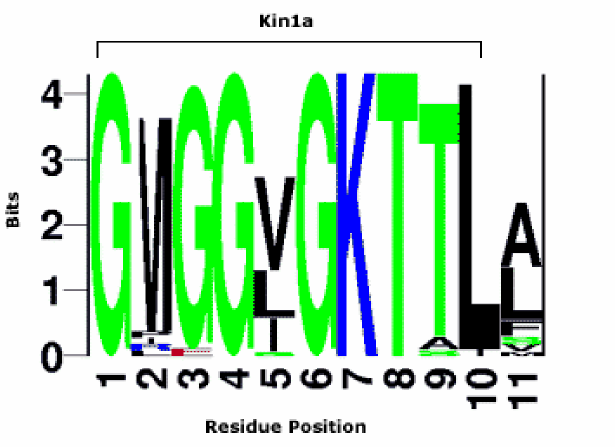

Figure 7. Logos representing relative entropy of Block1 in cluster-2. Kin1a start position 1 and up position 10.

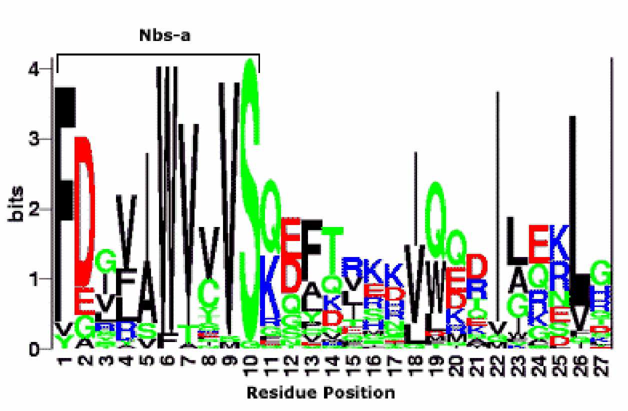

Figure 8. Logos representing relative entropy of Block2 in cluster-2. NBS-a motif start at position 1 and en up position 10.

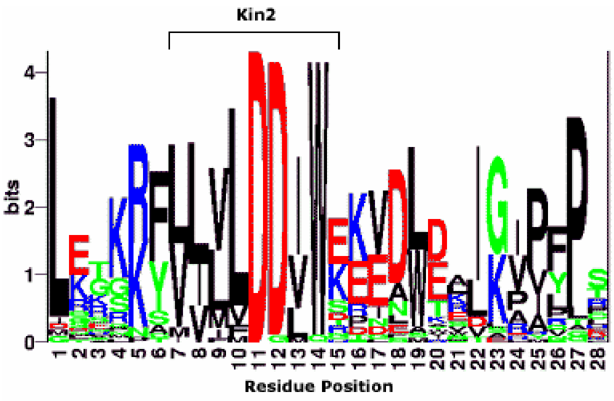

Figure 9. Logos representing relative entropy of Block3 in cluster-2. Kin2 motif start at position 7 and end up position 14.

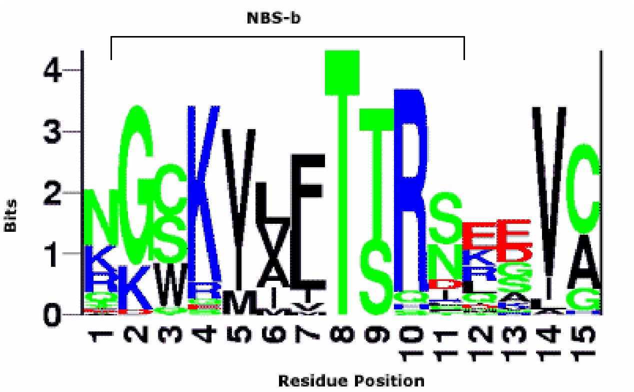

Figure 10. Logos representing relative entropy of Block4 in cluster-2. NBS-b motif start at position 2 and end up position 11.

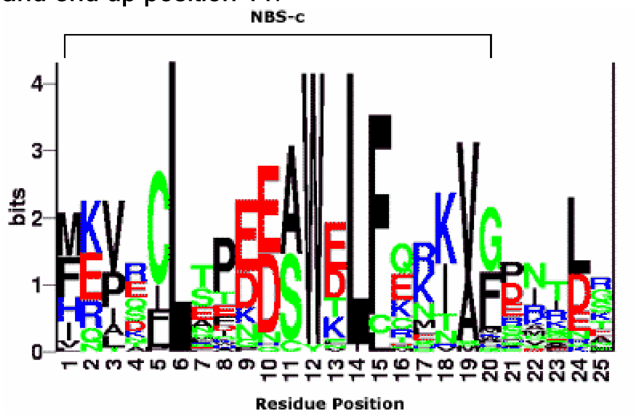

Figure 11. Logos representing relative entropy of Block5 in cluster-2. NBS-c motif start at position 5 and end up position 15.

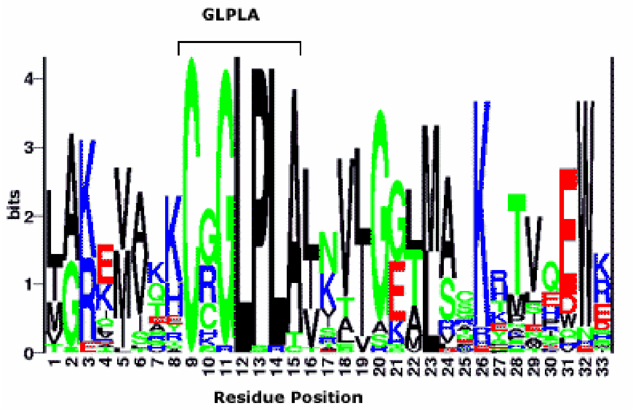

Figure 12. Logos representing relative entropy of Block6 in cluster-2. GLPLA motif start from at position 9 and end up position 15.

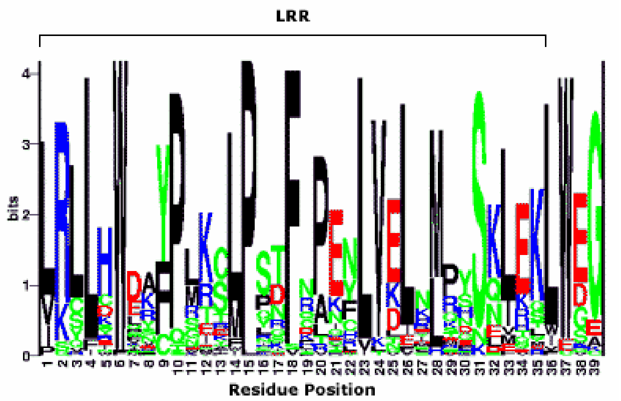

Figure 13. Logos representing relative entropy of Block7 in cluster-1. Leucine reach repeats shown.

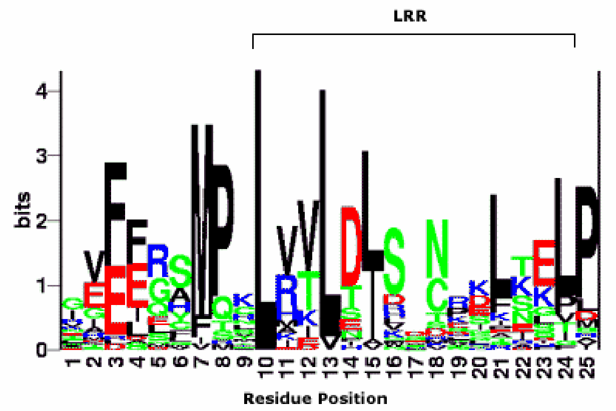

Figure 14. Logos representing relative entropy of Block9 in cluster-2. Leucine reach repeats shown.

4. Conclusion

Our framework demonstrates that sensible combination of tools provides an excellent mechanism for motif detection. At the core of this framework is the identification and use of clustering to help improve the performance of other, well-known tools. Future work includes generalizing this framework and further refinement of the clustering component.

5. Acknowledgements

This work was supported in part by INGEN (Indiana Genomics Initiatives), IBM SUR Grant, and NSF CAREER Award DBI-0237901

5. References

1- Kim, S. (2003) ``Graph Theoretic Sequence Clustering Algorithm and Their Applications to Genome Comparisons,” Computational Biology and Genome

Informatics edited by Jason T. L. Wang, Cathy H. Wu, and Paul Wang, World Scientific, 2003 (http://bio.informatics.indiana.edu/sunkim/BAG).

2- Eddy, S.R. (2001).HMMER: Profile hidden Markov models for biological sequence analysis (http://hmmer.wustl.edu/).

3- Henikoff, S., Henikoff, J.G, Alford, W.J, and Pietrokovski, S. (1995), Automated construction and graphical presentation of protein blocks from unaligned sequences, Gene 163:GC17-26.

4- Pietrokovski S, Henikoff JG, Henikoff S(1996), "The Blocks database—a system for protein classification," Nucleic Acids Research 24(19) 3836-3845.

5- Timothy L. Bailey and Charles Elkan, "Unsupervised Learning of Multiple Motifs in Biopolymers using EM," *Machine Learning*, 21(1-2):51-80, October, 1995.

6- Inge Jonassen*, "*Efficient discovery of conserved patterns using a pattern graph,"
CABIOS 13, 509-522 (1997).

7- Meyers, C.B., A.W. Dickerman, R.W.Michelmore, S. Sivaramakrishnan, B.W. Sobral and N.D. Young (1999). Plant disease resistance genes encode members of an ancient and diverse protein family within the nucleotide binding superfamily. The Plant Journal. 20(3), 317-332

8- Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D.J. (1997) "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic Acids Res. 25:3389-3402

9-Cynthia Gibas, Per Jambeck Developing bioinformatics computer skills.

10- J. Gorodkin, L. J. Heyer, S. Brunak and G. D. Stormo. Displaying the information contents of structural RNA alignments: the structure logos. Comput. Appl. Biosci., Vol. 13, no. 6 pp 583-586, 1997.