DNA Sequence Analysis using Hierarchical ART-based Classification Networks

Cathie LeBlanc
Dept of Computer Science
Florida State Univ
Tallahassee, FL 32306-4019
leblanc@cs.fsu.edu

Charles R. Katholi
Dept of Biostat and Biomath
Univ of Alabama at Birmingham
Birmingham, AL 35294
katholi@cis.uab.edu

Thomas R. Unnasch
Div of Geographic Medicine
Department of Medicine
Univ of Alabama at Birmingham
Birmingham, AL 35294

Susan I. Hruska
Dept of Computer Science
Florida State Univ
Tallahassee, FL 32306-4019
hruska@cs.fsu.edu

February 5, 1996

Abstract

Adaptive resonance theory (ART) describes a class of artificial neural network architectures that act as classification tools which self-organize, work in real-time, and require no retraining to classify novel sequences. We have adapted ART networks to provide support to scientists attempting to categorize tandem repeat DNA fragments from *Onchocerca volvulus*. In this approach, sequences of DNA fragments are presented to multiple ART-based networks which are linked together into two (or more) tiers; the first provides coarse sequence classification while the subsequent tiers refine the classifications as needed. The overall rating of the resulting classification of fragments is measured using statistical techniques based on those introduced by Zimmerman, et al. (1994) to validate results from traditional phylogenetic analysis. Tests of the Hierarchical ART-based Classification Network, or HABclass network, indicate its value as a fast, easy-to-use classification tool which adapts to new data without retraining on previously classified data.

1 Introduction

One of the major leading causes of infectious blindness in humans is human onchocerciasis, which is caused by *Onchocerca volvulus*, a parasite found in sub-saharan

Africa, South America, and Central America (Zimmerman, et al., 1994) (Thylefors, 1978). The need for rapid, automated, and adaptable analysis and classification of *Onchocerca volvulus* DNA fragments is the motivation of the work reported here. While the resulting methodology is much more widely applicable to the class of problems in classification of DNA sequences, we focus on its development and testing in this specific problem domain.

In earlier studies of the evolutionary history of this parasite, scientists relied on phylogenetic analysis augmented by time-consuming manual clustering methods and statistical comparisons to determine evolutionary similarities among fragments found in different parts of the world. The work presented here shows that use of artificial intelligence techniques in this problem solution speeds up the recognition of patterns which signal evolutionary similarities. An added advantage is the rapid adaptability of the tool to new data.

The artificial intelligence technique built into this tool is a class of artificial neural networks adapted from Adaptive Resonance Theory (ART) (Carpenter & Grossberg, 1991a). An example of these is an ART2 network which allows real-time, unsupervised classification of sequences. Because these networks use competitive learning, they do not allow hierarchical classifications, that is, every sequence belongs to one and only one category (Grossberg, 1987). In certain applications, hierarchical classifications (that is, categories within categories) are desirable (Nigrin, 1993), this being one. To this end, the Hierarchical ART-based Classification network, or the HABclass network, was built to employ several ART-based networks placed into two (or more) layers to allow various similarity levels in the final set of categories.

Once the networks have been trained, the fitness of the resulting set of categories can be rated using a modified F statistic (Zimmerman, et al, 1994). Several factors can affect the set of categories resulting from use of the tool and we must determine which set of categories is "best." The statistic measures inter-category similarity and intra-category similarity and combines these into one number which represents the overall fitness of the particular set of categories.

The effectiveness of this technique on the problem considered is evident in the results. The HABclass network is also significantly faster than previously used methods, typically requiring no more than 15 minutes (clock time) to complete a run, including integration of new data. This technique is a promising tool for scientists working to quickly determine similarities among large numbers of sequences of DNA.

2 The Problem

The problem that we have chosen to address regards the spread of *Onchocerca volvulus* around the world. Zimmerman, et al. (1994) used statistical techniques to suggest that the spread of *O. volvulus* to the New World occurred relatively recently, probably as a result of the slave trade. The statistical techniques involved in this study proved successful although time-consuming and computationally-intensive, and allowed statistical assessment of results in addition to traditional phylogenetic analysis. Because the evolutionary history of a parasite such as *O. volvulus* is important in determining the source of illnesses as well as cures, a faster, easier to use tool is needed.

In determining evolutionary history, we begin with a set of aligned DNA fragments taken from *O. volvulus* from around the world (Zimmerman, et al, 1994). Our goal is to determine which fragments are related to which other fragments. This problem is, first, one of sequence categorization. Researchers have used a variety of techniques to categorize biological sequences, particularly protein sequences (Harris, Hunter & States, 1992) (Ferran, Ferrara & Pflugfelder, 1993). Most of these techniques disregard the positional information of subfragments because the presence or absence of a motif is more important than where the motif might be found. In the problem presented in this paper, the position of a subfragment is very important. The adaptability and speed of artificial neural networks are particularly suited to such categorization problems.

Once categories have been determined through training the neural network on a set of training sequences, we then must determine into which category a new DNA fragment fits. This second phase of the problem becomes one of pattern recognition. Once again, pattern recognition has benefited greatly from artificial intelligence techniques. And, again, artificial neural networks are particularly suited to pattern recognition problems (Schalkoff, 1992).

Because the "right" grouping into categories is unknown, choice of an unsupervised learning technique for the artificial neural network is indicated. The tool developed should also be easily adaptable, that is, if confronted with sequences that are heretofore unseen, training on those sequences should occur without retraining on the entire set of previously learned sequences. ART networks were chosen to address this problem because they self-organize (that is, learning proceeds unsupervised), adapt easily to new sequences and when confronted with a sequence that is unlike any others seen by the network, they will create singleton categories (those with only one pattern). This last feature is important since the number of sequences initially available was small and may not have adequately covered the input space.

Because certain groups of DNA fragments may be of more interest to researchers than others, a category refinement mechanism is provided by layering several ART networks in the HABclass tool. This tool is of use in classification tasks where the sequences are aligned and can be represented numerically, where no "correct" grouping is known, where easy adaptability is necessary and where the input space may not be uniformly sampled.

3 Adaptive Resonance Theory

Adaptive resonance theory (ART) (Carpenter & Grossberg, 1990) (Carpenter & Grossberg, 1991a) (Carpenter & Grossberg, 1991b) (Carpenter, Grossberg & Rosen, 1991) describes a class of artificial neural network architectures that use competitive,

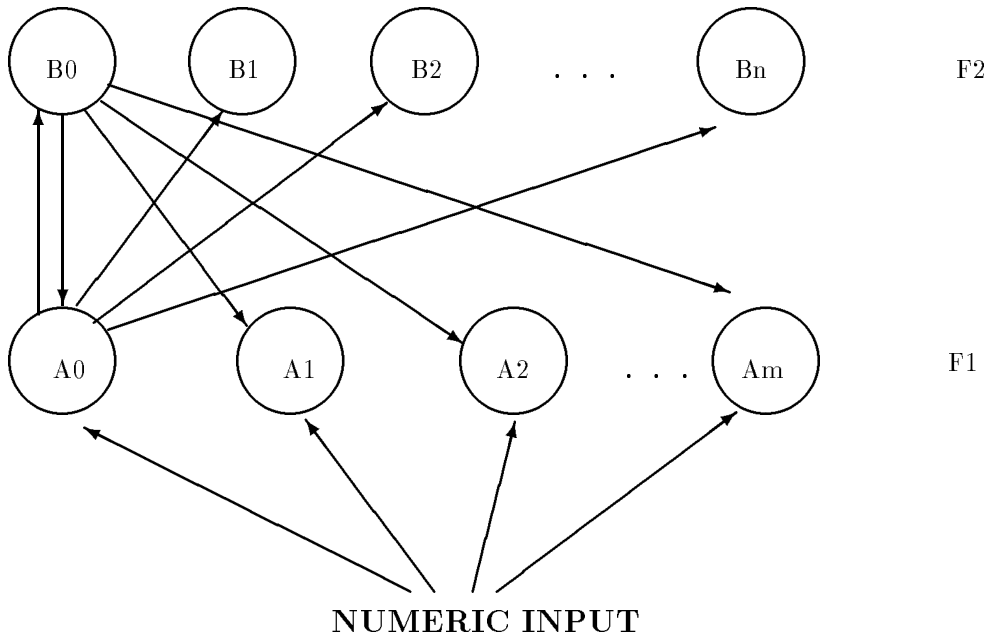

Figure 1. ART2 Network Architecture

unsupervised learning to classify patterns. ART networks work in real-time and their developers have made great strides toward solving the stability-plasticity dilemma, that is, how the network can be plastic enough to learn relevant new information yet remain stable enough to ignore irrelevant new information that would otherwise wash away previous learning (Grossberg, 1987).

ART networks contain two layers of units, F1 and F2. The F1 layer accepts input patterns and is called the feature representation field while the F2 layer will represent the categories learned by the network and is called the category representation field. Each F1 node A_i $(i = 0 \ldots m)$ is connected to every F2 node B_j $(j = 0 \ldots n)$ by means of a bottom-up weight or long term memory (LTM) trace, z_{ij}. Each F2 node B_j is connected to every F1 node A_i by means of a top-down weight or LTM trace, z_{ji}. Figure 1 shows the architecture of an ART2 network although, for the sake of simplicity, not all LTM traces are shown. The bottom-up LTM traces encode the input patterns while the top-down LTM traces encode learned expectations (Carpenter & Grossberg, 1991a). A vigilance parameter, ρ, controls the similarity of the patterns that will be placed into a particular category in F2.

When presented with a pattern to be classified, the network will search the category representation field, F2, for a potentially matching category. The network will measure the similarity between the input pattern and the previously learned expectation of the category node. If the similarity is good enough, that is greater than ρ, the learned expectation will be changed to incorporate the input pattern. If the similarity is not good enough, F2 is searched for a better match (Carpenter & Grossberg, 1991a).

There will be one F2 layer node for each possible category in the application. Each input pattern is encoded in exactly one category in the network. Because competitive learning is used, generic ART networks are not applicable for use in creating hierarchical categorization, that is, for forming subcategories within categories. (Also see Carpenter& Grossberg, 1990.)

4 The Technique

The overall goal of this adaptive method is to find a superior classification network for categorization of a set of DNA sequence fragments in order to analyze the similarities and differences among fragments in those categories and to serve as a classification instrument for new sequences of unknown origin. The overall technique used in HABclass is to start by presenting a set of training sequences to an ART-based network with a vigilance parameter set relatively low. From this first network, a set of categories is obtained that places fragments that are fairly similar into the same category. Some of these categories will be targeted for further breakdown, and new ART-based networks are trained for each of these categories. This second network runs on only the sequences in the category of interest and will create a set of subcategories for each category of interest. The set of categories and subcategories, taken collectively, is then evaluated to rate the performance of the classification network. Figure 2 shows the HABclass network tool. This method may be applied a number of times to the same set of data, with presentation order, vigilance parameters, etc., changed. A testing methodology is described below which allows choice of the best classification network among the possibilities.

4.1 Data Representation

In our data set [1], each DNA fragment is an aligned sequence of 106 nucleotides, i.e., a sequence of symbols from the set {`A, C, T, G, blank` (for skips) }. Since ART networks require numeric input, each nucleotide in the sequence is transformed into a series of three zeroes and one one, where the position of the one designates which nucleotide resides at that position in the nucleotide sequence. For this data, an `A` is encoded as 1000, a `G` as 0100, a `T` as 0010, and a `C` as 0001. A skip (deletion) in a particular position is encoded as 0000. Therefore, each sequence of 106 nucleotides will be represented as a sequence of 106*4=424 zeroes and ones. The F1 layer of each of the layered ART-based networks will have 424 nodes.

[1] The DNA sequences have been deposited in GENBANK under accession number U02590 to U02594 and U02731 to U02875.

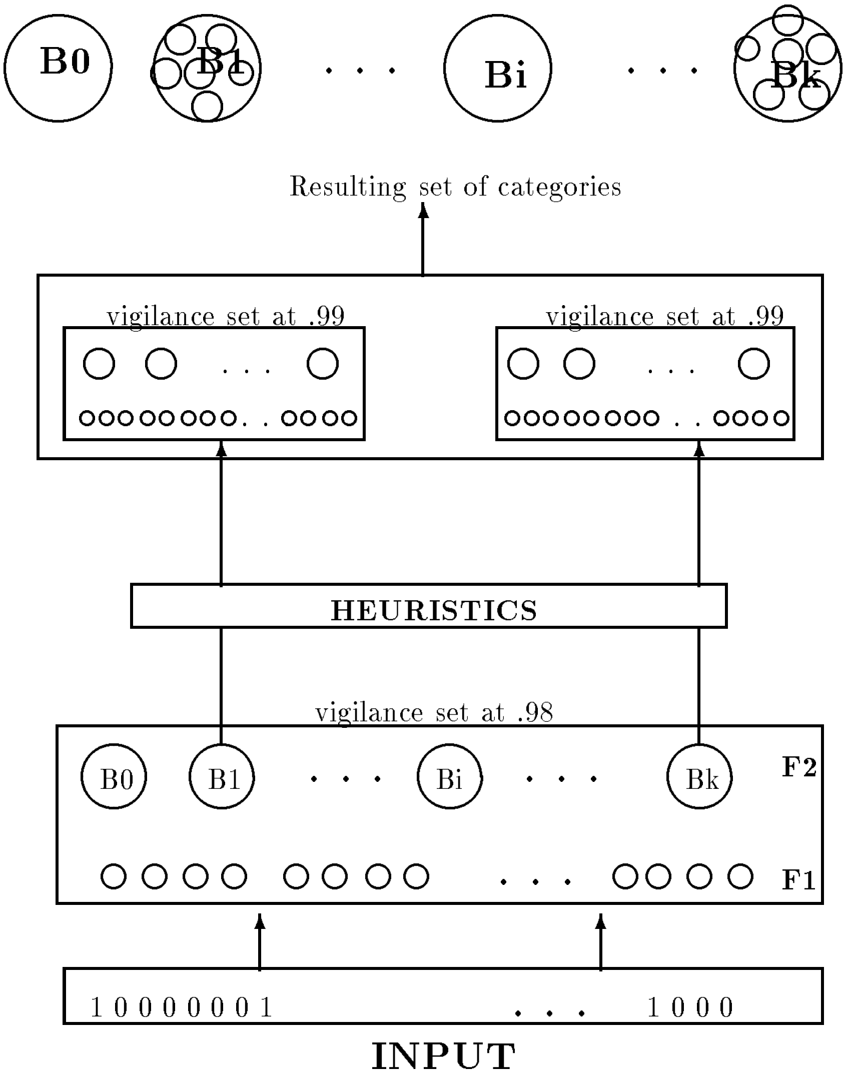

Figure 2. The HABclass Network Architecture

4.2 The Algorithm

Initially, there are p sequences which are used to train the first ART-based network, N_0. These sequences are presented to N_0 using slow learning (Carpenter & Grossberg, 1991a) until the network has stably categorized all p sequences. The vigilance parameter, ρ_0, for N_0 is set relatively low for the application domain so that sequences which are "relatively similar" will be clustered into one category. Once A_0 has been trained, there will be a resulting set of k categories, Ω. Note that $k \leq p$ and that if $k = p$, each sequence has been put into its own individual category, indicating too high a setting for vigilance.

The next step is to separate Ω into two subsets, Ω_I, the "interesting" categories and Ω_U, the "uninteresting" categories, for purposes of further sub-categorization.

The tagging of interesting categories is currently performed using a heuristic combination of the number in each class (with a tendency to target larger classes for further subdivision) and the within category variation, M, described in the Evaluation section below (categories with elevated variability may be targeted for further breakdown). Ω_I now contains c categories, where $0 \leq c \leq k$, and for $i = 1 \ldots c$, $p_i \leq p - k + 1$, where p_i is the number of sequences in category i.

For each of the categories in Ω_I, a new ART-based network, N_i, $i = 1 \ldots c$, is trained on the p_i sequences using slow learning until stable categorization has occurred. For each N_i, the vigilance parameter, ρ_i, is set greater than ρ_0 from network N_0. In this way, for each category i in Ω_I, a set of subcategories, Ω_{s_i}, is defined.

The result is then a new set of categories to be evaluated for performance in categorizing new DNA sequences:

$$\Omega_{New} = \Omega_U \bigcup (\bigcup_{i \epsilon I} \Omega_{s_i}).$$

Iterative application of the process can be used to yield even more finely discriminated categories.

4.3 Evaluation

There are two goals in evaluation of the category sets produced by HABclass networks. The first is to determine whether or not a set of categories derived using this technique is truly significant. Since we are using an ART-based mechanism, different runs through the data (with different presentation orders, for example) may produce different classification networks. The second evaluation goal is then to determine which among several possible classification networks is the best.

A modified F statistic is used to evaluate whether the sum of the square errors for some classification network is significant or not (Zimmerman, et al., 1994). For a particular set of p sequences, we first determine a consensus sequence pattern, $Y_{*,*}$, which uses a positional nucleotide mode of the observed frequency distribution of symbols. That is, we determine which nucleotide occurs most frequently in each position of the sequences and use that nucleotide in the consensus sequence. For example, if there are ten sequences for which we are attempting to determine a consensus sequence pattern and eight of those sequences have an `A` in the first position while two have a `T`, the consensus pattern would have an `A` in the first position. In the case of ties in the frequency count for two nucleotides in a particular position, we randomly choose one of the two nucleotides for the consensus pattern.

Next, a consensus pattern, $Y_{i,*}$, $i = 1 \ldots k$, is formed for each category containing two or more patterns, using the same method described above (singleton categories are disregarded). The variation among categories, T, is

$$T = \sum_{i=1}^{k} p_i [d(Y_{i,*}, Y_{*,*})^2] \tag{1}$$

where p_i is the number of patterns in category i, $d(W, Z)$ is the Hamming distance between sequence W and sequence Z, and k is the number of categories. The variation within categories, M, is

$$M = \sum_{i=1}^{k} \sum_{j=1}^{p_i} [d(Y_{i,j}, Y_{i,*})^2] \tag{2}$$

where $Y_{i,j}$ is the j^{th} sequence of the i^{th} category. The test statistic, F, is given by

$$F = \frac{T/(k-1)}{M/(p-k)}. \tag{3}$$

To determine whether a particular set of categories is significant, we generate a set of random groupings (with the number of categories, k, fixed) and calculate the modified F statistic for those random groupings. The number of F values greater than the F value for this categorization is calculated. The ratio of this number to the total number of random groupings used is taken as the probability of observing a test statistic larger than that for this categorization (i.e., the p-value for the test of the null hypothesis of no difference in the classification networks) (Zimmerman, et al., 1994).

The above method can only be used to compare networks that result in the same number of categories. Since the ART-based networks generate varying numbers of categories depending upon, among other things, the level of the vigilance parameter, a different method must be used to compare categorizations from various HABclass networks. Suppose we want to compare two sets of categories, A and B, each determined by different runs of the hierarchical ART network classification method described above. We first calculate the within category variation, M, for both A and B. Call these M_A and M_B. Both A and B have a certain number of categories, k_A and k_B where $k_A \neq k_B$. We will then run the randomization test described above for A. For each of the random groupings, we will calculate the within category variation, M.

Examination of the observed values of M for each such random trial has shown that the distribution of the observed M under the randomization process is well approximated by a chi-square distribution. The degrees of freedom (df) for this approximation are estimated by the method of moments, i.e., the degrees of freedom are taken to be the nearest integer to the mean of the observed values of M.

The process is then repeated for categorization B. Suppose that $M_A > M_B$. To determine whether M_A is significantly greater than M_B, use the ratio

$$\frac{M_A/df_A}{M_B/df_B}$$

compared to values from a standard F table to check for significance. It should be noted that this technique for assessment of what is "best" is in the nature of a heuristic statistical procedure rather than a formal statistical test.

5 Results

The techniques described above were tested on 107 sequences, each containing information on 106 nucleotides (Zimmerman, et al., 1994). Tests for significance of the sets of categories from HABclass networks yielded excellent results, with significance levels typically at 0.0001 or smaller. These levels of significance are similar to those reported for the manually-derived categorization in (Zimmerman, et al., 1994); significance of the categorization resulting from use of HABclass is clear.

In comparing the performance of two classification networks, the testing methodology is illustrated by the following example test: Setting the vigilance parameter to .97, the top-level HABclass network yielded eight categories for the inputs, with four singletons falling out. The within category variation for this clustering was $M_A = 3113$. Running 1000 permutations of the random groupings results in an average within cluster variation $M_{A_{ave}} = 13720.617$, therefore the degrees of freedom for this grouping are estimated as $df_A = 13720$.

With the vigilance parameter set to .98, the input sequences fell into nineteen categories with $M_B = 794$, with nine singletons falling out. Running 1000 permutations of the random groupings results in $M_{B_{ave}} = 12143.722$ and $df_B = 12143$.

To determine whether the error in network A is significantly greater than the error in network B , look at the ratio

$$\frac{M_A/df_A}{M_B/df_B} = 3.46995.$$

Looking in a standard F test table, we find this ratio is significantly large at the 0.005 level of significance and we can therefore say that the evidence indicates that network B is better than network A in the way it categorizes these sequences.

The results from testing HABclass with a variety of vigilance levels are reported in Table 1. Note that all two-way tests of the networks at vigilance levels adjacent to one another in the table indicate that the network which uses the higher vigilance level yields an error which is significantly smaller than that of the lower vigilance level (with p-value < 0.005). A more appropriate multi-way test for differences is under study. Note that the effect that raising the vigilance level has on classification error is somewhat predictable: the absolute minimum error can be obtained by forcing each sequence into its own individual category. This effect must be offset by other factors including a goal of keeping the number of singletons small and the number of categories significantly less than the number of sequences being categorized.

The effect of varying the presentation order during training is another interesting aspect of this problem. Our data was taken from four isolates, labeled Forest, Savannah, Brazil, and Guatemala. Table 2 shows the effect of presenting the isolates in five different orders. Tests of significance indicated that Presentation orders 1 and 5 yielded classification networks which had significantly lower error than Presentation orders 2, 3, or 4. Getting significant differences in such a test indicates that presentation order does need to be taken into account, and that choosing the best network

Vigilance Level	Number of Categories	Number of Singletons	Error/ df
.95	2	0	.42023
.97	8	4	.22690
.98	19	9	.06539
.99	32	23	.02565

Table 1: Effect of changing vigilance level

among those resulting from several runs of this technique with differing presentation orders of the data is a good first step.

The hierarchical-network feature of HABclass has also been shown to be effective for the purposes outlined. A typical use of the complete HABclass methodology is reflected in the following experiment. Using presentation order 1 (Table 2) and a vigilance level of $\rho = 0.98$ the ART-based network yielded 19 categories, 10 of which were singletons. Three of the non-singleton categories, categories B0, B1, and B5, were singled out as "interesting," due in part to the large number of individual sequences which ended up in these categories. The sequences from these three categories were run through ART again with $\rho = 0.99$. Table 3 shows the set of categories that resulted from this complete run of HABclass. [2]

Comparing this to "flat" networks coming from application of ART with vigilance set at 0.98 and at 0.99, analysis shows that the error from HABclass is significantly lower than that of the 0.98 flat network, and significantly higher than that of the 0.99 flat network (both with p-values < 0.005). However, HABclass yielded 13 categories and 14 singletons, whereas the 0.99 flat network yielded 32 clusters with 23 singletons. The error analysis currently in use is definitely skewed to favor a larger number of categories (which also tends to be associated with a higher number of singletons), and this must be balanced with the desirability of maintaining a reasonable limit on number of categories (and keeping the number of singletons relatively low) to aid in realistic classification.

6 Conclusion

We have developed an adaptive tool for DNA sequence classification which uses adaptive resonance theory networks to form a hierarchical categorization system. The HABclass tool is fast and easy to use. It allows hierarchical classification of sequences, allowing the user to target certain categories for finer breakdown while

[2] Individuals beginning with `Z, M,` or `E` are from the Savannah isolate, those beginning with `B` are from Brazil, those beginning with `L or 1` are from the Forest isolate, and those beginning with `G` are from Guatemala.

Presentation Order		Number of Categories	Number of Singletons	Error	Degrees of Freedom (Est.)	Error / df
1	Forest Savannah Brazil Guatemala	19	10	786	12114	0.06488
2	Savannah Brazil Guatemala Forest	18	11	1475	12538	0.11764
3	Guatemala Forest Brazil Savannah	16	10	1682	12796	0.13145
4	Savannah Brazil Forest Guatemala	18	9	1469	12471	0.11779
5	Brazil Savannah Guatemala Forest	20	11	763	12028	0.06344

Table 2: Effect of presentation order (vigilance = 0.98)

holding others to a somewhat looser similarity constraint. This tool also provides a statistically-based heuristic method for evaluating the performance on various classification networks. Work to completely automate the analysis, including identification of "interesting" categories to target for further breakdown in the hierarchical network, testing multiple network results to determine significant differences in performance, and a good way of handling singletons in the analysis, is ongoing.

Experiments indicate that the method introduced here can result in promising categorization of data in a fraction of the time traditional heuristic methods would have taken. In addition, adding new sequences to those already learned by the network will not require retraining on all previously learned sequences. This adaptive tool holds tremendous promise for easing the time and computational burden of many tasks in DNA sequence analysis.

7 Acknowledgements

This work was supported in part by the National Institutes of Health (NIAID AI33008).

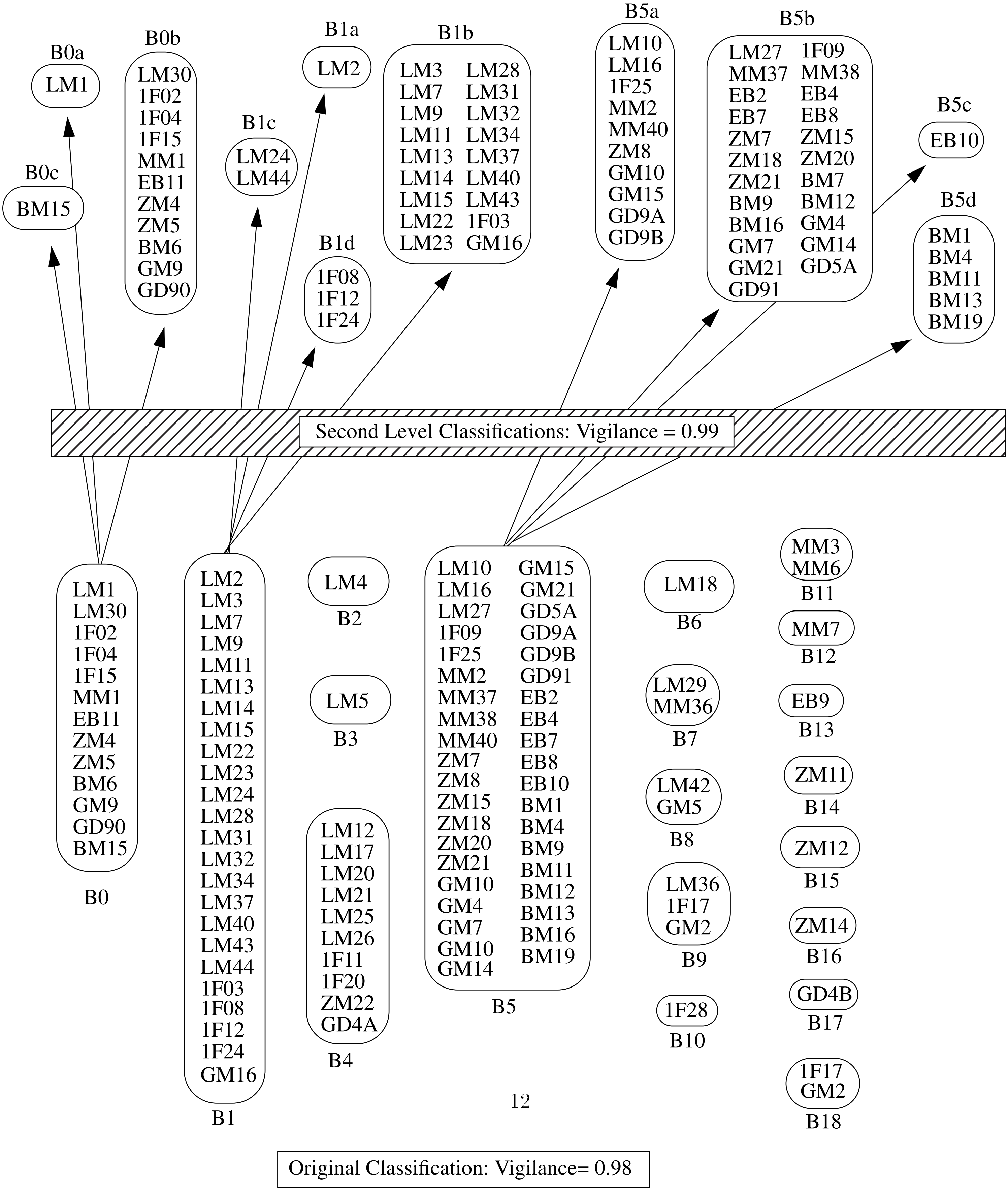

Table 3. Sample Run of the HABclass Network Tool

8 References

Gail A. Carpenter and Stephen Grossberg. 1990. ART3: Hierarchical search using chemical transmitters in self-organizing pattern recognition architectures. *Neural Networks*, 3:129–152.

Gail A. Carpenter and Stephen Grossberg. 1991a. ART2: Self-organization of stable category recognition codes for analog input patterns. In Gail A. Carpenter and Stephen Grossberg, editors, *Pattern Recognition by Self-organizing Neural Networks*, chapter 12, pages 397–423. MIT Press, Cambridge, Massachusetts.

Gail A. Carpenter and Stephen Grossberg. 1991b. A massively parallel architecture for a self-organizing neural pattern recognition machine. In Gail A. Carpenter and Stephen Grossberg, editors, *Pattern Recognition by Self-organizing Neural Networks*, chapter 10, pages 313–382. MIT Press, Cambridge, Massachusetts.

Gail A. Carpenter, Stephen Grossberg, and David B. Rosen. 1991. ART2-a: An adaptive resonance algorithm for rapid category learning and recognition. *Neural Networks*, 4:493–504.

Edgardo Ferran, Pascual Ferrara, and Bernard Pflugfelder. 1993. Protein classification using neural networks. In *Proceedings of the First International Conference on Intelligent Systems for Molecular Biology*, 127-135. AAAI Press, Menlo Park, California.

Stephen Grossberg. 1987. Competitive learning: From interactive activation to adaptive resonance. *Cognitive Science*, 11:23–63.

Nomi Harris, Lawrence Hunter, and David J. States. 1992. Mega-classification: discovering motifs in massive datastreams. In *Proceedings of the Tenth National Conference on Artificial Intelligence*, 837-842. AAAI Press, Menlo Park, California.

Albert Nigrin. 1993. *Neural Networks for Pattern Recognition.* MIT Press, Cambridge, Massachusetts.

Robert J. Schalkoff. 1992 *Pattern Recognition: Statistical, Structural and Neural Approaches.* John Wiley and Sons, Inc., New York.

B. Thylefors. 1978. Ocular onchocerciasis. *Bulletin of the World Health Organization*, 56:63–72

Peter Zimmerman, Charles R. Katholi, Michael C. Wooten, Naomi Lang-Unnasch, and Thomas Unnasch. 1994. Recent evolutionary history of American *Onchocerca volvulus* based on analysis of a tandemly repeated DNA sequence family. *Molecular Biology and Evolution.* Forthcoming.