Region Identification on a Trained Growing Self-Organizing Map for Sequence Separation between Different Phylogenetic Genomes

Johannes Reinhard[1] **Chon-Kit Kenneth Chan**[2]
Saman K. Halgamuge[2] **Sen-Lin Tang**[3] **and Rudolf Kruse**[1]
[1]*Department of Computer Science, University of Magdeburg, Germany*
[2]*Mechatronics Research Group, Department of Mechanical and Manufacturing Engineering, The University of Melbourne, Australia*
[3]*School of Veterinary Science, The University of Melbourne, Australia*
Email: joreinha@cs.uni-magdeburg.de, ckkc@mame.mu.oz.au, saman@unimelb.edu.au, tangsl8@hotmail.com and kruse@iws.cs.uni-magdeburg.de

ABSTRACT: The Growing Self-Organizing Map (GSOM), an extended type of the Self-Organizing Map, is a widely accepted tool for clustering high dimensional data. It is also suitable for the clustering of short DNA sequences of phylogenetic genomes by their oligonucleotide frequency. The GSOM presents the result of the clustering process visually on a coloured map, where the clusters can be identified by the user. This paper describes a proposal for automatic cluster detection on this map without any participation by the user. It has been applied with good success on 20 different data sets for the purpose of species separation.

1 INTRODUCTION

Genome sequencing becomes one of the most important approaches for understanding environmental microorganisms at molecular level. However, sequencing the genomes of environmental microorganisms is difficult because they are uncultivable. One possible solution is to directly isolate DNAs from environmental samples, ligate these predigested DNAs into suitable vectors and transform them into bacterial clones. Then the whole genome shotgun sequencing can be applied to these clones, as carried out by Venter et al. [1]. However, a drawback of this strategy is that the minority groups of microbacteria in the sample present much less DNA. This will cause many fragmented and unclassified sequences since there is a poor coverage among them and no homolog can be found in current databases. In Venter's case, there are 17.7% meaning that 177 million DNA sequences are unclassified [1]. A tool which can classify these sequences with biological meaning will make these immense data useful. The Self-Organizing Map (SOM) is a useful tool for associating the sequences to their correct phylogenetic genomes as previously tested and found by Abe et al. [2]. The SOM could cluster short bacteria sequences by using different nucleotide frequencies as the training features.

In contrast to Abe et al., we chose the Growing Self-Organizing Map (GSOM) instead of the SOM because of its advantages over the SOM. In GSOM, there is no need for predetermined map structure and therefore no prior knowledge on inherent structure of data is needed [3]. This feature of the GSOM is important in our case since for the environmental genome sequencings, there is no prior knowledge available. In addition, GSOM has also been successfully used in bioinformatics [4].

In order to make the SOM-based algorithm practically applicable to cluster species, identification of clusters from the distance map is an important step. The genomic sequences of some species appear overlapped in the cluster boundary of the trained GSOM map. This overlapping between clusters makes the distance map not so obvious to visually identify the clusters. However this task was not addressed before. We present an add-on process to the SOM-based algorithm for species separation by automatic identification of the clusters on the distance map.

We proposed a new three-step approach to this sequence separation. The next section will describe these 3 steps. The experimental results for this approach are presented in section 3. Finally the discussion section will conclude this paper.

2 METHODS

In this section, the three-step approach to sequence separation is described. The first subsection shortly explains how to preprocess the DNA sequences so that GSOM can process them. In the second subsection, the GSOM algorithm is described. The third step comprises the step of automatic region identification and is described in detail.

2.1 Preprocessing

The data available for species separation are short DNA sequences. The sequences are preprocessed by a method described by Abe et al. [2] to provide the nucleotide frequency. The nucleotide frequency is determined by moving a sliding window of a fixed size ($n \in \{2,3,4\}$) along the sequence. It starts at the beginning of the sequence and moves base by base towards the end. Along the way, it is counted how often each of the 4^n permutations (with repetition) occurs in the sequence. The frequency vectors generated with a window size of 2, 3 and 4 are denoted as di-, tri- and tetranucleotides respectively. The vectors are 16-, 64- and 256-dimensional respectively and have to be normalized to a range between 0 and 1 as required by GSOM.

2.2 The Growing Self-Organizing Map

The Self-Organizing Map as well as its extended version, the Growing Self-Organizing Map, is known as a neural network used for clustering high dimensional data. This is achieved by projecting the data onto a two or three dimensional feature map with lattice structure and every lattice point representing a neuron or a node in the map. The mapping preserves the data topology, so that similar samples in the input space can be found quite close to each other on the grid map. The process of clustering can be described as a training procedure, in which a sample is presented to the neural network and a 'winning' neuron, which has the smallest distance to the presented sample, is identified. This 'winning' neuron and its neighbours adapt to the sample by a certain amount of distance. After several training epochs (epoch = each sample has been presented once), the clusters will be formed in the neural network structure.

One problem of the SOM is that the structure (i.e., length and width) of the lattice has to be defined before training. Therefore, prior knowledge is needed or the structure must be found by several trials. The GSOM overcomes this limitation by dynamically adjusting the structure during training. It creates new nodes at the edge of the map whenever a node has to represent a part of the input space with too many and too different samples. The size of the map can be controlled by a parameter, the *Spread Factor* ($\in [0,1]$). A higher Spread Factor results a bigger map [3].

The GSOM training can be divided into three steps: An initialization of the map, a growing phase and a smoothing phase. We chose hexagonal topology which is known to have better topology preservation than a rectangular grid [5].

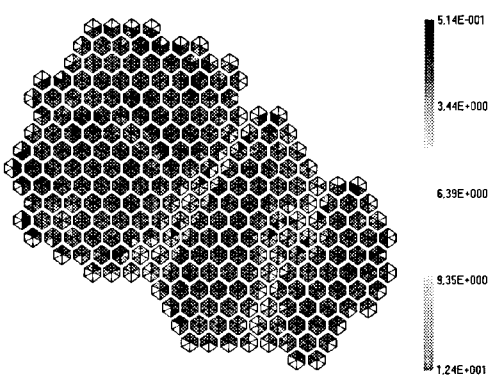

Figure 1: Gray scaled image of a (in practice coloured) Distance Map (3 species, set 3)

After training, the trained GSOM can be displayed as a *distance map* (as shown in Figure 1). The distance map visualizes the distance of a node's weight vector to each of its six neighbour nodes. Therefore the hexagon is divided into six triangles and each triangle symbolizes the distance between the node to whose hexagon it belongs and the node with whose hexagon an edge is shared. The distance value is displayed by colour code, whereas the coherence between the colour and the scalar distance values is displayed on a scale, located on the right hand side of the map.

2.3 Automatic Cluster Detection

For the following descriptions of the automatic cluster detection algorithm, two terms are introduced, which are quite descriptive and thus will help in the understanding of the presented approach. The term *region* is defined as a *cohesive area of hexagons which have low distance values between nodes inside this area.* Due to the topology preserving nature of the GSOM, close data in the high dimensional input space (which are likely to belong to the same cluster) are projected on close (i.e., cohesive) areas. Furthermore the distance values inside this area are low compared to the distance value at a border between two areas on the map, which in fact should represent two different clusters in the input space. Hence, the detection of *clusters* in the high dimensional input space corresponds to the detection of *regions* on the two dimensional surface of the distance map.

The second term introduced is the *centre* of a region which is defined as: *A hexagon belongs to the centre of a region, if it has a relatively very low distance to its surrounding hexagons.* As it is difficult to assign the hexagons in the overlapping zones, our approach begins with assigning the hexagons in the centre of a region which are not affected by any overlapping and then deals with the more difficult assignable hexagons in the border zones later.

The basic idea of the proposed algorithm is to use the well known graphical algorithm Flood Fill or Region Growing to detect the centre of the regions first. When the seed point and a homogeneity threshold are given to the algorithm, it can detect the homogenous region in respect to that threshold. It works as follows: The flood starts at the seed point and recursively fills around it as long as the homogeneity threshold is not exceeded. By doing this, the algorithm fills a whole region and only stops at the borders which exceed the threshold. In our case, we used the distance value as the homogeneity criteria. Applying this algorithm several times and always starting with a seed point outside the filled region, the map could finally be divided into different regions.

As we have to deal with overlapping clusters whose borders are not clear, this Flood Fill algorithm has a high chance of leakage. Leakage will result in two different clusters being identified as one. Our solution to this problem is to find the centres of the regions first. Finding the centres means applying the Flood Fill with a "tough" threshold value, which only fills hexagons with low distance values (i.e., that are located in the centre) and leaves much space unfilled. The unfilled space will be handled later.

Our approach for automatic region identification is divided in seven steps which are described in detail below.

2.3.1 Creating an average distance map

The algorithm first concentrates on detecting the centre of each region. The definition of the centre demands for a measure for a hexagon's distance to its surrounding hexagons. We chose the mean of the six distance values as this measure. The resulting map, which is referred to as average distance map (Figure 2), only displays one value for each hexagon, instead of six as the 'explicit' distance map does. Using the Flood Fill on the average distance map determines the centre of each region.

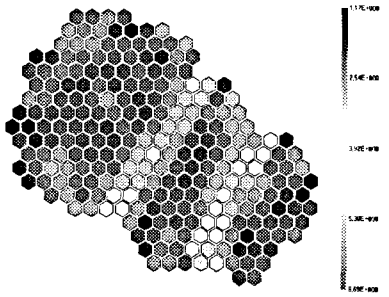

Figure 2: Average Distance Map (3 species, set 3)

2.3.2 Plotting histogram of distance values

The Flood Fill needs a threshold value to be determined. In search for an accurate threshold value for the Flood Fill, the distribution of the distance values is visualized in a histogram. After investigating many of the histograms, it became obvious, that (for a sufficient high Spread Factor value, which we assume for our approach) they all share a common shape (Figure 3). The shape can be described as the sum of two overlapping Gaussian bell-shaped curves: The "left" one with low mean and low variance, the "right" one with higher mean and high variance (Figure 4). One could imagine the left curve as describing the distribution of the low inner region distances, and the right curve as describing the distribution of the higher border distances.

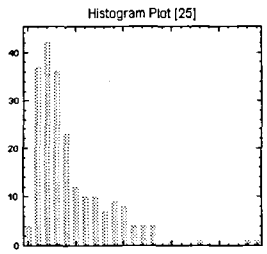

Figure 3: Typical histogram

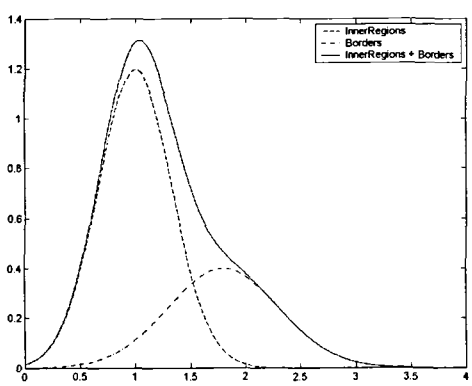

Figure 4: Histogram curve as sum of two overlapping Gaussian bell-shaped curves

The discovered regularity in the shape of the histograms encourages determining the threshold by the histogram. A characteristic point which is also "tough" enough for our purpose is the maximum of the histogram. According to our assumption concerning the Spread Factor for GSOM and the resulting size of map, the hexagons inside a region outweigh the hexagons located at the border. Thus the distance value describing the maximum of the histogram belongs to the inner region distances. It is also strict enough as a threshold for the Flood Fill, as all distances on the right slope of the histogram are ignored, many of which belong to the inner region distances according to the model of two overlapping Gaussian bell-shaped curves. Applying the Flood Fill with the maximum as the threshold results the detection of the centre of a region.

2.3.3 Identifying the maximum in the histogram

The shape of a histogram depends highly on its resolution, determined by the used interval range. As the appropriate interval range is difficult to decide as it depends on many factors and even on the favoured purpose, the maximum is not determined by deciding this issue. Instead our approach works on the raw list of distance values. The list is sorted and could be imagined as a sorted array with the lowest distance value stored at the first index (which is 0 in most programming languages), the highest distance value at the last index. When the distance values are plotted as points on a scaled axis, the maximum of the histogram must be situated at a location where the points are most dense. This location is determined as follows:

(1) Determine, on which half of the scaled axis, the majority of points are located. This is equal to determining the higher bar in a two-bar-histogram.
(2) Decrease the absolute range of the axis by simply cutting off the outer half of the "losing" side. For example, if the left bar is higher (i.e. the winning side), the last fourth of the axis is cut off, which is the outer half of the losing side. In respect of the common shape of the histograms, we assume a monotonously gradient left and right of the maximum. Thus we consider the maximum as not affected by the cut.
(3) Continue at step (1) until the lower bound of the axis equals the upper bound. This distance is considered as the maximum of the histogram.

Tests have shown that the determined maximum value is quite a good choice for many resolutions of the histogram. However it is not always desirable for our purpose. Imagine a situation where the peak of the histogram is located very early and is followed by quite a big number of nearly equal-high bars. All the nodes with these distances will not be filled by the flood, although we might consider them as part of the frequent occurring 'inner region' (thus centre) distances. On the other hand a peak located too late, increases the risk of leakage between regions with unclear borders. Therefore we believe that a smoothed maximum belonging to a bar, located somewhere in the middle of the relatively high bars, represents better our situation, where we are looking for the most occurring distances in respect to all regions. The smoothing is done in a second step, which uses the determined maximum from the previous step as initial value.

(1) As we are determined to find a more appropriate new maximum, the maximum determined by the step before is named the *old* maximum.
(2) A *range* is determined, by subtracting the smallest distance in the whole list which is stored at index 0 from the distance of the old maximum.
(3) The mean value is calculated, considering all the values located in the calculated range around the old maximum.
(4) The deviation between the mean and the old maximum is checked:
    a. If the corresponding index values differ by 2 or less, the algorithm stops returning the mean as the new maximum

b. Else, the range is reduced by 25%, the old maximum is reassigned with the value of the mean and the algorithm continues at step (3)

2.3.4 Flood Fill the centres of the regions

The Flood Fill algorithm is now applied on the average distance map. The distance as the homogeneity criteria and the strict maximum, determined in the previous step as the demanded strict threshold, lead in the detection of the centre of each region. There is one additional condition in the Flood Fill algorithm: Whenever a hexagon didn't gather any hits (empty node), the hexagon is neither filled by the flood, nor chosen as a seed point for the flood. This is done because empty nodes are often at the borders between two regions but nonetheless have quite low distance values and hence have a high risk of being responsible for leakage. The reason for this bad property of the empty nodes is that during growth the GSOM generates new neighbours at all free positions of the growing node. By this, some redundant nodes are created which represent an empty part of the input space.

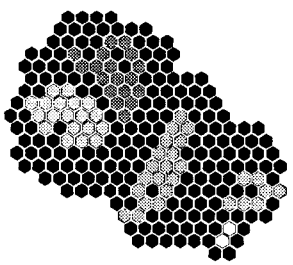

Figure 5: Flood Fill detects 5 centres of regions (3 species, set 3)

However after appliance of the Flood Fill the results are unsatisfying (As shown in Figure 5). Especially for large regions, the Flood Fill often creates several centres where in fact only one should be found. Therefore we developed a method for fusion of similar region centres.

2.3.5 Fusion of similar centre regions using the variance of data

The process of the fusion of similar centres of a region is based on two assumptions:

1. Every region has something like an "ideal" sample that describes best all members of the region.
2. The members of the region have a certain distribution and the standard deviation can be used as a measure of dispersion for this distribution.

In the following, an ideal exemplar for every centre of region is determined, and the deviation between the nodes belonging to the centre of region and the ideal exemplar. The similarity of two region centres is investigated based on these two values. If the similarity is high, the two region centres are assumed to in fact belong to the same region and thus are fused. Mathematically, the ideal exemplar vector is determined by calculating for each dimension the mean of the weight vectors that belong to the centre of the region. The standard deviation of each centre of a region is calculated accordingly. So the result is a vector which contains in each dimension the standard deviation of the weight vectors of the centre from the ideal vector. On the basis of this vector a maximal permitted distance is determined, which defines how far away an ideal exemplar of another centre is allowed to be, to be considered as similar. The maximal permitted distance is calculated by calculating the norm of the standard deviation vector multiplied by a factor, the *fusion greed*. As we chose the Manhattan-distance as distance measure, the norm is defined accordingly as:

$$\|\cdot\| : V \to \Re,\ \|x\| = \sum_{i=1}^{n} |xi| \qquad (1)$$

Experiments show that the value 1.5 as a rule of thumb for the fusion greed gives quite good results for many maps, but is also strict enough not to lead to a fusion of two region centres where there are in fact two region centres. After having calculated the ideal exemplar and the maximal permitted distance for every region, the distances between the ideal exemplars are determined. If the calculated distance between the ideal exemplar of a centre A and a centre B is less than the maximal permitted distance of at least one of the centres, both centres are fused. The described procedure is applied on the resulting centres again and again, until no fusions occur anymore. Figure 6 shows the map after this fusion step.

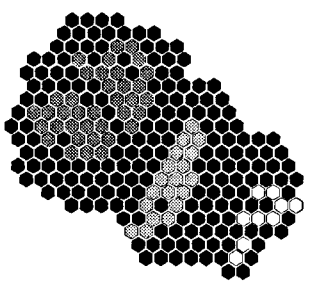

Figure 6: After fusion there are 3 centres left

2.3.6 Assigning the remaining nodes to one of the centre region of clusters

In this step all those nodes are assigned that don't belong to a centre yet. After the assignment procedure all nodes are assigned, so that the evolving areas are no longer be denoted as centres of a region but as regions. The procedure needs the information about the distance of every node to its six surrounding nodes, i.e. the information that are visualized on the explicit distance map. The procedure works as follows: In search for a centre of a region, a remaining node joins the node that is closest it. Both again join the node which is closest to them, and so on, until they finally join a node that belongs to a centre of a region. The group of searching nodes is then assigned to this centre. This final clustered GSOM is shown in Figure 7.

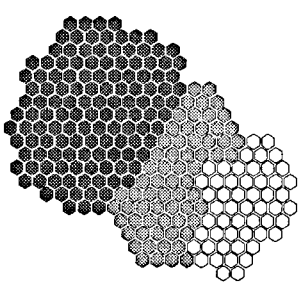

Figure 7: Assignment of the remaining hexagons

2.3.7 Analysing the resulting borders and joining weak-border clusters

In a last step weak-border regions are joined. This step is necessary, because of too less fusions in step 5, due to the strictly chosen fusion greed. If two regions have weak borders they should be joined. Therefore the explicit distance map is analysed and those distances are determined that form the current border between two regions. The mean of these distance values is determined, which provides information about the average strength of the border. Whenever the average strength of a border is less than a threshold, the determination of which is described below, the two regions are joined. The threshold doesn't need to be fixed that strict as for the Flood Fill, because there is no risk of leakage anymore. Leakage could happen, if only *one* border distance is quite low, however the decision whether or not to join two regions is based on *all* border distances.

The threshold is again determined by the shape of the histogram. It should be situated somewhere at the right slope, somewhere around the point where the bars have half the height of the maximum. One characteristic point, which fulfils the demand and is easy to determine, is the inflection point of the Gaussian distribution representing the inner distances (according to our model of the two overlapping Gaussian bell-shaped curves). Considering the assumption of the normal distribution of the inner distances the inflection point can easily be determined if the variance of the distribution is known. We assume that there is negligible overlapping on the left half of the inner distances curve and hence the variance of the distribution can be calculated only using the mean (i.e. the maximum of the histogram) and the distance values left of it. Adding the standard deviation (square root of the variance) to the mean distance, results in the inflection point. Whenever the average strength of a border is below the distance representing the inflection point, the two regions are joined.

3 RESULTS

3.1 Data for Evaluation

To evaluate the proposed approach, we generated 10 sets of 3 species and 10 sets of 10 species. The species were randomly chosen out of a set of 64 species (50 bacteria and 14 archaea) from the NCBI complete genome database [6]. The DNA sequences of these species are cut into non-overlapping short sequences of 10,000 bases. In this experiment, we used the tetranucleotide frequency as the training feature. We preprocess these short sequences as tetranucleotide frequency vectors and used them as the GSOM input vectors. As the DNA sequences of the different species differ in their length, we obtain different number of input vectors for each species.

3.2 Method of Evaluation

We adapt the concept of correctness ratio from Tomida et al. [7] into our application for the evaluation of the clustering results. In the case of fewer detected regions than the number of species in the data set, every region is assigned to one species only. A region will be assigned to the species which has the majority nodes containing this species. If one species is eligible for more than one region, the species will assign to the region which contains the highest number of nodes containing that species. Then other regions will be assigned to the species having the second majority nodes and so forth. In the other case in which the number of species is fewer than the detected region, the assigning process will be continued in a similar way for more than one round of assigning until all regions are assigned to species. After this assigning process, the number of correctly assigned nodes of a particular species will be the number of nodes contained in the region that has been assigned to the species.

We use the name 'Accuracy' instead of correctness ratio for a more meaningful representation in our context. The overall accuracy is determined according to the formula:

$$Accuracy = \frac{\sum_i ci}{\sum_i ni} \times 100 \qquad (2)$$

Where c_i is the number of correctly assigned nodes of species i and n_i is the total number of nodes containing species i.

3.3 Evaluation Results

Each set contains 3 randomly chosen species Training feature: tetranucleotide frequency					
Set number	1	2	3	4	5
Accuracy (%)	88.8	97.9	98.9	92.9	99.5
Set number	6	7	8	9	10
Accuracy (%)	79.0	83.0	99.7	82.0	95.1

Table 1: Evaluation results for the 10 sets with 3 randomly chosen species

Each set contains 10 randomly chosen species Training feature: tetranucleotide frequency					
Set number	1	2	3	4	5
Accuracy (%)	84.3	86.8	78.6	64.3	71.5
Set number	6	7	8	9	10
Accuracy (%)	77.5	71.5	90.2	62.2	83.6

Table 2: Evaluation results for the 10 sets with 10 randomly chosen species

3.4 Interpretation

The results are quite good for some sets (especially the 3-species sets in Table 1), considering the fact that with our method of evaluation 100% accuracy is not achievable if overlapping (i.e., nodes containing more than one species) occurs. However, overlapping occurs only for three (sets 4, 7 & 9 in Table 1) of the 3-species sets and for nine (sets 1-9 in Table 2) of the 10-species sets.

The partly unsatisfying results of the evaluation are due to the strict evaluation method which is based on the assumption that every species is situated in only one region and which evaluates very strictly in case of violation against this assumption. However, this assumption is indeed violated against in many sets. On the one hand, the GSOM using the tetranucleotide frequency as training feature splits several species in two regions. This fact, which was already

observed by Abe et al., is referred to as *intraspecies separation*. Abe et al. ascribes it to the transcription polarity of protein coding sequences [2]. In the case of intraspecies separation only one of the regions (in most cases the bigger one) could be regarded as correctly assigned while the other region's nodes count as detected wrong. On the other hand two closely related species may contain a very similar nucleotide frequency and they may appear entirely overlapped in the map. According to our method of evaluation only one species is allowed to cover a region, hence for one species all nodes are counted as wrongly assigned. The evaluation method however is justified as the algorithm is developed for species separation and the evaluation results show its accuracy for this purpose.

As far as the maps with only 3 species are concerned, only in 4 cases there was no intraspecies separation observed at all. In two cases (sets 1 & 2) the species split. However in set 2, the centres of the split species could be fused by step 5 of our algorithm. This leads to the assumption that distortion occurred for this map. In 4 other cases (sets 4, 6, 7 & 9) intraspecies separation did not result in a split but in a quite clear border inside a region. Our algorithm detects the regions with quite clear inner borders as two different regions.

The maps with 10 species are an even higher challenge for our algorithm: Intraspecies separation with split occurs in almost every map (except set 1) for up to 4 species (only set 4). In addition, the entire overlapping of two species happened in sets 7 & 9. Furthermore the regions are not that big as for the 3-species maps and the algorithm has also to deal with very small and narrow regions. For them, there is no hexagon, which is completely surrounded by hexagons containing the same species samples and therefore such a region does not have a real low distance hexagon in the average distance map. In some cases the strict threshold for the Flood Fill is lower than the lowest distance of that region, hence no centre is detected and the whole region will be assigned to the centre of another region.

4 DISCUSSION

In this paper we introduced an algorithm which automates the process of visual cluster detection on the SOM-based distance map. It has good detection rates for maps with large regions but the detection for maps with small regions could still be improved. Nevertheless the evaluation results don't reflect its accuracy for this task. They evaluate its accuracy for the purpose of grouping species according to their oligonucleotide frequency. As with this feature, an appropriate clustering with only one cluster for each species is not possible, and as the algorithm lacks of a way to detect the split of a species, we receive these little worse results, if splits occur.

To improve the results, research on a more appropriate feature must be done. In this context, we refer to Tyson et al. who suggests a three-step approach for species separation [8].

But improvements can also be done on the proposed algorithm itself. It has difficulties in detecting small regions. This is due to an over-strict threshold for the Flood Fill. During the development and testing of our approach, it shows that the choice of an appropriate threshold is crucial but nonetheless difficult. Therefore our approach disperses the influence of the choice of a threshold to a several-step algorithm which contains the finding of the threshold for the Flood Fill, the Fusion Greed for the fusion and the threshold for joining the weak borders. The setting for these three parameters is important and future work could be done on the determination of more suitable parameters. However we recommend developing a different approach that does not based on one global threshold adapted from the histogram but on several local thresholds (one corresponds to each region). This could improve the results and in addition make the algorithm suitable for other data sets. The current algorithm works only on datasets with similar dispersion of the data inside each cluster. Otherwise the histogram would not have this characteristic shape of two overlapping Gaussian bell-shaped curves but several peaks for each region's inner distances. Future work will concentrate on making the algorithm applicable for such data sets.

REFERENCES

[1] J. C. Venter, K. Remington, J. F. Heidelberg, A. L. Halpern, D. Rusch, J. A. Eisen, D. Wu, I. Paulsen, K. E. Nelson, W. Nelson, D. E. Fouts, S. Levy, A. H. Knap, M. W. Lomas, K. Nealson, O. White, J. Peterson, J. Hoffman, R. Parsons, H. Baden-Tillson, C. Pfannkoch, Y.-H. Rogers, and H. O. Smith, "Environmental Genome Shotgun Sequencing of the Sargasso Sea," *Science*, vol. 304, pp. 66-74, 2004.

[2] T. Abe, S. Kanaya, M. Kinouchi, Y. Ichiba, T. Kozuki, and T. Ikemura, "Informatics for Unveiling Hidden Genome Signatures," *Genome Res.*, vol. 13, pp. 693-702, 2003.

[3] D. Alahakoon, S. K. Halgamuge, and B. Srinivasan, "Dynamic self-organizing maps with controlled growth for knowledge discovery," *Neural Networks, IEEE Transactions on*, vol. 11, pp. 601-614, 2000.

[4] A. L. Hsu, S.-L. Tang, and S. K. Halgamuge, "An unsupervised hierarchical dynamic self-organizing approach to cancer class discovery and marker gene identification in microarray data," *Bioinformatics*, vol. 19, pp. 2131-2140, 2003.

[5] T. Kohonen, *Self-Organizing Maps*, 2nd Edition ed. Berlin, Heidelberg, New York: Springer, 1997.

[6] "NCBI Database: http://www.ncbi.nlm.nih.gov/," 2004.

[7] S. Tomida, T. Hanai, H. Honda, and T. Kobayashi, "Analysis of expression profile using fuzzy adaptive resonance theory," *Bioinformatics*, vol. 18, pp. 1073-1083, 2002.

[8] G. W. Tyson, J. Chapman, P. Hugenholtz, E. E. Allen, R. J. Ram, P. M. Richardson, V. V. Solovyev, E. M. Rubin, D. S. Rokhsar, and J. F. Banfield, "Community structure and metabolism through reconstruction of microbial genomes from the environment," *Nature March*, vol. 428, pp. 37-43, 2004.