International Journal of Bio- Science and Bio- Technology
Vol. 1, No. 1, December, 2009

A New Generalized Growth Threshold for Dynamic SOM for Comparing Average Mutual Information and Oligonucleotide Frequency as a Species Signature

Chon-Kit Kenneth Chan and Saman Halgamuge

MERIT theme Biomedical Engineering, Melbourne School of Engineering,
The University of Melbourne, Australia
cckenneth@gmail.com, saman@unimelb.edu.au

Abstract

The average mutual information (AMI) known from information theory has been reported as a strong genome signature in some literature and we have reported the use of oligonucleotide frequencies as a genome signature. In this work we improve the use of AMI as a training feature for Growing Self Organising Maps (GSOM). Although the range of k is considered as an important parameter in AMI, no standard range for k is proposed. Our first contribution is to introduce a new growth threshold (GT) for GSOM and use it to identify the best range of k for clustering prokaryotic sequence fragments of 10 kb. We then, compare the results using the best k range of AMI against our previously published results using oligonucleotide frequencies. These experiments showed that the newly proposed GT equation makes GSOM able to efficiently and effectively analyse different data features for the same data. The results also emphasize our use of oligonucleotide frequencies as opposed to AMI.

Keywords: *GSOM, AMI, Species Separation, Prokaryotic Sequences.*

1. Introduction

Average mutual information (AMI) is a well-known measure of dependence of two variables in information theory, which has increasingly been used for analysing DNA sequences. Grosse et al. [1] showed that the probability distributions of AMI are significantly different in coding and non-coding DNA; Slonim et al. [2] used AMI to study the relationships between genes and their phenotypes; and Swati [3] applied a mutual information function to quantify the similarities and differences between bacterial strains. The investigation of AMI on DNA sequences has been also extended to DNA fragments. Otu and Sayood [4] revealed that fragments coming from the same regions of the target sequence have similar AMI profiles. Bauer et al. [5] found that the AMI profile could separate the fragments, whose sizes vary between 200 base pair (bp) and 10 000 bp, of two eukaryotes and claimed that the AMI profile can be used to discriminate between DNA fragments from different species. This growing evidence supports the hypothesis that AMI may also be able to distinguish DNA fragments according to their phylogenetic relationship. Therefore, this paper investigates the use of AMI as a training feature in Growing Self-Organising Maps (GSOMs) for separating DNA sequence fragments, and directly compares the results with AMI to the results with oligonucleotide frequencies obtained in our previous work [6] on the same datasets.

In the first development of GSOM, Alahakoon et al. [7] claims that the spread factor (SF) introduced to the Growth Threshold (GT) equation in GSOM is independent of the dimensionality of the data and that the same SF can be used to generate maps for different dimensionalities. In our previous experiments [6], the same SF was tested for datasets with different dimensions. However, a significantly lower map resolution was observed as the order of oligonucleotide frequency (i.e. dimensions) was increased. In order to rectify such unexpected behaviour of GSOM in the early tests, the SFs were experimentally determined for datasets with different oligonucleotide frequencies to achieve similar map resolutions for comparison. Although the experiments were completed successfully by spending more time and effort to experimentally determine the same map resolution for datasets with different dimensionalities, it cannot be repeated for the analysis of large datasets, for which a large computing time is required. This problem raised a question in the accuracy of the original GT equation. In this paper we propose a new GT equation generalised to a wider range of distance functions, as well as rectifying the problem of different dimensionalities so that the intended purpose of introducing SF in [7] can be achieved. Then the proposed GT equation is applied to investigate the AMI for DNA sequence separation.

The remainder of the paper is organised as follows: Section 2 describes the background of GSOM and present the new generalised GT equation. Section 3 shows the results generated using AMI as a training parameter to GSOM. Finally, Section 4 provides a summary and conclusion of this paper.

2. Backgrounds

Growing Self-Organising Map (GSOM) [7] is an extension of Self-Organising Map (SOM) [8]. GSOM also considers as dynamic SOM overcomes the weakness of the need for user defined static map structure of SOM. Both SOM and GSOM are used for clustering high dimensional data. This is achieved by projecting the data onto a two or three dimensional feature map with lattice structure where every point of interest in the lattice represents a node in the map. The mapping preserves the data topology, so that similar samples can be found close to each other on the 2D/3D feature map.

2.1. The problem in the original Growth Threshold (GT) equation in GSOM

While SOM uses a pre-specified number of nodes and map structure in training, GSOM uses a threshold to control the spread of a map in the growing phase so that the resolution of the map will be proportional to the final number of nodes generated at the end of the growing phase. This threshold can be any constant value and should be inversely proportional to the amount of detail in the hidden data structure that the user would like to observe in the map. Nevertheless, in order to standardise this threshold into a parameter which is easy to use, Alahakoon et al. [7] introduced the growth threshold (GT) equation to determine the threshold:

$$GT = -D*\ln(SF) . \quad (1)$$

The user conveniently selects SF between zero and one as a standardised referencing measure for the map resolution. The GT equation also includes the dimensionality (D) of input data vectors intended to make it versatile for analysing datasets with different

dimensions while using the same SF as a referencing measure. However, this was not observed in our previous experiment [6].

In the first development of GSOM, Euclidean distance was used as the similarity measure. When an input vector is compared with a node, the accumulated error of the node is increased by the distance difference, which is defined as:

$$dist(\mathbf{x},\mathbf{w}) = \left(\sum_{d-1}^{D} |x_d - w_d|^2 \right)^{\frac{1}{2}},$$

where **w** is the weight vector of the node, and **x** is the input vector (**w**, $\mathbf{x} \in \Re^D$ where D is the dimensionality of data).

The growing mechanism in GSOMs depends on the comparison of the accumulated error in a node and the GT value, which is determined prior to the start of training. In order to achieve the same spread for different dimensionalities, when the GT value is increased with increased data dimensionality, the generated error should also be increased proportionally to the increment of the GT value. However, this is not the case in the original GT equation. A simple example can effectively illustrate this problem. Let us consider GT = GT_D where D indicates the number of dimensions. To train a 1D dataset with SF=0.1, the original GT equation gives GT_1=2.3. Using the standard practice in artificial neural networks that all dimensions are normalised to values between zero and one, the maximum possible error (maxErr) is one, as shown in Figure 1(a). However, for a 2D dataset (Figure 1(b)) using the same spread factor (SF=0.1), the GT value is doubled to GT_2=4.6 but the maximum possible error is only $\sqrt{2}$ which is less than double the maximum possible error in the 1D case: GT_2 = 2*GT_1, but maxErr_2D < 2*maxErr_1D. The disproportion between GT value and generated error appears whenever the dimensionality of data is changed. Consequently, the resultant map will be smaller for dataset with a higher data dimensionality.

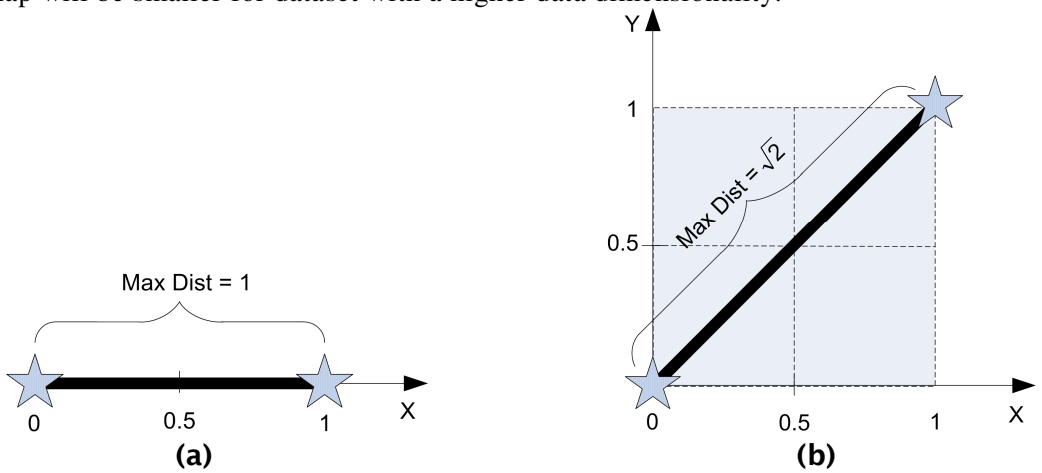

Figure 1. Illustration of maximum Euclidean distance. (a) 1D, (b) 2D input space. Stars represent the vectors locating in the corresponding normalised input space.

2.2. A generalised GT equation

The GT equation is directly related to the generated errors which depend on the distance function used. Hence, the GT value is implicitly linked to the distance function. The Minkowski distance, a general form of some other distance functions, is chosen to be used in the modification.

Considering the Minkowski distance function:

$$dist(\mathbf{x}) = \left(\sum_{d=1}^{D} |x_d|^p \right)^{\frac{1}{p}}, \tag{2}$$

where p is the order of the distance. If it is multiplied and divided by $D^{1/p}$:

$$dist(\mathbf{x}) = \frac{D^{\frac{1}{p}}}{D^{\frac{1}{p}}} \left(\sum_{d=1}^{D} |x_d|^p \right)^{\frac{1}{p}} = D^{\frac{1}{p}} \left(\frac{\sum_{d=1}^{D} |x_d|^p}{D} \right)^{\frac{1}{p}} = D^{\frac{1}{p}} AVG \tag{3}$$

,

where AVG represents a constant value over all dimensions of data and for a large D, AVG is less sensitive for change in D.

Since GT is related to dist(**x**) and for the GT equation to account for dimensionality, the simplest solution is to make GT proportional to the dimensionality related part in dist(x) assuming D is large and therefore AVG is constant: $GT \propto D^{1/p}$. Then using the same standardising control measure SF as in the original GT equation, the generalised GT equation becomes:

$$GT = -D^{\frac{1}{p}} \ln(SF), \tag{4}$$

where $SF = e^{-AVG}$. Considering that AVG can take values between zero and infinity, we obtain $0 \le SF \le 1$. In this light we observe that the original GT equation was defined to suit only p=1, which is the Manhattan distance, but not the intended Euclidean distance (p=2).

2.3. Average mutual information for DNA sequences

Mutual information, which measures the dependency of two random variables, is originally from the field of information theory. In this paper, the average mutual information (AMI) reported in Bauer et al. [5], is adopted. In a DNA sequence, if X is taken to be the base at location i and Y to be the base at location j, which is k positions downstream from i (i.e. j = i+k), the AMI function (I_k) is defined as:

$$I_k = \sum_{X \in A} \sum_{Y \in A} p_k(X,Y) \log \left(\frac{p_k(X,Y)}{p(X)p(Y)} \right),$$

where A is the set of nucleotides $\{A,C,G,T\}$; p(X) is the marginal probability of X and is defined by dividing the total number of times the nucleotide X occurs by the total number of bases in the sequences; and $p_k(X,Y)$ is the joint probability for the nucleotides occur k bases apart and is defined as:

$$p_k(X,Y) = \frac{n_k(X,Y)}{\sum_{I \in A} \sum_{J \in A} n_k(I,J)},$$

where $n_k(X,Y)$ is the number of times two bases k apart take on the values $X \in A$ and $Y \in A$.

By calculating the AMI for different values of k for a sequence fragment, an input vector for the GSOM training can be created. Each k value will correspond to a single dimension in the input vector and the dimensionality of the vector depends on the number of different k used.

2.4. Quality measurement of the clustering performance in a mixing region

To evaluate a clustering algorithm's ability to group DNA sequence fragments into species-specific or "pure" clusters, we define two criteria that measure the clustering quality in a mixing region: intensity of mix (IoM) and level of mix (LoM), where the former measures the percentage of mixing and the later indicates the taxonomic level of ambiguity for a given pair of clusters [6].

The IoM is evaluated based on the concept of mixed pair described below. Let A and B be sets of vectors belonging to species A and B, respectively, and n(X) is the number of elements in set X. If A and B is a mixed pair, then the percentage of A in the mixing region of the two classes is n(A ∩ B | A)/n(A) and the percentage of B is n(A ∩ B | B)/n(B). For k number of species, there can be up to k (k − 1)/2 mixed pairs. Additionally, a pair of clusters is only considered to be truly mixed when both clusters are heavily overlapped. We use TH = 5% for the threshold of being truly mixed meaning that, statistically, we have a non mixing confidence of 95%. The IoM measures the amount of mixing sequences and it is nonlinearly categorised into five levels: low (L) 5%–10%, medium low (ML) 10%–20%, medium (M) 20%–40%, medium high (MH) 40%–60%, and high (H) 60%–100%.

To evaluate clustering results of species, we use LoM to describe the taxonomic level of the mixed species. Because of the evolution of organisms, nucleotide composition of genomes belonging to the same lower taxonomic levels can be very similar. Clustering organisms at higher level of taxonomy should be easier than at lower level of taxonomy. Therefore, if truly mixed pair occurs, lower LoM (e.g., Species) is more acceptable and more desirable than higher LoM (e.g., Kingdom). In summary, the proposed two measures are defined as

- IoM $\in$ {L,ML,M,MH,H},
- LoM $\in$ {Species, Genus, Family, Order, Class, Phylum, Kingdom}.

The two proposed measures, IoM and LoM, are only defined for truly mixed pairs to evaluate the clustering quality in the mixing regions of a map by the following steps.

(i) Find truly mixed pairs for all pairs of species where if $n(X \cap Y \mid Y)/n(Y) \geq$ TH and $n(X \cap Y \mid X)/n(X) \geq$ TH, then X and Y is a truly mixed pair.

(ii) If X and Y are truly mixed, determine IoM according to $\min\{n(X \cap Y \mid Y)/n(Y), n(X \cap Y \mid X)/n(X)\}$.

(iii) Identify LoM of *X* and *Y*.

Clustering results can now be assessed based on three criteria: number of truly mixed pairs, IoM, and LoM. However, the criterion associated with the higher priority may vary between applications. Therefore, in our assessment, one result is better than another only when it is superior on at least two of the three measures.

3. Results

This work investigates whether the average mutual information (AMI) can be used to separate short DNA sequence fragments, and compares the results generated by the AMI with the results created by oligonucleotide frequencies. The AMI, which was used in Bauer et al. [5], is adopted and summarised in Section 2.3. In order to compare the oligonucleotide frequency results produced in our previous work [6], the same two sets of species genomes were used here. Similar data preprocessing was applied to produce datasets for these experiments (i.e. using a fragment length of 10 kb), except that the input vectors were created by calculating the AMI for a series of k values instead of calculating the oligonucleotide frequencies. For convenience, datasets produced using the k values ranging from X to Y will be denoted as k:X-Y. For example, k:1-100 represents the datasets which were generated using the AMI with k values ranging from 1 to 100.

Different ranges of k values have been used in literature depending on preference, but no standard range of k value is proposed. Therefore, this investigation was also aimed at finding a proper range for k for the task of species separation. To do this, the generalised GT equation can be conveniently applied with the same SF for all different ranges of k values. The same training settings are used as in our previous work [6] (i.e. learning length, learning rate, etc.). SF=0.01 was found to produce a similar resolution to the maps generated for the oligonucleotide frequencies and therefore, it is used in this work allowing direct comparison. As Bauer [5] used the range of k:5-512 in his experiment to successfully separate the short DNA fragments and a longer range of dependencies between two nucleotides are improbable from the biology perspective, such range of k was used here as maximum range of k for the investigation. Four datasets were created for each of the two sets of species genomes. These datasets will be referred as long-range k values in the following discussion. They are: k:1-100, k:5-300, k:201-500 and k:5-512. The evaluation method for the mixing regions, introduced in our previous work [6] was adopted here to evaluate results and a summary is described in Section 2.4.

The results generated by the AMI with the four long-range k values for Set1 and Set2 are tabulated in Table 1 and Table 2 respectively. For convenience, the results generated by tetranucleotide frequency in our previous work [6] were also included in the tables. From these results, the AMI performed very badly for both Set1 and Set2 since there are large numbers of mixed pairs. However, it was noticed that the shorter range of k values, i.e. k:1-100, produced fewer mixed pairs compared to other long-range k values which was observed consistently in both sets. Therefore, it was concluded that even shorter ranges of k values may perform better.

To test the above hypothesis, another five datasets for each of Set1 and Set2 were generated using the short range of k values: k:1-5, k:1-10, k:1-16, k:1-25 and k:1-50. The

results for Set1 and Set2 of each of the five datasets are shown in Table 3 and Table 4 respectively. As expected, these short-range k values provide better results than the long-range k values. For example, the numbers of mixed pairs for any short-range k values in Table 3 are smaller than those for the long-range k values in Table 1, and similarly for Table 4 and Table 2. The best results are shown where k:1-16 for both Set1 and Set2. However, although it shows the best results of all the tested ranges of k values, this range still performs less effectively than the tetranucleotide frequency.

Table 1. Results of using AMI with long-range k values for Set1. 'Tetra' represents using the tetranucleotide frequency as training feature; 'k:X-Y' denotes using the AMI with k ranges from X to Y as the training features. If a specific 'IoM' is more than one, its name is displayed and followed by a colon and number of times it should appear.

	Tetra	**k:1-100**	**k:5-300**	**k:201-500**	**k:5-512**
#OfMix	4	34	41	42	40
Kingdom	--	ML:4, M, L:2	H:3, MH, ML:3, L	H:3, MH, M, ML:2, L	H:3, MH, ML:4
Phylum	--	MH, M:2, ML:3, L:5	H:3, MH:2, M:4, ML, L:3	H:3, MH, M:6, ML:2, L:2	H:3, MH:2, M:4, ML, L:3
Class	--	H, MH, M:3, ML:2, L:2	H:4, MH, M, ML:4, L:2	H:4, MH, M:4, ML:3	H:4, MH, ML:5, M, L
Order	L:2	H, MH, M, ML:2	H, MH:2, M:2	H, MH:2, M:2	H, MH:2, M, ML
Family	--	--	--	--	--
Genus	--	--	--	--	--
Species	ML, L	MH, M	H, MH, L	H, MH, ML	H, MH

Table 2. Results of using AMI with long-range k values for Set2.

	Tetra	**k:1-100**	**k:5-300**	**k:201-500**	**k:5-512**
#OfMix	0	25	43	45	44
Kingdom	--	--	--	--	--
Phylum	--	H, MH:2, M:3, ML:5, L:6	H:6, MH:8, M:8, ML:4, L:2	H:8, MH:11, M:7, ML:3, L	H:8, MH:6, M:8, ML:7
Class	--	MH, M, L:2	H:2, MH, M:3, ML:2, L	H:2, MH:3, M:2, ML:2	H:2, MH, M:3, ML:2, L
Order	--	M, ML:2, L	H, M:3, ML, L	H, M:2, ML:3	H, M:3, ML, L
Family	--	--	--	--	--
Genus	--	--	--	--	--
Species	--	--	--	--	--

Table 3. Results of using AMI with short-range k values for Set1.

	Tetra	**k:1-5**	**k:1-10**	**k:1-16**	**k: 1-25**
#OfMix	4	31	29	27	30
Kingdom	--	ML:4, L:2	ML:4, L	M:2, ML:3	M:2, ML:3
Phylum	--	MH, M, ML:4, L	M:2, ML:2, L:5	MH, M, ML:3,	M:3, L:7

				L:3	
Class	--	H, MH:2, M:3, L:3	H, MH, M:2, ML:2, L:2	H, M:5, L	H, M:4, ML, L:2
Order	L:2	H, MH:2, M, ML:2	H, MH, M, ML:2	H, M:2, ML:2	H, M:2, ML:2
Family	--	--	--	--	--
Genus	--	--	--	--	--
Species	ML, L	MH:2, L	MH:2	MH, M	H, M

Table 4. Results of using AMI with short-range k values for Set2.

	Tetra	**k:1-5**	**k:1-10**	**k:1-16**	**k: 1-25**
#OfMix	0	24	19	19	21
Kingdom	--	--	--	--	--
Phylum	--	MH:2, M:3, ML:6, L:5	H, M:4, ML:3, L:5	MH:2, M:2, ML:4, L:4	MH:2, M:4, ML:2, L:5
Class	--	MH:2, M, L:2	MH, ML:2, L	M, ML, L:3	M, ML, L:4
Order	--	M, L:2	M, L	M, ML	M, ML
Family	--	--	--	--	--
Genus	--	--	--	--	--
Species	--	--	--	--	--

4. Conclusion

This paper identified a key deficiency in the original GT equation through experimental analysis. Consequently, a generalised GT equation was proposed to suit a wider range of distance functions and to give GSOM the ability to analyse datasets with different numbers of dimensions through a single SF value. Then the proposed GT equation was used to effectively investigate the AMI as applied to the separation of short sequence fragments. The long-range k values performed less than the short-range ones, perhaps because the short sequence fragments are not long enough to provide a good estimation for long-range k values [9]. The results showed that k:1-16 performed better than other short ranges of k values, such as k:1-10, and better than the longer ranges of k values, such as k:1-25. This may also be because the short ranges of k values did not provide enough signals; the amount of stored signal was limited by the short length of sequences for the longer ranges of k values. Although the best range of k values (k:1-16) for the AMI could be identified after intensive tests, the results were inferior in comparison to the excellent results achieved by using oligonucleotide frequencies [6]. The poor results for the AMI may be due to the noise from the non-coding region of the fragments, as Grosse et al. [1] showed that the probability distributions of AMI are significantly different in coding and non-coding DNA. Therefore, the results should be improved if only the sequences with sufficient portions of coding DNAs are employed in the clustering.

References

[1] Grosse I., Herzel H., Buldyrev S.V., Stanley H.E.: Species independence of mutual information in coding and noncoding DNA. Phys Rev E J1 - PRE, 61(5), 5624 LP - 5629 (2000)

[2] Slonim N., Elemento O., Tavazoie S.: Ab initio genotype-phenotype association reveals intrinsic modularity in genetic networks. Molecular Systems Biology, 2, (2006)
[3] Swati D.: In silico comparison of bacterial strains using mutual information. J Biosci, 32(6), 1169-1184 (2007)
[4] Otu H.H., Sayood K.: A divide-and-conquer approach to fragment assembly. Bioinformatics, 19(1), 22-29 (2003)
[5] Bauer M., Schuster S.M., Sayood K.: The average mutual information profile as a genomic signature. BMC Bioinformatics, 9(48), doi:10.1186/1471-2105-1189-1148 (2008)
[6] Chan C.-K.K., Hsu A.L., Tang S.-L., Halgamuge S.K.: Using Growing Self-Organising Maps to Improve the Binning Process in Environmental Whole-Genome Shotgun Sequencing. Journal of Biomedicine and Biotechnology, Volume 2008, Article ID 513701, 10 pages. doi:10.1155/2008/513701 (2008)
[7] Alahakoon L.D., Halgamuge S.K., Srinivasan B.: Dynamic self-organizing maps with controlled growth for knowledge discovery. IEEE Transactions on Neural Networks, 11(3), 601-614 (2000)
[8] Kohonen T.: Self-Organizing Maps, 2nd edn, Springer, Berlin, Heidelberg, New York (1997)
[9] Deschavanne P., Giron A., Vilain J., Dufraigne C., Fertil B.: Genomic signature is preserved in short DNA fragments. In: Proceedings of the 1st IEEE International Symposium on Bioinformatics and Biomedical Engineering, pp. 161-167. (2000)

Authors

Kenneth Chan received the BE (Hons.) degree (2003) in mechatronics engineering, the BCS degree (2003) in computer science, and the Ph.D degrees (2008) in bioinformatics, all from the University of Melbourne, Australia.

From 2004 to 2008, he was a part-time guest lecturer, tutor and demonstrator in the engineering department at the University of Melbourne. Since 2009, he has been working as a bioinformatics scientist in Monsanto Biotechnology Research (Beijing) Co., Ltd., China. His research interests are in unsupervised learning, DNA sequence assembly and analysis, and metagenomics.

Saman Halgamuge is a Professor of Melbourne School of Engineering of University of Melbourne, Australia working in research areas of Biomedical Engineering and Sustainability and Energy Systems. He received the B.Sc. degree in electronic and telecommunication engineering from the University of Moratuwa, Sri Lanka, in 1985 and the Dipl.-Ing. and Dr.-Ing. Degrees in electrical engineering from Darmstadt University of Technology, Germany, in 1990 and 1995, respectively. Since 1996 he has been with the Department of Mechanical Engineering, University of Melbourne and also participating in the MERIT Research theme Biomedical Engineering. He is a coauthor of about 200 conference/journal papers and has contributed to books in the areas of data mining and analysis and mechatronics. His research interests are in bio-inspired methods of Pattern Recognition and Optimization focusing on problems in Mechanical Engineering and Bioengineering.