Fuzzy Clustering of Stochastic Models for Molecular Phylogenetics

TUAN D. PHAM[1], DOMINIK BECK[2], and DENIS I. CRANE[3]
[1]School of Information Technology
James Cook University, Townsville, QLD 4811, Australia.
[2] Department of Biotechnology and Bioinformatics
University of Applied Sciences Weihenstephan
Weihenstephan, 85350 Freising, Germany.
[3]School of Biomolecular and Biomedical Science
Nathan Campus, Griffith University, QLD 4111, Australia.

Abstract: A new method for the study of molecular phylogenetics based on fuzzy c-means clustering of Markov models is proposed. This approach is able to cluster whole sequences or genomes into groups whose boundaries overlap, and to reconstruct the phylogenetic trees that graphically describe the evolutionary relationships between organisms. The method is applied to examine the similarities and evolutionary relationships of a large data set of complete mammalian mitochondrial genomes.

Key Words: Fuzzy c-means; Markov models, molecular phylogenetics, systems biology.

1. Introduction

By comparing two or more sequences or genomes one is able to infer the evolutionary relationships between them. The comparison is based on the assumption that two sequences that diverged in the recent past would be expected to be more similar than a pair of sequences whose common ancestor is more ancient. The objective of most molecular phylogenetic analyses is to reconstruct the tree-like pattern that decribes the evolutionary relationships between organisms [4]. A phylogenetic tree reconstruction (PTR) can be carried out from comparative analysis of data such as protein and DNA sequences. However, DNA yields more phylogenetic information than protein because variability in both the coding and non-coding regions of the genome can be examined [4].

Conventionally, a DNA-based phylogenetic tree reconstruction involves the alignment of the DNA sequences so that nucleotide differences can be scored to obtain the comparative data; this is followed by conversion of the comparative data into a reconstructed tree. There are two categories of method for the reconstruction of phylogenetic trees using molecular data: distance-based, and character-based (maximum parsimony) [18]. Distance-based methods convert the sequence information into a distance matrix that show the evolutionary distances between all pairs of sequences in the data set. These evolutionary distances are used to establish the lengths of the branches connecting two sequences in the reconstructed tree.

The earliest described distance-based method for tree reconstruction is the unweighted-pair-group method using arithmetic averaging (UPGMA) [26]. The UPGMA method requires a distance matrix and begins by clustering the two species with the smallest distance, separating them into a single, composite group. After the first clustering, a new distance matrix is calculated with the distance between the new group. The species separating by the smallest distance in the new matrix are then clustered to make another new composite species. This iterative procedure is carried out until all species have been grouped. If scaled branch lengths are to be used on the tree to represent the evolutionary distance between species, branch points are located at a distance halfway between each of the species being grouped. Another popular distance-based approach is the neighbor-joining method developed by Saitou and Nei [24]. To start the reconstruction, this method assumes that there is just one internal node from which branches leading to all species coming off in a star-like tree. A pair of sequences (species) is chosen at random, removed from the star, and attached to a second internal node connected by a branch to the center of the star. The distance matrix is then used to calculate the total branch length in the new tree. The sequences are then returned to their original positions and another pair attached to the second internal node, and the total branch length is again calculated. The process is repeated until all possible pairs have been examined, and the pair that makes a tree with the overall smallest branch lengths are grouped as neighbors, so that a new star with one branch fewer than the original can be created, and a new distance matrix generated. The whole process

of pair selection and branch-length calculation is iterated so that all subsequent neighbors are found that minimize the total length of branches on the tree. The result is a complete reconstructed phylogenetic tree.

The maximum likelihood methods [12, 13, 14] represent an alternative statistical approach of phylogenetic tree reconstruction. This approach computes the probabilities for every individual nucleotide substitution in a set of sequence alignments. The tree with the highest aggregate probability is the most likely true phylogenetic tree.

The advantage of the distance-based approach is that the handling of data is relatively easy because the information of multiple alignment has been reduced to its simplest form. However, the distance-based approach attends to a few or just one of many possible phylogenetic trees by considering the overall similarities between sequences and progressively grouping those that are most alike [18]. For the maximum likelihood methods, the calculation of the probabilities is complicated by the fact that the sequence of the common ancestor to the sequences being considered is generally not known, and the number of possible trees for even a modest number of sequences makes the computation intensive. The character-based or maximum parimony method [16] assumes that evolution follows the shortest possible path and that the correct phylogenetic tree is the one that requires the minimum number of nucleotide changes to produce the observed differences between the sequences. Therefore the trees are randomly constructed and the number of nucleotide changes calculated until all possible topologies have been considered, with the one requiring the smallest number of steps the most likely inferred phylogenetic tree. The maximum parsimony method is more rigorous in comparison with the neighbor-joining method. However, the parsimony analysis becomes computationally intensive with multiple alignments involving 20 or more sequences with just five sequences there are only 15 possible unrooted trees, for 10 sequences there are 2,027,025 unrooted trees, and for 50 sequences the number of trees exceeds the number of atoms in the universe [11].

One of the most popular software packages for phylogenetic analysis is Clustal, which was orginally developed in 1988 and subsequently upgraded [17]. Clustal primarily carries out multiple alignments of protein or DNA sequences, and works effectively provided that the sequences do not contain extensive internal repeat motifs. Clustal is usually used in conjunction with a program for tree reconstruction such as the neighbor-joining method. More comprehensive software packages that utilize a variety of different methods for tree reconstruction include PAUP [27], and PHYLIP [15].

Due to the limitations of phylogenetic tree reconstruction using mutiple sequence alignments, particularly with whole genome phylogeny, some computational methods have been developed to compute the distance matrices without the use of a multiple sequence alignment. Li *et al.* [19] applied the notion of Kolmogorov complexity to introduce a distance measure between two unaligned sequences and evaluated the method by comparing a set of whole mitochondrial genomes. Almeida *et al.* [1] introduced the chaos game representation for the analysis of genomic sequences. Vinga *et al.* reviewed several alignment-free methods [28], and evaluated word composition distance methods for the recognition of SCOP relationships [29]. Otu and Sayood [22] recently proposed a similar approach to the distance measure by Li *et al.*. This method calculates a distance between two sequences based on the Lempel-Ziv complexity. However, the similarity or dissimilarity measures obtained from these alignment-free methods are only able to be interpreted in conjuction with other program that analyze distance-matrix data for phylogenetic tree reconstruction such as the neighbor-joining [24], UPGMA [26], and hypercleaning [5]. To date, however no single ideal method for phylogenetic tree reconstruction has been developed.

In this paper, a new method for DNA-based phylogenetic tree reconstruction is presented using the concepts of Markov models, and the fuzzy c-means algorithm. Unlike clustering data sets with features, this proposed method transforms each nucleotide sequence into a Markov model, and then iteratively applies the fuzzy c-means to cluster the sequences represented by their corresponding Markov models into exhaustive sub-groups. The proposed method can reconstruct a phylogenetic tree of complete mitochondrial genomes without the requirements of sequence alignment and programs for distance-matrix data.

2. Markov Models for DNA Sequences

Let $\{t_k\}$ for $k = 0, 1, 2, \ldots$, be the discrete points in time, and ξ_{t_k} be the random variable that characterizes the state of the system at t_k. Let $s_1, s_2, \ldots, s_N$ represent the finite states of a system at any time. The system may be in any of these states at time t_0. Let π_i, $1 \leq i \leq N$, be the initial probability that the system is in state s_i at t_0.

We now define $a_{ij} = P\{\xi_{t_n} = j | \xi_{t_{n-1}} = i\}$, as the first-order transition probability of going from state i at t_{n-1} to state j at t_n and assume that these probabilities are stationary over time. Thus these transition probabilities going from s_i to s_j can be expressed as

$$\mathbf{A} = [a_{ij}],\ 1 \le i, j \le N \tag{1}$$

The matrix $\mathbf{A}$ is called a stochastic matrix because all the transition probabilities a_{ij} are fixed and independent of time, and must satisfy the following conditions:

$$a_{ij} \ge 0\ \forall i, j \tag{2}$$

$$\sum_{j=1}^{N} a_{ij} = 1\ \forall i \tag{3}$$

The transition matrix $\mathbf{A}$ and the initial probability vector $\pi = \{\pi_i, 1 \le i \le N\}$ associated with states $\{s_i, 1 \le i \le N\}$ completely define a Markov chain that can be denoted in a compact form as

$$\lambda = (\mathbf{A}, \pi) \tag{4}$$

With the concept of a Markov chain defined, we now wish to study the behavior of unaligned nucleotide sequences using Markov chain analysis. Let $\mathcal{G}$ be a sequence of nucleotides, and $\{\text{a,c,g,t}\} \in \mathcal{G}$ be the set of four different bases used in DNA molecules: adenine (a), cytosine (c), guanine (g), and thymine (t). To model the information contained in $\mathcal{G}$ in the context of a Markov chain, we define these four bases as the four Markov states, thus giving the number of states $N = 4$. The initial probability π_i is computed as the frequency (number of times) the system is in state i at time t_0. There is one sequence for each Markov model, thus the initial probabilities can be assumed to be equiprobable, that is

$$\pi_i = \frac{1}{N}\ \forall i \tag{5}$$

The state transition probabilities can be estimated as

$$a_{ij} = \frac{N_{ij}}{N_i},\ 1 \le i, j \le N \tag{6}$$

where N_{ij} is the number of transitions from state (base) i to state j, and N_i is the number of transitions from state i.

Since the initial probabilities $\{\pi_i\}$ for an unaligned sequence of nucleotides are assumed to be equiprobable, the distance (dissimilarity) measure between two Markov models λ_1 and λ_2 can be defined in the context of the state transition probability matrices using a Euclidean norm:

$$\begin{aligned} d_E(\lambda_1, \lambda_2) &\approx d_E(\mathbf{A}_1, \mathbf{A}_2) \\ &= ||\mathbf{A}_1 - \mathbf{A}_2||_2 \\ &= \left[\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(a_{ij}^{(1)} - a_{ij}^{(2)})^2\right]^{\frac{1}{2}} \end{aligned} \tag{7}$$

where $a_{ij}^{(1)} \in \mathbf{A}_1$, and $a_{ij}^{(2)} \in \mathbf{A}_2$.

3. Fuzzy Clustering of Markov Models

Let $J : M \times \Re^{cp} \rightarrow \Re^{+}$, and $\mathbf{U} \in M$ be a fuzzy c-partition of a collection of Markov models having been defined in (4). That is $\mathbf{X} = (\lambda_1, \lambda_2, \ldots, \lambda_n) \approx (\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_n)$. To find the fuzzy prototypes or cluster centers of $\{\mathbf{A}_k, k = 1, \ldots, n\}$, the fuzzy c-means clustering algorithm aims to minimize the following objective function [2]:

$$J(\mathbf{U}, \mathbf{V}) = \sum_{k=1}^{n}\sum_{i=1}^{c}(u_{ik})^m (d_{ik})^2 \tag{8}$$

where $m \in [1, \infty)$ is the weighting exponent; and particularly when $m = 1$, the FCM becomes identical to the hard c-means algorithm.

$$\mathbf{V} = (\mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_c) \in \Re^{cp} \tag{9}$$

in which $\mathbf{V}_i \in \Re^p$ is the cluster center or prototype of u_i, $1 \le i \le c$;

$$(d_{ik})^2 = (d(\mathbf{A}_k, \mathbf{V}_i))^2 \tag{10}$$

in which $d(\cdot)$ can be d_E being defined in (7).

The summation of the fuzzy membership grades is subject to the unity constraint:

$$\sum_{i=1}^{c} u_{ik} = 1, \forall k \tag{11}$$

Conditions for the objective function $J(\mathbf{U}, \mathbf{V})$ being defined in (8) to reach a minimum can be found by forming a new function J^* as follows:

$$J^*(\mathbf{U}, \mathbf{V}, \alpha) = J(\mathbf{U}, \mathbf{V}) + \sum_{k=1}^{n}\alpha_k(\sum_{i=1}^{c} u_{ik} - 1) \tag{12}$$

where α_k, $k = 1, \ldots, n$ are the Lagrange multipliers for the n constraints expressed in (11).

After some mathematical rearrangements and differentiating J^* with respect to all input arguments giving

$$\mathbf{V}_i = \frac{\sum_{k=1}^{n}(u_{ik})^m \mathbf{A}_k}{\sum_{k=1}^{n}(u_{ik})^m} \tag{13}$$

and

$$u_{ik} = \frac{1}{\sum_{j=1}^{c}\left(\frac{d_{ik}}{d_{jk}}\right)^{2/(m-1)}} \tag{14}$$

Thus, the FCM for clustering Markov-model based sequences has now been described.

It has been mentioned in the description of FCM that c, the number of clusters, needs to be given. In many practical cases, c is unknown. It is reasonable to expect cluster substructure at more than one value of c, and therefore necessary to estimate the most plausible value of c for the cluster analysis. This problem is known as cluster validity [10]. It is very difficult to formulate the cluster validity problem in a mathematically tractable manner, because the basic question is imposed on the definition of a cluster. For fuzzy clustering, one should examine which pairs of fuzzy groups/classes overlap, and this leads to the question of how fuzzy a fuzzy c-partition is. A heuristic solution to this problem is to calculate the measure of fuzziness in $\mathbf{U}$, and then assign c as the most valid value that has the least fuzzy partitions.

The first functional designed for cluster validity measure is the partition coefficient [3]. This partition coefficient of a fuzzy c-partition of $\mathbf{U} \in M$ of $\mathbf{X}$ is expressed as

$$F(\mathbf{U}; c) = \frac{1}{n} \sum_{k=1}^{n} \sum_{i=1}^{c} (u_{ik})^2 \tag{15}$$

Another equivalent expression for (15) that emphasizes various properties of F is the Euclidean inner product for two matrices $\mathbf{I}$, $\mathbf{J} \in \mathbf{V}_{cn}$ is $< \mathbf{I}, \mathbf{J} >= \mathrm{Tr}(\mathbf{I}\mathbf{J}^T)$, where Tr is the trace of a matrix, and $\mathbf{J}^T$ is the transpose of $\mathbf{J}$. And (15) has alternative forms

$$F(\mathbf{U}; c) = \frac{\mathrm{Tr}(\mathbf{U}\mathbf{U}^T)}{n} = \frac{< \mathbf{U}, \mathbf{U} >}{n} = \frac{||\mathbf{U}||^2}{n} \tag{16}$$

Now it can be analyzed that: if $F(\mathbf{U}; c) = 1$ then $\mathbf{U}$ contains no fuzzy clusters ($\mathbf{U}$ consists of only zeros and ones); if $F(\mathbf{U}; c) = 1/c$ (all elements in $\mathbf{U}$ is equal to $1/c$) then $\mathbf{U}$ is completely fuzzy; and in general $1/c \leq F(\mathbf{U}; c) \leq 1$. As $F(\mathbf{U}; c)$ increases, the partition of the data sets is more effective. Thus the formal strategy for selecting the most valid c^* is as follows. Let Ω_c represents any finite set of optimal $\mathbf{U}$'s $\in M$, and $c = 2, 3, \ldots, n-1$. The optimal c^* is determined by direct search

$$c^* = \arg\max_{c} [\max_{\Omega_c} F(\mathbf{U}; c)] \tag{17}$$

Fuzzy c-partitions can also be formulated using the concept of Shannon's entropy [25, 9], because the fuzzy membership grades in $\mathbf{U}$ subjected to the constraint (11) are identical to the probabilities of the Shannon's entropy S:

$$S = -\sum_{i} p_i \log_y(p_i) \tag{18}$$

where logarithmic base $y \in (1, \infty)$ and $p_i \log_y(p_i) = 0$ whenever $p_i = 0$.

As an alternative, the entropy-based partition coefficient of any fuzzy c-partition $\mathbf{U} \in M$ of $\mathbf{X}$ is given by [2]

$$H(\mathbf{U}; c) = -\frac{1}{n} \sum_{k=1}^{n} \sum_{i=1}^{c} u_{ik} \log_y(u_{ik}) \tag{19}$$

where logarithmic base $y \in (1, \infty)$ and $u_{ik} \log_y(u_{ik}) = 0$ whenever $u_{ik} = 0$.
Using the partition entropy criteria, the optimal c^* is given by

$$c^* = \arg\min_{c} [\min_{\Omega_c} H(\mathbf{U}; c)] \tag{20}$$

4. Fuzzy Phylogenetic Tree

After carrying out the fuzzy clustering process, the sequences under study can be grouped into fuzzy clusters. However, we wish to further cluster these subgroups according to the concept of cluster validity, until all sequences can be split into a cluster of at most two sequences after a fuzzy-hardening process (the term "fuzzy-hardening" is equivalent to "defuzzification", which means the membership value of a sequence belonging to a fuzzy cluster is assigned to unity if it is maximum with respect to other fuzzy clusters; otherwise it is zero). In other words, we perform a top-down fuzzy clustering approach to exhaustively identify all fuzzy subgroups at the level when no further clustering is allowed, that is when the number of "data points" is at most three. The information obtained at this level will be utilized for the tree reconstruction using a bottom-up approach.

Let $\Omega_l = \{\mathbf{V}_i^l, i = 1, 2 \ldots, c^l\}$ be the space of fuzzy clusters obtained at nested or hierarchical level l, $l = 1, 2, \ldots, L$. That is $\Omega_L \subset \Omega_{L-1} \subset \Omega_{L-2} \subset \ldots \subset \Omega_1$. The distance between two groups of fuzzy clusters in Ω_l, denote as $\mathbf{G}_1^l$ and $\mathbf{G}_2^l$, can be determined as

$$d_E(\mathbf{G}_1^l, \mathbf{G}_2^l) = d_E(\bar{\mathbf{G}}_1^l, \bar{\mathbf{G}}_2^l) \tag{21}$$

where d_E has been defined in (7), and

$$\bar{\mathbf{G}}_1^l = \frac{1}{|\mathbf{G}_1^l|} \sum_{i}^{c^l} \mathbf{V}_i^l \tag{22}$$

in which $|\mathbf{G}_1^l|$ is the cardinality of $\mathbf{G}_1^l$, $\mathbf{V}_i^l$ is a fuzzy-hardening cluster center of $\mathbf{G}_1^l$, $\bar{\mathbf{G}}_1^l = \mathbf{V}_1^l$ if $|\mathbf{G}_1^l| = 1$.

Using (21), our method joins the fuzzy clusters based on the minimum distance in the same way as the UPGMA approach does. This cluster joining is carried out in a hierachical, bottom-up manner from $l = 1$ to $l = L$; and therefore unlike the UPGMA, it does not need to calculate the distances between the new group and all other remaining groups in all levels, but only between the new and other remaining clusters of the same level.

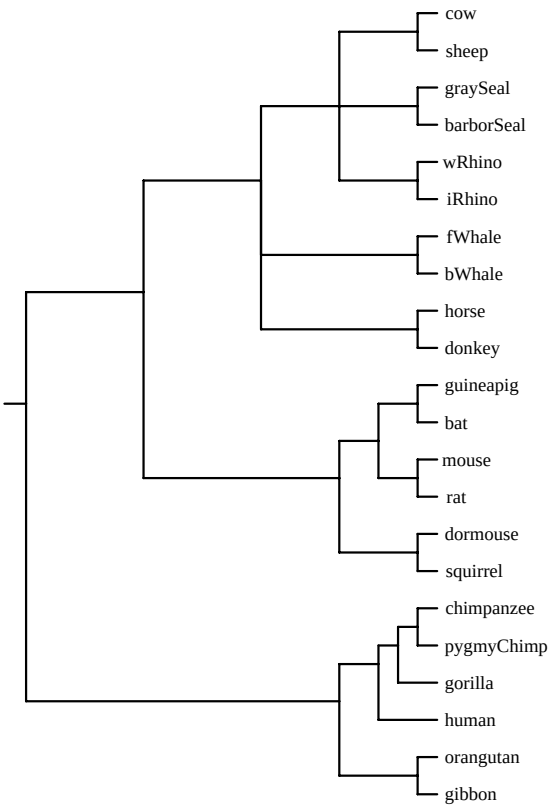

Figure 1. Phylogenetic tree reconstructed by fuzzy clustering using $F(\mathbf{U}, c)$.

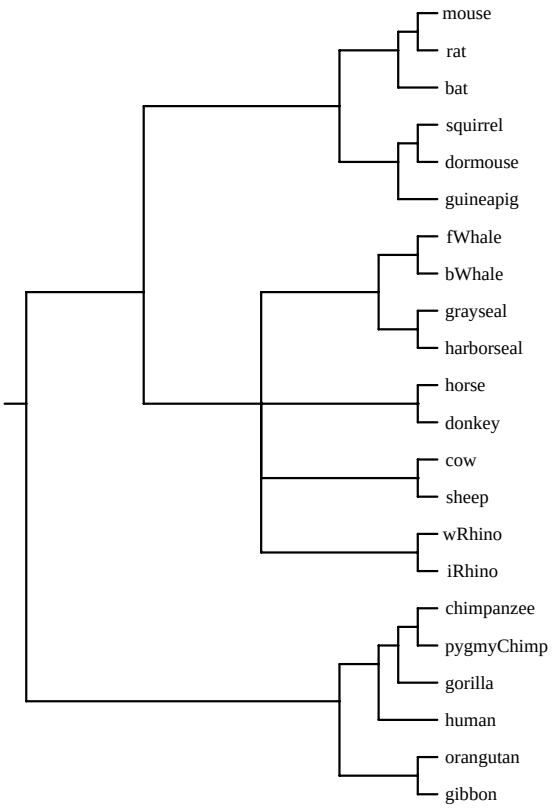

Figure 2. Phylogenetic tree reconstructed by fuzzy clustering using $H(\mathbf{U}, c)$.

5. Results

We have selected a data set of twenty-two complete mitochondrial genomes to test the proposed method. This set is a subset of the larger data set that has been the subject of many independent phylogenetic studies [6, 7, 23, 20, 21, 19, 22], and led to conflicting findings regarding the phylogeny of eutherian orders. We therefore selected more distinct species to lessen this controversial issue. The mtDNA sequences were obtained from the public-domain database of the National Center for Biotechnology Information (NCBI) (www.ncbi.nlm.nih.gov/Entrez/). The genomes consist primates, rodents, and ferungulates. The primates consist of human, chimpanzee, pygmy chimpanzee, gorilla, orangutan, and gibbon. The rodent group includes rat, mouse, dormouse, squirrel, guinea pig, and fruitbat. The ferungulates include harbor seal, gray seal, white rhino, indian rhino, blue whale, finback whale, cow, sheep, donkey, and horse.

The fuzzy c-means clustering of the Markov models of the above 22 complete mtDNA genomes was carried out in a fully automatic procedure using $F(\mathbf{U}, c)$ as the partition criterion. Figure 1 shows the phylogenetic tree obtained from the proposed method using $F(\mathbf{U}, c)$.

The fuzzy c-means clustering of the Markov models of the above 22 complete mtDNA genomes was also carried out in a fully automatic procedure using $H(\mathbf{U}, c)$ as the partition criteria. Figure 2 shows the phylogenetic tree obtained from the proposed method using $H(\mathbf{U}, c)$.

From the topologies of the two fuzzy phylogenetic trees as shown in Figures 1 and 2, we can see that both fuzzy clusterings using $F(\mathbf{U}, c)$ and $H(\mathbf{U}, c)$ can cluster the primates, ferungulates, and rodents into three groups. Both cluster-validity criteria give the same grouping of the primates: (((chimpanzee, pygmy chimpanzee), gorilla, human), (orangutan,

gibbon)). However, there are some differences in the sub-groupings of the ferungulates and the rodents. Whales and rhinos are closest to each other for the clustering using $F(\mathbf{U}, c)$; whereas whales and seals are for $H(\mathbf{U}, c)$. From Figure 2, we found that the fuzzy clustering using the validity criterion $H(\mathbf{U}, c)$ groups the squirrel with the non-murid rodents: (squirrel, dormouse); whereas the squirrel is not direclty located with the non-murid rodents, although it is clustered in the rodent group. Traditional molecular phylogenetic studies hypothesized the monophyly of Rodentia. However, this view has been challenged by several phylogenetic analyses, and a recent study of complete mtDNA genomes of 16 mammalian species has established that the guinea pig is not a roden [8]. As a result, the guinea pigs is still very controversial [23], and other alignment-free methods [19, 22] found that the guinea pig groups with neither the murid (mouse, rat) nor the non-murid rodents.

6. Conclusions

We have presented a new method for clustering biological sequences with application to the phylogenetic study of complete mammalian mtDNA genomes. The proposed method can classify the relationship among primates, ferungulates, and rodents and reconstruct the phylogenies using their complete mtDNA genomes in a fully automatic procedure without relying on any evolutionary model. The results obtained from experiments carried out without prior knowledge of the numbers of clusters have shown the consistency of the proposed computational model for molecular phylogenetics.

References

[1] J.S. Almeida, J. A. Carrico, A. Maretzek, P. A. Noble, and M. Fletcher, Analysis of genomic sequences by chaos game representation, *Bioinformatics*, 17 (2001) 429-437.

[2] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, New York, 1981.

[3] J.C. Bezdek, Numerical taxonomy with fuzzy sets, *J. Mathematical Biology*, 1 (1974) 57-71.

[4] T.A. Brown, *Genomes*, second edition, John Wiley & Sons, New York, 2002.

[5] V. Berry, D. Bryant, T. Jiang, P. Kearny, M. Li, T. Wareham, and H. Zhang, A practical algorithm for recovering the best supported edges of an evolutionary tree, *Proc. 11th Ann. ACM-SIAM Symp. Discrete Algorithms*, San Francisco, CA, pp. 287-296, 2000.

[6] Y. Cao, N. Okada, and M. Hasegawa, Phylogenetic position of guinea pigs revisited, *Mol. Biol. Evol.*, 14 (1997) 461-464.

[7] Y. Cao, A. Janke, P.J. Waddell, M. Westerman, O. Takenaka, S. Murata, N. Okada, S. Paabo, and M. Hasegawa, Conflict among individual mitochondrial proteins in resolving the phylogeny of eutherian orders, *J Mol. Evol.*, 47 (1998) 307-322.

[8] A.M. D'Erchia, C. Gissi, G. Pesole, C. Saccone, and U. Arnason, The guinea-pig is not a rodent, *Nature*, 381 (1996) 597-600.

[9] A. DeLuca, and S. Termini, A definition of a nonprobabilistic entropy in the setting of fuzzy sets theory, *Information and Control*, 20 (1972) 301-312.

[10] R. Duda, and P. Hart, *Pattern Classification and Scene Analysis*. John Wiley & Sons, New York, 1973.

[11] D.J. Eernisse, A brief guide to phylogenetic software, *Trends Genet.*, 14 (1998) 473-475.

[12] J. Felsenstein, Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters, *Syst. Zool.*, 22 (1973) 240-149.

[13] J. Felsenstein, Evolutionary trees from DNA sequences: A maximum likelihood approach, *J. Mol. Evol.*, 17 (1981) 368-376.

[14] J. Felsenstein, and G.A. Churchill, A hidden Markov model approach to variation among sites in rate of evolution, *Mol. Bio. Evol.*, 13 (1996) 93-104.

[15] J. Felsenstein, PHYLIP (Phylogeny Inference Package) version 3.5c. Distributed by the author. Department of Genetics, University of Washington, Seattle, 1993.

[16] W.M Fitch, On the problem of discovering the most parsimonious tree, *Am. Nat.*, 111 (1977) 223-257.

[17] F. Jeanmougin, J.D. Thompson, M. Gouy, D.G. Higgins, and T.J. Gibson, Multiple sequence alignment with Clustal X, *Trends Biochem. Sci.*, 23 (1998) 403-405.

[18] D.E. Krane, and M. L. Raymer, *Fundamental Concepts of Bioinformatics*, Benjamin Cummings, San Francisco, CA, 2003.

[19] M. Li, J. H. Badger, X. Chen, S. Kwong, P. Kearney, and H. Zhang, An information-based sequence distance and its application to whole mitochondrial genome phylogeny, *Bioinformatics*, 17 (2001) 149-154.

[20] O. Madsen, M. Scally, C.J. Douady, D.J. Cao, W.R. DeBry, R. Adkins, H.M. Amrine, M.J. Stanhope, W.W. de Jong, and M.S. Springer, Parallel adaptive radiationsin two major clades of placental mammals, *Nature*, 409 (2001) 610-618.

[21] W.J. Murphy, E. Eizirik, S.J. O'Brien, O. Madsen, M. Scally, C.J. Douady, E. Teeling, O.A. Ryder, M.J. Stanhope, W.W. de Jong, and M.S. Springer, Resolution of the early placental mammal radiation using Bayesian phylogenetics, *Science*, 294 (2001) 2348-2351.

[22] H.H. Otu, and K. Sayood, A new sequence distance measure for phylogenetic tree construction, *Bioinformatics*, 19 (2003), 2122-2130.

[23] A. Reyes, C. Gissi, G. Pesole, F.M. Catzeflis, and C. Saccone, Where do rodents fit? Evidence from complete mitochondrial genome of Sciurus vulgaris, *Mol. Biol. Evol.*, 17 (2000) 979-983.

[24] N. Saitou, and M. Nei, The neighbor-joining method: A new method for reconstructing phylogenetic trees, *Mol. Bio. Evol.*, 4 (1987) 406-425.

[25] C.E. Shannon, Mathematical theory of communication, *Bell Syst. Tech. Journal*, 3 (1948) 379-423.

[26] R. Sokal, and C.D. Michener, Statistical method for evaluating systematic relationships, *Univ. Kansas Sci. Bull.*, 38 (1958) 1409-1438.

[27] D.L. Swofford, *PAUP: Phylogenetic Analysis using Parsimony*, Illinois Natural History Survey, Champaign, IL, 1993.

[28] S. Vinga, and J. Almeida, Alignment-free sequence comparison - a review, *Bioinformatics*, 19 (2003) 513-523.

[29] S. Vinga, R. Gouveia-Oliveira, and J.S. Almeida, Comparative evaluation of word composition distances for the recognition of SCOP relationships, *Bioinformatics*, 20 (2004) 206-125.