Using Hierarchical Cluster and Factor Analysis to Classify and Built a phylogenetic Tree Species of ND1 Mitochondria

Ali A Ibrahim[*1], Zainib H. A. Al Rikabi[**] and Najwa Sh. Ahmed[***]
[*]College of Science, Biotechnology, Al-Nahrain University, Baghdad-Iraq.
[**] Genetic Engineering and Biotechnology Institute, Baghdad University, Baghdad-Iraq.
[***] Biotechnology Research Center, Al-Nahrain University, Baghdad-Iraq.
[1]E.mail: dr_ali9@yahoo.com.

Abstract

The research amied to classify and characterize 11 different organisms according to amino acid sequencing of ND1 mitochondria gene by using the two analysis methods of factor analysis and cluster analysis, in order to achieve this goal, we collected amino acid sequencing from Database of national center biotechnology information (NCBI) for eleven organisms (human, chimpanzee, monkeys, cow, horse, mouse, rat, fish, frog, chicken and rabbit) and this was one of the most important of the results which was as follows: A: By using factor analysis the organisms was classified into six groups. B: By using cluster analysis the organisms was classified into four groups. Each group did included similar number of organism, they were identical with each other. This was a result of a similar chain peptide multiple for ND1. Results of the two methods (cluster and factor analysis) were cluster organisms (human and chimpanzee and monkey) in one group and was the lowest value between humans and chimpanzees, with distance equal to 0.09, the great affinity between the human and chimpanzee and monkey and clustering of chicken as a single group and reached the highest value between human and chicken with distance equal to 0.642012 and to move away genetic traits to humans from chicken. The conclusion factor analysis method was able to give more precise details of the method from cluster analysis through the. identification and clustering of new groups with genetic dimension for the other group.

Introduction

Mitochondria DNA as a genetic material has its own replication system separate from that of the nuclear genome, it is also only inherited from the mother [1]. Mitochondria are about 0.5–1 mm in diameter and up to 7 mm long [2]. The genes found within the mitochondria contain the information that codes for the production of many of the important enzymes that drive the biochemical reactions to produce the body's source of energy: a chemical called ATP (*Adenosine Triphosphate*) that is used by the body to drive the various reactions essential for the body to function, grow and develop [3]. The cells in the body, especially in organs such as the brain, heart, muscle, kidneys and liver, cannot function normally unless they are receiving a constant supply of energy [4]. Four of these five complexes contain mtDNA-encoded subunits; complex II, the succinate-ubiquinone oxidoreductase complex, is the exception. The ND1, ND2, ND3, ND4, ND4L, ND5, and ND6 subunits of complex I (NADH-ubiquinone oxidoreductase) are all core subunits localized in the membrane arm of the enzyme. The ND1 subunit is implicated in ubiquinone binding [5]. Changes in any of these mitochondrial genes that make them faulty can result in biochemical problems due to absence of enzymes involved in the respiratory chain, or enzymes that are impaired and do not work properly. This leads to a reduction in the supply of ATP, and may result in problems with the body's functions. Mitochondria play a pivotal role in cellular metabolism. About 1% of patients with diabetes mellitus had the 3243 mutation, a typical mutation of MELAS, facilitated a search for patients with common diseases caused by mtDNA mutations [6]. MtDNA mutations have also been found in patients with other common diseases such as cardiomyopathy, migraine, cluster headache, and deafness [7].

Similarities and divergence among related biological sequences revealed by sequence alignment often have to be rationalized and visualized in the context of phylogenetic tree. Thus, molecular phylogenetics is a fundamental aspect of bioinformatics .

Cluster analysis is an approach that finds structure in data by identifying natural groupings (clusters) in the data. A cluster is simply a collection of cases that are more 'similar' to each other than they are to cases in other clusters [8]. In engineering and computer science, classification is usually called pattern recognition. Some writers use the term "Classification Analysis" to describe cluster analysis, in which the observations are clustered according to variable values rather than into predefined groups, which is also called data segmentation, is unsupervised learning approach [9]. The problem that classification analysis is designed is to solve the following one: given nimages (Objects), each of which has score on variables, devise a scheme for grouping the objects into classes so that 'similar' ones are in the same class. The method must be completely numerical and number of class is unknown [10 & 11]. Cluster analysis is the name of multivariate techniques, whose primary purpose is to identify similar entities from characteristics they posses. It identifies and classifies objects or variables, so that each object is very similar to others in its cluster, with respect to some predetermined selection criteria. The resulting object clusters should then exhibit high internal (within-cluster) homologyeneity and high external (between-clusters) heterogeneity. Thus, if the classification is successful, Then objects within clusters will be close together when plotted geometrically, and the objects in different clusters will before apart [12].

The goals of this research paper are:

- To built phylogenetic tree of mitochondria gene for the different species..
- To find an optimal grouping for which the observations or objects within each cluster are similar, but the clusters are dissimilar to each other.
- To compare between the results of cluster analysis and factor analysis.

Materials and Methods

Sample collection

Collection of sequenced amino acids was made from Database of the National Center Biotechnology Information (NCBI) online at (http:// www.ncbi.nlm.nih.gov) for eleven organisms (human, chimpanzee, monkeys, cow, horse, mouse, rat , fish, frog, chicken and rabbit) [13].

Bioinformatics Programs

To achieve the research objectives, it was used the following three software's and package:

-*ClustalX*: the raw data (mitochondrial gene) entered in *ClustalX* package to align the mitochondria sequences of eleven species,

-*dnadist_PHYLIP:*, the alignment sequences entered in *dnadist_PHYLIP* software to convert raw data (which is DNA sequences) into a distance matrix,

- *SPSS*, the distance matrix entered in *SPSS* software to implement the cluster analysis and factor analysis.

Hierarchical Cluster Analysis

Cluster analysis technique used to search for patterns in a data set by grouping the (multivariate) observations into clusters. Hierarchical cluster is one of cluster analysis techniques, since it takes a part of grouping and classifying data into clusters and these clusters are very similar to each other considering similarity within cluster, but it is different or dissimilar between clusters. Hierarchical cluster analysis is considered as a good, active way and a tool to analysis data in different methods. This method (Hierarchical cluster) used to analysis matrix of data consists of (n) of elements, and every on of it have number of variables (p).

There are two basic approaches for generating a hierarchical clustering:

Agglomerative:

Start with the points as individual clusters and, at each step, merge the closest pair of clusters.

Divisive:

Start with one, all-inclusive cluster and, at each step, split a cluster until only singleton clusters of individual points remain. In this case, the need is to decide which cluster to split at each step and how to do the splitting. A hierarchical clustering is often displayed graphically using a tree-like diagram called a dendrogram, which displays both cluster-subcluster relationships and the order in which the clusters were merged or split. For sets of two dimensional points. These points were clustered using the single-link techniques are

variations on a single approach: starting with individual points as clusters, successively merge the two closest clusters until only on cluster remains. This approach is expressed more formally in Fig.(1).

Step 1. compute the distance matrix.
Step 2. repeat
Step 3. Merge the closest two clusters.
Step 4. Update the distance matrix to reflect the proximity between the new cluster and the original clusters.
Step 5. Until Only one cluster remains.

Fig.(1) Agglomerative hierarchical clustering algorithm.

From Fig.(1), can be explained as follows:

1. Calculating descent distance that decides and specifies closing degree between every two types of different elements according to the following equation :

$$d_{ij} = \sqrt{\{\sum_{k=1}^{p}(x_{ik} - x_{jk})^2\}}$$

2. Then searching for the shortest descent distance in matrix (d_{ij}) to make relationship between the element (i) and the element (j) , x_{ik} is the value of variable X_k for individual I and x_{jk} is the value of the same variable for individual j.
3. Continuing of Agglomerate process or grouping depending on the possible shortest descent distance, until making connection of the last two groups (or cluster) to form the final groups.

Factor Analysis

Factor analysis is a statistical method used to describe variability among observed, correlated variables in terms of a potentially lower number of unobserved variables called factors. In other words, it is possible, for example, that variations in three or four observed variables mainly reflect the variations in fewer unobserved variables. Factor analysis searches for such joint variations in response to unobserved latent variables. The observed variables are modeled as linear combinations of the potential factors, plus "error" terms. The information gained about the interdependencies between observed variables can be used later to reduce the set of variables in a dataset. Computationally this technique is equivalent to low rank approximation of the matrix of observed variables. Factor analysis originated in psychometrics,.The aim of factor analysis is to simplify a complex data set by representing the set of variables in terms of a smaller number of underlying (hypothetical or unobservable) variables, known as factors or latent variables. The technique is a branch of multivariate analysis and may also be described as unsupervised learning and an exercise in modeling [11]. Spearman proposed the idea that the test scores are all of the form

$$X_i = a_i F + e_i,$$

Where X_i is the ith standardized score with a mean of zero and a standard deviation of one, a_i is a constant, F is a 'factor' value, which has mean of zero and standard deviation of one for individuals as a whole, and e_i is the part of X_i that is specific to the i th test only. There are three stages to a factor analysis. To begin with, provisional factor loadings a_{ij} are determined. One way to do this is to do a principal component analysis and neglect all of the principal components after the first m, which are themselves taken to be the m factors. The factors found in this way are the uncorrelated with each other and are also uncorrelated with the specific factors. Whatever way the provisional factor loadings are determined, it is possible to show they are not unique. If $F_1, F_2, \cdots; F_m$ are the provisional factors, then linear combinations of these of the form

$$F_1' = d_{11}F_1 + d_{12}F_2 + \cdots + d_{1m}F_m$$
$$F_2' = d_{21}F_1 + d_{22}F_2 + \cdots + d_{2m}F$$
$$F_m' = d_{m1}F_1 + d_{m2}F_2 + \cdots + d_{mm}F_m$$

can be constructed that are uncorrelated and 'explain' the data just as well. There are an infinite number of alternative solutions for factor analysis model, and this leads to the second stage in the analysis, which is called factor rotation. Thus the provisional factors are transformed in order to find new factors that are easier to interpret. To 'rotate' in this context means essentially to choose the d_{ij} values in the above equations. The last stage of an

analysis involves calculating the factor scores. These are the values of the factors $F_1, F_2, \cdots; F_m$ for each of the individuals. Generally the number of factors (m) is up to the
factor analyst, although it may sometimes be suggested by the nature of the data [12]. This approach is expressed more formally in Fig.(2).

Step 2. Generate a variance covariance matrix of the observed variables.
Step 3. Select number of factors.
Step 4. Extract your initial set of factors.
Step 5. Perform factor rotation to a terminal solution.
Step 6. Interpret the factor structure.
Step 7. Construct factor scores to use in further analysis

Fig.(2) Factor algorithm.

Results and Discussion

Cluster Analysis

After entering the distance matrix (Table (1)) in *SPSS* and implement cluster analysis form Analysis menu of *SPSS* software the result was listed in Fig.(3).

Table (1)
Distance values for sequencing amino acid of ND1 mitochondrial gene of eleven organisms.

	Human	***Monkey***	***Horse***	***Frog***	***Chicken***	***Fish***	***Rat***	***Cow***	***Chimpanzee***	***Rabbit***	***Mouse***
Human	0.0	.258	.395	0.602	.642	.548	.416	.381	.091	.375	.417
Monkey	.258	0.0	.369	0.606	.655	.573	.440	.397	.245	.394	.434
Horse	.359	.369	0.0	0.525	.596	.501	.370	.258	.375	.339	.372
Frog	.602	.606	.525	0.0	.615	.439	.535	.540	.565	.520	.519
Chicken	.642	.655	.596	.615	0.0	.576	.624	.614	.616	.603	.620
Fish	.548	.572	.501	.439	.576	0.0	.526	.511	.527	.510	.515
Rat	.416	.440	.370	.535	.624	.526	0.0	.371	.418	.360	.216
Cow	.381	.397	.258	.540	.614	.511	.371	0.0	.368	.330	.360
Chimpanzee	.091	.254	.375	.565	.616	.527	.418	.368	0.0	.401	.423
Rabbit	.375	.394	.339	.520	.603	.510	.360	.330	.401	0.0	.354
Mouse	.417	.434	.372	.519	.620	.515	.216	.360	.423	.354	0.0

No. of clusters	*5*		*6*		*4*		*11*		*7*		*10*		*8*		*3*		*2*		*9*		*1*
1	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
2	X	X	X	X	X		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
3	X		X	X	X		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
4	X		X	X	X		X	X	X	X	X	X	X	X	X		X	X	X	X	X
5	X		X		X		X	X	X	X	X	X	X	X	X		X	X	X	X	X
6	X		X		X		X	X	X		X	X	X	X	X		X	X	X	X	X
7	X		X		X		X	X	X		X		X	X	X		X	X	X	X	X
8	X		X		X		X	X	X		X		X	X	X		X		X	X	X
9	X		X		X		X	X	X		X		X		X		X		X	X	X
10	X		X		X		X		X		X		X		X		X		X	X	X

Fig.(3) The steps of clustering of sequencing of amino acid of ND1 mitochondria gene of eleven organism.

The tree of cluster analysis of mitochondria gene of eleven species constructed from Fig.(3) and as show in Fig.(4) bellow:

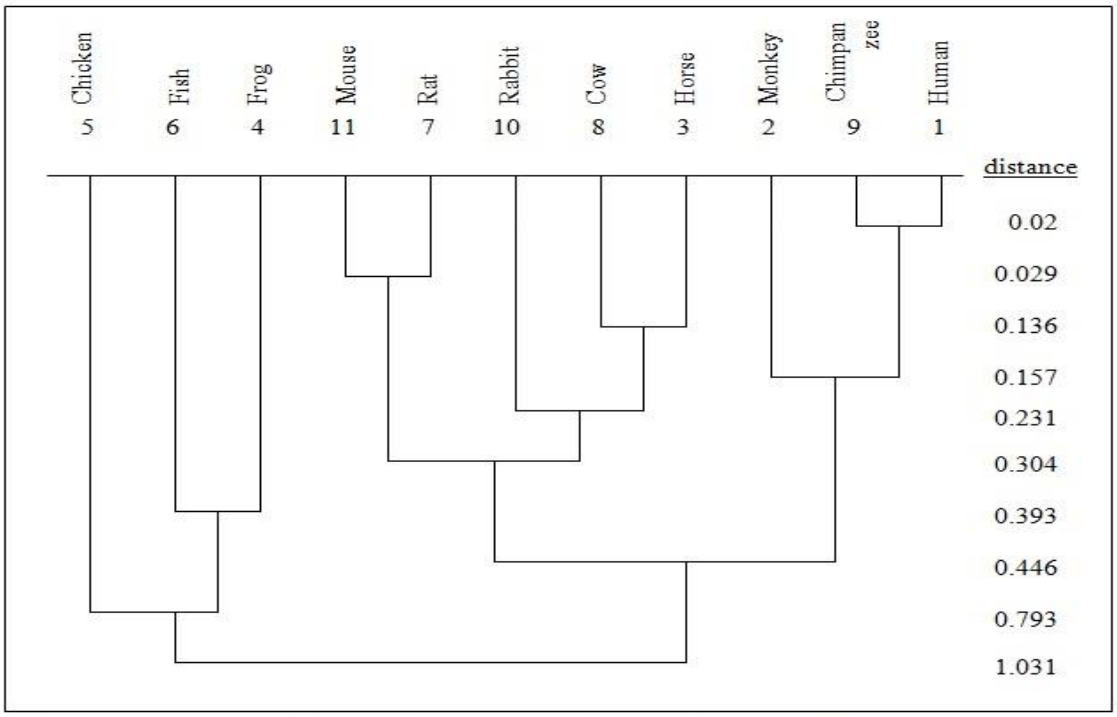

Fig. (4) The clustering tree of sequencing of amino acid of ND1 mitochondria gene of eleven organisms.

From Fig.(4), hierarchical cluster analysis determine the following clusters: Cluster one: Human, Chimpanzee, and Monkey, Cluster two: Horse, Cow, Rabbit, rat, and Mouse, Cluster three: Frog, and Fish, Cluster four: Chicken.

Factor Analysis

Fig.(5), shows the steps followed in this application of factor analysis techniques. The starting point in factor analysis, as with other statistical techniques, is the research problem.

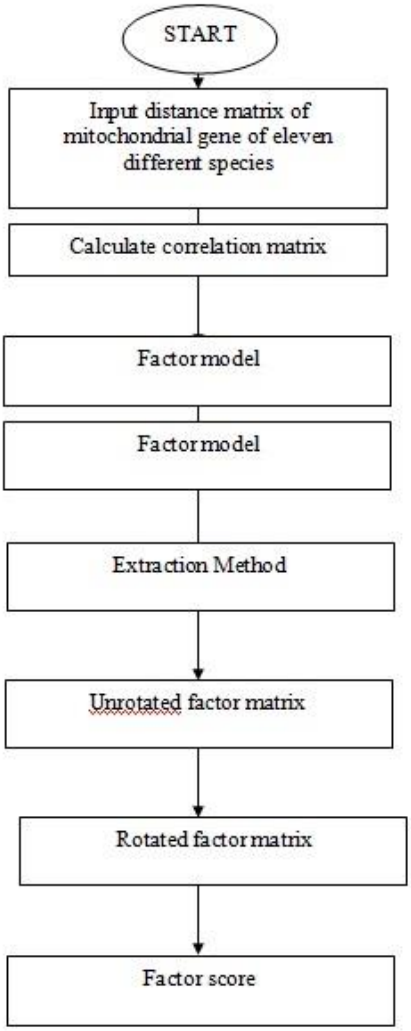

Fig. (5) Implementation of factor analysis diagram.

After entering the distance matrix (Table (1)) in *SPSS* and implement Factor analysis form Analysis menu of *SPSS* software the result was listed in Table (2), and as follows:

Table (2)
Total Variance Explained for Initial Eigenvalues and Extraction Sums of Squared Loadings.

Component	*Initial Eigen values*			*Extraction Sums of Squared Loadings*			*Rotation Sums of Squared Loading*		
	Total	*% of variance*	*Cumulative %*	*Total*	*% of Variance*	*Cumulative %*	*Total*	*% of variance*	*Cumulative %*
1	4.918	44.708	44.708	4.918	44.708	44.708	3.076	27.953	27.953
2	1.941	17.642	62.350	1.941	17.642	62.350	2.351	21.372	49.325
3	1.308	11.890	74.239	1.308	11.890	74.239	2.089	18.993	68.317
4	.897	8.156	82.396	.897	8.156	82.396	1.164	10.578	78.895
5	.693	6.301	88.696	.693	6.301	88.696	1.078	9.801	88.696
6	.516	4.689	93.386						
7	.287	2.606	95.992						
8	.249	2.267	98.259						
9	.168	1.523	99.782						
10	.024	.218	100.000						
11	-1.13E-17	-1.03E-16	100.000						

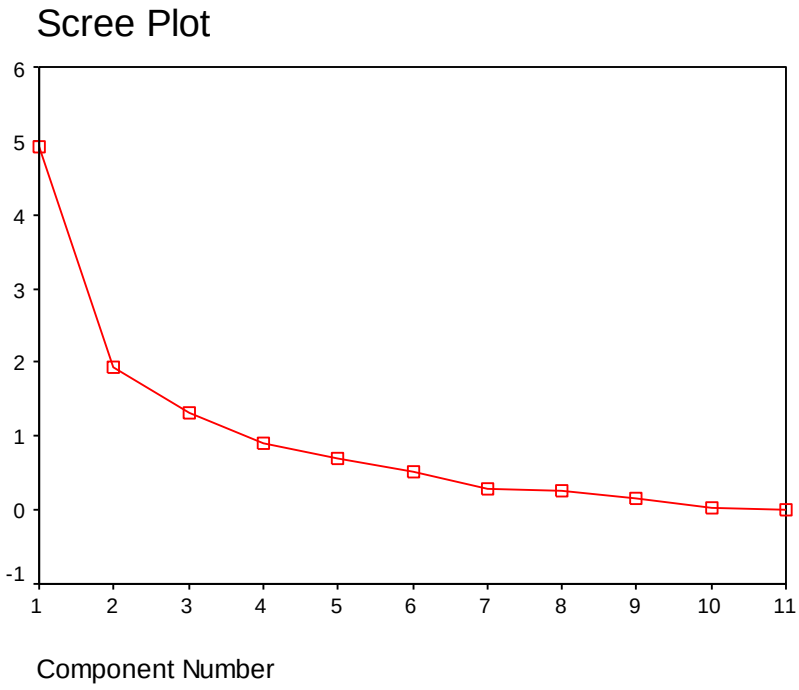

Fig. (6) Number of eigenvalue against component number which constructed from Table (2), and as follows:

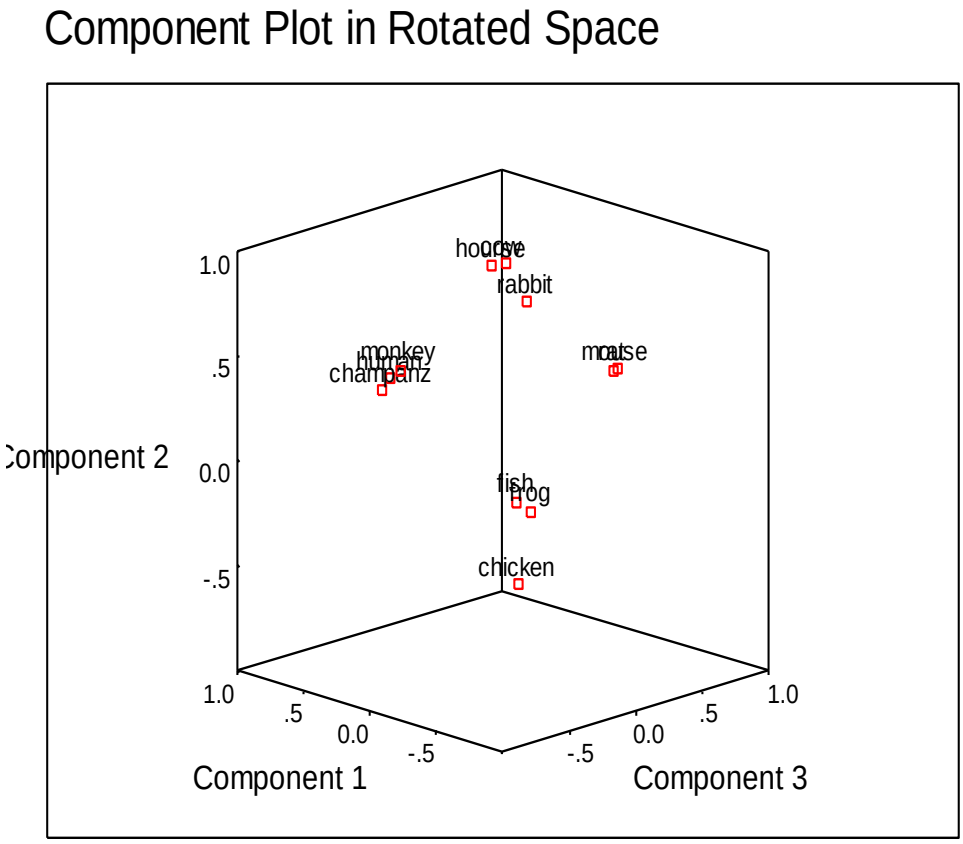

Fig.(7) The location of mitochondria gene of eleven species determine by three components of factor analysis.

From Fig.(7), Factor analysis determine the following Clusters: Cluster one: Horse, and Cow, Cluster two: Human, Monkey, and Chimpanzee, Cluster three: Mouse, and rat, Cluster four: Fish, and Frog, Cluster five: Rabbit, Cluster Six: Chicken.

Comparing between two analysis

Classification based on the principle of amino acids and strains variations that have occurred in sequence and led to be classified into groups according to methods used in the research. The results showed there was no congruence between the two methods (cluster and factor analysis) in the classification of various organisms, class into four groups in a way cluster analysis (first group of human and chimpanzee, monkeys, and the second group horse and cow and mouse, rat and rabbit and the third group of fish and frog and the fourth group chicken) and to six groups by factor analysis and six groups in a way factor analysis (the first group of human and chimpanzee, monkeys; the second group cow and horse; the third group mouse, rat; the fourth group of fish and frog; group Five of chickens and group six of rabbit). Factor analysis was method of giving more exact than cluster analysis way through the separation of some of the groups in the new totals closest genetic convergence and the dimensions of the totals of her genetic dimension.

Results of the two methods (cluster and factor analysis) were cluster organisms (human and chimpanzee and monkey) in one group and was the lowest value between humans and chimpanzees, amounting to 0.09 as in Table (1) and Fig. (4 & 7) of the great affinity between the human and chimpanzee and monkey and clustering of chicken as a single group and reached the highest value between human and chicken a mounting to 0.642012 and to move away genetic traits to humans from chicken. Factor analysis way of clustering with rat and mouse and clustering with horse and cow however, clustering rabbit alone clustering, results similar research as class researcher Mouchaty *et al*. when used three statistical (maximum parsimony, neighbor joining and maximum likelihood) analyses of the cytochrome b gene [13]. The clustering of fish and frog because the frog amphibian and fish aquatic organisms evidence of similar peptide chains of ND1 mitochondria and different from other groups. Comparing factor analysis with cluster analysis means to approach a data set from two complementary perspectives. The underlying logic of both procedures is classification. Classification in either approach is based on homogeneity. Factor analysis, in contrast, concentrates on the homogeneity of variables resulting from the similarity of values assigned to variables by respondents. From the perspective of the data matrix, variables are located in the columns of the matrix and are classified into factors or dimensions. In this paper we refer to cluster analysis in this specific meaning. Obviously, cluster analysis and factor analysis yield different information about the data. While factor analysis and especially structural equation modeling implies the aspiration of establishing a theoretically based causal relationship between indicators (items) and a latent variable (the factor or dimension), the goal of cluster analysis is to find an empirical classification or an a priori theoretically defined cluster structure [12]. The conclusion, factor analysis method is more exact than cluster analysis in determining of this paper refer that of genetic convergence between different organisms.

Reference

[1] Schwartz, M., and Vissing, J. “Paternal inheritance of mitochondrial DNA”. N. Engl. J. Med, 347, 576–580, 2000.

[2] Gray, M.W., Burger, G., and Lang, B.F. “The origin and early evolution of mitochondria”. Genome Biol, 2, 1011-1018, 2001.

[3] Rube, D.A., and van der Bliek, A.M. “Mitochondrial morphology is dynamic and varied”. Mol. Cell. Biochem., 256-257, 331-339, 2004.

[4] Kato, T. “The other, forgotten genome: mitochondrial DNA and mental disorders”. Molecular Psychiatry, 6, 625–633, 2001.

[5] Schultz, B.E., and Chan, S.I. “Structures and proton-pumping strategies of mitochondrial respiratory enzymes”. Annu. Rev. Biophys. Biomol. Struct, *30*, 23–65, 2001.

السمك , الضفدع , الدجاج والأرنب) وكانت من أهم نتائج هذا البحث الآتي: أ : باستخدام التحليل العاملي صنفت الكائنات الى ست مجاميع، ب : باستخدام التحليل العنقودي صنفت الكائنات الى اربعة مجاميع. تحتوي كل مجموعة منها على أعداد مختلفة من الكائنات الحية تتماثل فيما بينها , من جراء تماثل السلسة الببتيدية المتعددة ل ND1 وتختلف الكائنات في المجموعة الواحدة عن المجموعات الأخرى وكانت نتائج الطريقتين (العنقودي والعاملي) الكائنات الكتلة (الإنسان والشمبانزي وقرد) في مجموعة واحدة، وكان أقل قيمة بين البشر والشمبانزي، تبلغ المسافة إلى ٠,٠٩، وتقارب كبير بين الإنسان والشمبانزي وقرد و بلغ تجميع الدجاج كمجموعة واحدة وأعلى قيمة بين الإنسان والدجاج والبالغة ٠,٦٤٢٠١٢ المسافة لبتعاد الصفات الوراثية للإنسان من الدجاج, ونستنتج من هذا البحث ان طريقة التحليل ألعاملي من أعطاء تفاصيل أكثر دقه من طريقة التحليل العنقودي وذلك من خلال تحديد وتعنقد مجاميع جديدة ذات البعد الوراثي عن المجموعة الأخرى.

[6] Susumu Suzukia, S. and Okaa, Y. "Clinical features of diabetes mellitus with the mitochondrial DNA 3243 (A–G) mutation in Japanese". Maternal inheritance and mitochondria-related complications, 59 (3): 207–217, 2003.

[7] Kogelnik AM, Lott MT, Brown MD, Navathe SB, Wallace DC. "MITOMAP: a human mitochondrial genome database2004 update". Nucleic Acids Res, **33**, D611–D613, 2005.

[8] Zenko B. "Learning predictive clustering rules" PhD THESIS, Faculty of computer and information science, University of Ljubljana, 2007.

[9] Bryan, F.; Manly, J. "Multivariate Statistical Method APRIMER", University of Otago, Newzeland, Chapman and Hall, pp, 214, 2005.

[10] Maurice Kendall, Sc. D., F.B.A. "Multivariate Analysis" Charles Griffin and company LTD, London and High Wycobe, 1975.

[11] Joseph, F. and Hair, Jr. "Multivariate Data Analysis With Readings", Macmillan Puplishing Company, New York, pp,757, 1995.

[12] Fieller, N. "Further Multivariate Analysis : Working Notes", NRJF, University of Sheffield, 2001.

[13]http://www.ncbi.nlm.nih.gov/books/bv.fcgi?call=bv.View..ShowTOC&rid=immTOC&depth=10.

[14] Mouchaty, S., Gullberg A., Janke A., and Arnason U. "The phylogenetic position of the Talpidae within eutheria based on analysis of complete mitochondrial sequences". Mol. Biol. Evol,17, 60- 67, 2000.

الخلاصة

يهدف البحث الحالي الى تصنيف وتميز ١١ كائنا من الكائنات الحية المختلفة استنادا الى تتابع الأحماض الامينية لجين المايتوكندريا ND1 باستخدام طريقتي التحليل ألعاملي والتحليل العنقودي وللوصول للهدف تم جمع تتابع الأحماض الامينية من قاعدة البيانات للموقع الوطني للمعلومات التقنيات الاحيائية NCBI ل احدعشر كائن والمتمثل (الإنسان , الشمبانزي , القردة والبقر, الحصان ,الفار ,الجرذ,