Optimization of Morphological Data in Numerical Taxonomy Analysis Using Genetic Algorithms Feature Selection Method

Yasin Bakış
Abant İzzet Baysal University
Faculty of Science,
Department of Biology,
14280 Bolu, TURKEY
+905358578118
bakis_y@ibu.edu.tr

O. Uğur Sezerman
Sabancı University
Faculty of Engineering and Natural Sciences, Orhanli,
Tuzla 34956 Istanbul,
TURKEY
+902124839513
ugur@sabanciuniv.edu

M. Tekin Babaç
Abant İzzet Baysal University
Faculty of Science,
Department of Biology,
14280 Bolu, TURKEY
+903742541000
babac_m@ibu.edu.tr

Cem Meydan
Sabancı University
Faculty of Engineering and Natural Sciences, Orhanli,
Tuzla 34956 Istanbul,
TURKEY
+902124839513
cemmeydan@su.sabanci-univ.edu

ABSTRACT

Studies in Numerical Taxonomy are carried out by measuring characters as much as possible. The workload over scientists and labor to perform measurements will increase proportionally with the number of variables (or characters) to be used in the study. However, some part of the data may be irrelevant or sometimes meaningless. Here in this study, we introduce an algorithm to obtain a subset of data with minimum characters that can represent original data. Morphological characters were used in optimization of data by Genetic Algorithms Feature Selection method. The analyses were performed on an 18 character*11 taxa data matrix with standardized continuous characters. The analyses resulted in a minimum set of 2 characters, which means the original tree based on the complete data can also be constructed by those two characters.

Categories and Subject Descriptors

J.3 [**Life and Medical Sciences**]: Biology and genetics;

General Terms

Algorithms, Measurement, Experimentation.

Keywords

Genetic algorithms, Optimization, Morphological Data, Phylogenetics, Biological Data Mining.

1. INTRODUCTION

Numerical taxonomy, also known as phenetics, is an attempt to classify organisms based on overall similarity, usually in morphology or other observable traits, regardless of their phylogeny or evolutionary relation [1]. Phenetic techniques include various forms of clustering and ordination. These are sophisticated ways of reducing the variation displayed by organisms to a manageable level. In practice this means measuring dozens of variables, and then presenting them as graphs. Much of the technical challenge in numerical taxonomy revolves around balancing the loss of information in such a reduction against the ease of interpreting the resulting graphs [2]. Since the studies in numerical taxonomy are carried out by the data with the number of characters as much as possible [1], some part of the data may be irrelevant or sometimes meaningless [3]. Recent advances in phyloinformatics have made possible to extract uninformative characters and exclude them from the data in parsimony analysis [4]. However, most of the techniques were implemented for the analysis of molecular sequences. Most recently, two new techniques have been described for inferring phylogenetic trees by using answer set programming [5] and by particle swarm optimization-aided fuzzy cloud classifier [6]. The both methods give optimum solutions to find a subset of characters with minimum number of features. In both methods, only the qualitative characters can be analyzed, since the method was based on character-based cladistics approach. However, morphological data may include various types of characters and can be analyzed by any of the procedures in phylogeny analyses varying on selected phylogenetic approach. If it would be possible to inform scientists about information content within the characters or subset of data with minimum set of characters that gives an acceptable approximate solution, then the work load over the scientist and labor to gathering data will decrease while efficiency in use of time increase. A suggestion to give an exact or most approximate solution to this issue is Genetic Algorithms (GA).

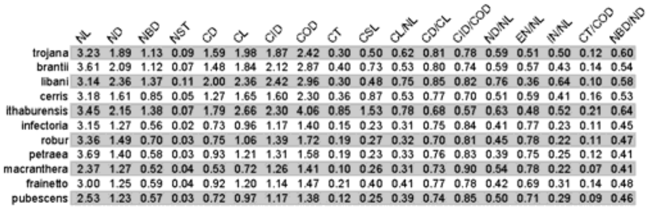
A Genetic Algorithm is a search technique used in computing to find exact or approximate solutions for optimization and search problems [3]. Genetic algorithms are categorized as global search heuristics and are a particular class of evolutionary algorithms (also known as evolutionary computation) that use techniques inspired by evolutionary biology such as inheritance, mutation, selection, and crossover (also called recombination) [3].

2. MATERIALS AND METHOD

A GA method Feature Selection – Subset Selection was used in the study to find the exact or most approximate solution with optimum number of characters. Data with morphological characters were obtained from Bakış 2005 [7]. Oaks are belongs to the family Fagaceae, currently includes nine genera, and *Quercus* is the largest genus among the genera. Cupule is one of the most characteristic and peculiar features of the Fagaceae. Acorns vary greatly in size between and within species, depending on the oak species and its environment [8].

Depending on the type of character encoding, there are plenty of different phylogeny analyzing techniques; only continuous characters were extracted from the data which composed large portion of data. The data with 18 characters and 11 Operational Taxonomic Units (OTUs) has been standardized within characters. Standardization computed for each character by setting minimum value to 0 and maximum to 1 for a 18*11 (Characters*OTUs) matrix. An algorithm was developed to optimize the dataset. It is running on C++, DOS Shell Scripts, and PHYLIP Package 3.67 [9] used for phylogenetic analysis.

Table 1: Morphological data used in the study. Morphological characters versus OTUs.



	NL	ND	NBD	NST	CD	CL	CID	COD	CT	CSL	CL/NL	CD/CL	CID/COD	ND/NL	EN/NL	IN/NL	CT/COD	NBD/ND
trojana	3.23	1.89	1.13	0.09	1.59	1.98	1.87	2.42	0.30	0.50	0.62	0.81	0.78	0.59	0.51	0.50	0.12	0.60
brantii	3.61	2.09	1.12	0.07	1.48	1.84	2.12	2.87	0.40	0.73	0.53	0.80	0.74	0.59	0.57	0.43	0.14	0.54
libani	3.14	2.36	1.37	0.11	2.00	2.36	2.42	2.96	0.30	0.48	0.75	0.85	0.82	0.76	0.36	0.64	0.10	0.58
cerris	3.18	1.61	0.85	0.05	1.27	1.65	1.60	2.30	0.36	0.87	0.53	0.77	0.70	0.51	0.59	0.41	0.16	0.53
ithaburensis	3.45	2.15	1.38	0.07	1.79	2.66	2.30	4.06	0.85	1.53	0.78	0.68	0.57	0.63	0.48	0.52	0.21	0.64
infectoria	3.15	1.27	0.56	0.02	0.73	0.96	1.17	1.40	0.15	0.23	0.31	0.75	0.84	0.41	0.77	0.23	0.11	0.45
robur	3.36	1.49	0.70	0.03	0.75	1.06	1.39	1.72	0.19	0.27	0.32	0.70	0.81	0.45	0.78	0.22	0.11	0.47
petraea	3.69	1.40	0.58	0.03	0.93	1.21	1.31	1.58	0.19	0.23	0.33	0.76	0.83	0.39	0.75	0.25	0.12	0.41
macranthera	2.37	1.27	0.52	0.04	0.53	0.72	1.26	1.41	0.10	0.26	0.31	0.73	0.90	0.54	0.78	0.22	0.07	0.41
frainetto	3.00	1.25	0.59	0.04	0.92	1.20	1.14	1.47	0.21	0.40	0.41	0.77	0.78	0.42	0.69	0.31	0.14	0.48
pubescens	2.53	1.23	0.57	0.03	0.72	0.97	1.17	1.38	0.12	0.25	0.39	0.74	0.85	0.50	0.71	0.29	0.09	0.46

1.1 Genetic Algorithm Method

Data Input: A matrix file containing 18*11 values and a file including character names have entered to program as input files. Delimiter between values is ';' (for the columns) and <ENTER> character delimits OTUs (for the rows). In the initialization part of the algorithm, code parses the file and converts it into an 2D array.

Creating Individuals of Population: For each iteration (generation), a population with certain number of individuals is created. Each individual (child) have different arrangement of chromosomes (characters).

Initializing: Before initialization, a primary population is being generated with individuals each composed of certain number of random chromosomes.

Elitist Selection: To pass the most successful individuals of each generation to the next generation, a certain number of children with lowest fitness score is killed and the parents with highest scores from the previous generation is replaced.

Generations: Children in the initialization use individuals in previous generation as parents. A child of current generation has chromosomes from a parent in previous generation by mutating chromosomes or doing cross-over between two parents.

Score Calculation: to predict which parents are more successful, we calculate fitness scores. The scores will be used to generate next generations (children).

Rank Selection: Children of the current generation will be produced by using the character set of previous generation's parents with a ratio depending on the each parent's fitness. Parent with higher fitness score will have a chance to be used to create children for next generation more than the parents that have fewer score.

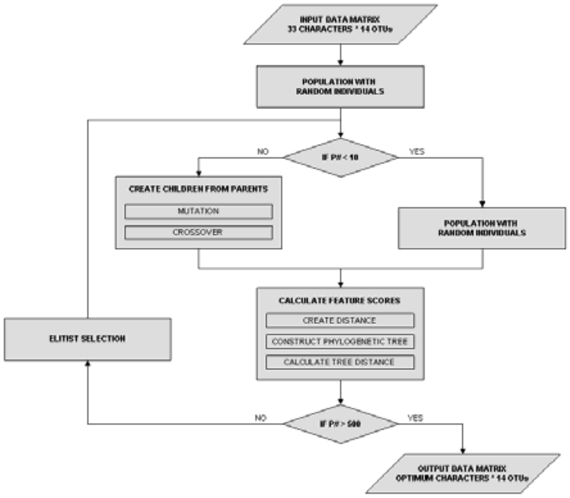


Figure 1: Flow diagram of the optimization algorithm.

1.2 Phylogeny Reconstruction

PHYLogenetic Inference Package (PHYLIP) version 3.67 was used for all analysis in phylogeny reconstruction [9]. For each individual in population, a distance matrix will be created from chromosomes, and then a distance matrix is calculated by using CONTML [10]. NEIGHBOR routine is used to construct phylogenetic tree from the distance matrix. Characters were considered without giving them weights while no out-group has been set. TREEDIST is used to calculate distance between tree with original data and tree with optimized data. Only topological distances between two trees have been calculated since the explanation; "we cannot say whether a larger distance is significantly larger than a smaller one" [9].

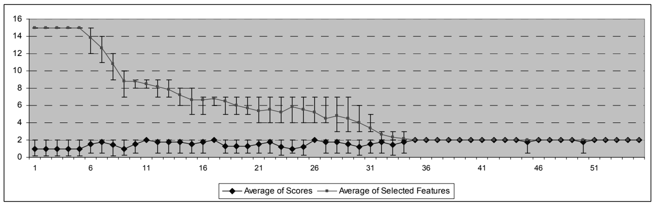


Figure 2: A sample run with 5 preprocessing plus 50 generations. Y error bars represents minimum maximum values.

3. RESULTS

An optimized in the study has been performed on the morphological data based on acorn characters of some Turkish Oaks. After 50 generations, optimization algorithm converges to a solution at average score of 2.0 (Figure 2) and a average number of features at 2, 3 and 4 which means by using only 2 characters one can built exactly the same tree (Table 2). Figure 2 represents a sample solution generated by optimization algorithm. First, a population consisting of randomly created individuals has been created. In the first 5 generation, random individuals are created with a fixed number of features (15) and elitists individuals of the population has been conserved and transferred to next generations. Since the algorithm aims to find a optimum solution, average number of selected features decrease. At a certain point, the algorithm converges to a solution, and no more change would occur after this point even some of the individuals were mutated.

Table 2: Randomly selected 20 solutions (set of features) from optimization algorithm, sorted by number of features in a set.

# OF FEATURES PER SET	NL	ND	NBD	NST	CD	CL	CID	COD	CT	CSL	CL/NL	CD/CL	CID/COD	ND/NL	EN/NL	IN/NL	CT/COD	NBD/ND
2							X							X				
2		X												X				
2		X												X				
2							X							X				
2		X												X				
3	X						X	X										
3	X							X								X		
3		X					X							X				
3	X							X										X
3	X							X								X		
3	X						X									X		
3				X			X								X			
3	X							X					X					
3	X						X	X										
3	X			X			X											
3	X							X								X		
3	X							X										X
3	X							X								X		
4	X	X	X					X										
4	X	X	X					X										
OCCURENCE	13	6	2	2	0	0	8	11	0	0	0	0	1	6	1	5	0	2

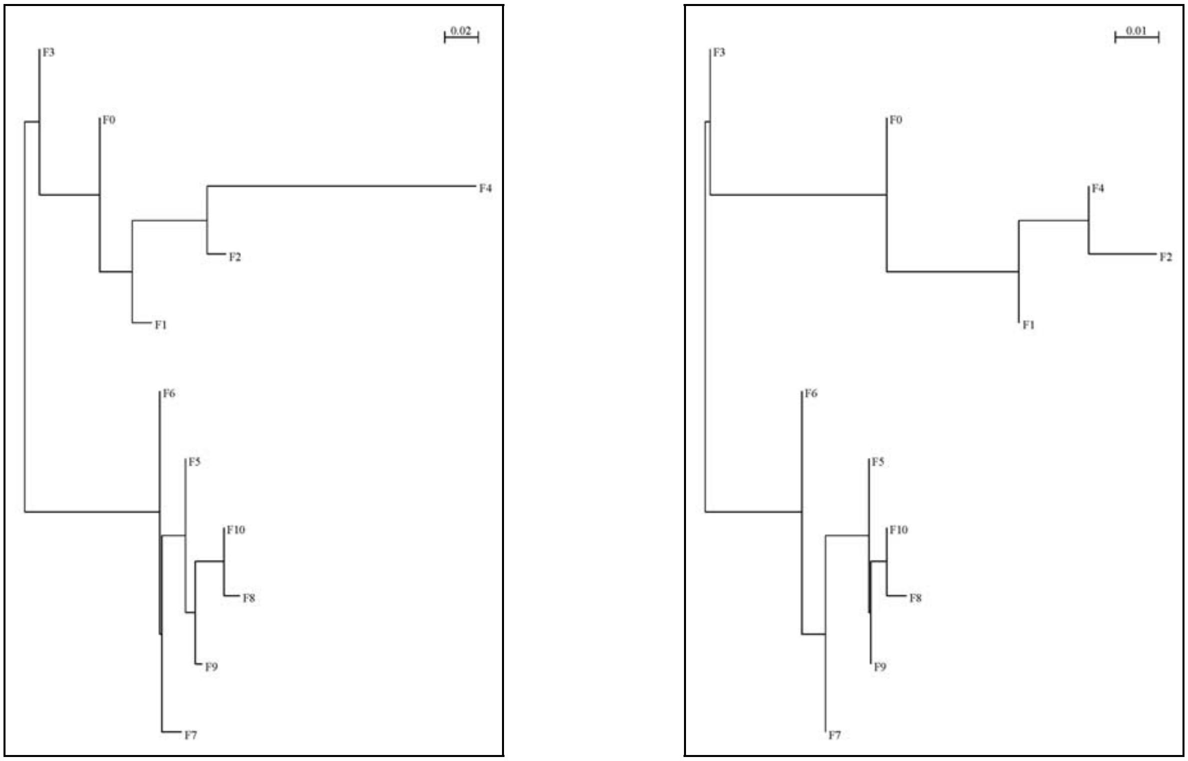
The resulted data (individual) has set of characters (chromosomes) as in table 2. It can be easily observed that some of the features were involved in the data sets many times, while some others were never occurred in any solution. The characters occurred in different sized sets were also showing differences.

Sample trees obtained from original dataset and evaluated data sets were placed in Figure 3. Even there some branch length distances occurs between the trees, they have exactly the same topology.

4. DISCUSSION

The morphological data based on acorn characters of some Turkish Oaks has been optimized in the study. Minimum solution set with 2 characters has been obtained. Table 2 shows the 20 sample run with occurrences of the characters. Some of the characters did never involve in any of the solution sets, which are mostly cupule based characters. Cup morphology had been represented in original data with 6 characters while nut had been represented with 4. It appears like some of the cup characters are not informative as some others are.

The most represented morphological characters, NL and COD, were derived from different parts of organ – which seems so reasonable – from fruit and from cup. However, these were not the ones that were involved in optimum data with 2 features. This is because the information contents of features of optimum data were overlapping, and thus represents whole data. ND/NL is the only one that is found in every 2 set solution. The reason would be the twice the information content of index characters and also the two dimension information of nut morphology, length and diameter.



```
((F3:0.000,(F6:0.000,(F7:0.011,(F5:0.000,(F9:0.004,(F10:
0.000,F8:0.009):0.016):0.006):0.014):0.001):0.089):0.035,
((F2:0.011,F4:0.159):0.044,F1:0.011):0.019,F0:0.000);
```

```
((F3:0.000,((F7:0.000,(F5:0.000,((F8:0.004,F10:0.000):
0.003,F9:0.000):0.001):0.010):0.005,F6:0.000):0.023):0.041
,(F1:0.000,(F4:0.000,F2:0.015):0.016):0.030,F0:0.000);
```

Figure 3: Phylogenies produced by original data (left figure and newick format) and by solution set (right figure and newick format).

An interesting result is nut diameter's and cupule inner diameter's occurrence with in the ND/NL index in a 2 character set solution. Both derived from the same origin actually, one is the diameter of nut, and another is the inner diameter of cup, which is diameter of nut at cup mouth. In any ways, the resultant solutions gave us two characters as representing whole dataset; the nut diameter and nut length.

5. REFERENCES

[1] R. R. Sokal, "Numerical taxonomy" *Scientific American,* vol. 215, no. 6, pp. 106-116;, 1966.

[2] W. J. L. Quesne, "A Method of Selection of Characters in Numerical Taxonomy" *Systematic Zoology* vol. 18 no. 2, pp. 201-205 1969

[3] M. Mitchell, *An introduction to genetic algorithms*, Cambridge, Mass.: MIT Press, 1996.

[4] D. Swofford. "PAUP* 4.0," 2009; http://paup.csit.fsu.edu/.

[5] D. R. Brooks, E. Erdem, S. T. Erdogan *et al.*, "Inferring phylogenetic trees using answer set programming," *Journal of Automated Reasoning,* vol. 39, no. 4, pp. 471-511, Dec, 2007.

[6] E. P. Hongfei Lu, Qiufa Peng, Lanlan Wang, Changjiang Zhang, "A particle swarm optimization-aided fuzzy cloud classifier applied for plant numerical taxonomy based on attribute similarity," *Expert Systems with Applications,* vol. 36, pp. 9388-9397, 2009.

[7] Y. Bakış, "Morphometric Analysis of Oak (Quercus L.) Acorns in Turkey," Graduate School Of Natural And Applied Sciences, Abant İzzet Baysal University, Bolu, 2005.

[8] R. J. Jensen, "The Quercus falcata Michx. Complex in Land Between The Lakes Kentucky and Tennessee; a Study of Morphological Variation," *American Midland Naturalist,* vol. 121, pp. 245-255, 1989.

[9] J. Felsenstein. "PHYLIP Home Page," 2009; http://evolution.genetics.washington.edu/phylip.html.

[10] J. Felsenstein, "Maximum-likelihood estimation of evolutionary trees from continuous characters," *Am J Hum Genet,* vol. 25, no. 5, pp. 471-92, Sep, 1973.