A Computational Model of the Spread of Ancient Human Populations Based on Mitochondrial DNA Samples

Peter Z. Revesz

***Abstract*—** The extraction of mitochondrial DNA (mtDNA) from ancient human population samples provides important data for the reconstruction of population influences, spread and evolution from the Neolithic to the present. This paper presents a mtDNA-based similarity measure between pairs of human populations and a computational model for the evolution of human populations. In a computational experiment, the paper studies the mtDNA information from five Neolithic and Bronze Age populations, namely the Andronovo, the Bell Beaker, the Minoan, the Rössen and the Únětice populations. In the past these populations were identified as separate cultural groups based on geographic location, age and the use of, decoration or shape of cultural artifacts.

***Keywords*—**Evolution, Mitochondrial DNA, Population Genetics, Similarity Measure, Phylogenetic Tree.

I. Introduction

Recent advances in biotechnology enable the extraction of ancient mitochondrial DNA (mtDNA) from human bones going back thousands of years. These advances already facilitated several studies of the origin and spread of various mtDNA types, called haplogroups. However, most human populations are highly heterogeneous in terms of their mtDNA haplogroup compositions. Hence even with the newly available mtDNA information, it is not obvious how human populations spread geographically over time. In particular, there are two main challenges for such studies.

The first challenge in studying the relationships among human populations is to develop an easy-to-compute and flexible similarity measure between pairs of human populations based on mtDNA samples from those two populations. Flexibility in this case means that the similarity measure has to accommodate mtDNA haplogroups that are defined to an arbitrary depth or level. For example, we need to be able to compare a relatively short haplogroup description, such as H5 with a long haplogroup description, such as H1a5b2. We define in Equation (2) below for any pair of populations an overall similarity measure that is both easy-to-compute and flexible.

Once a pairwise overall similarity measure is defined, it is possible to build a similarity matrix for all the populations for which mtDNA sample data is available. The second challenge is making valid inferences from the similarity matrix regarding the mutual interaction and spread of human populations. In the area of phylogenetics, which is the study of biological phyla, similarity matrices are used to derive a hypothetical evolutionary tree of the phyla [1]-[4]. However, the algorithms that build hypothetical evolutionary trees, such as Neighbor Joining [9], UPGMA [11] and the common mutations similarity matrix (CMSM) algorithm [5] may not be applicable to the study of human populations for several reasons. First, the time scale of phyla evolution is vast compared to the time scale of the development of human populations. The evolution of biological phyla may take millions of years [10], [12], while ancient human mtDNA samples do not go back more than about ten thousand years. Second, while biological phyla diverge from each other in genetic isolation, when human populations come in contact with each other, they tend to merge their genetic pool. Therefore, the set of mtDNAs in a human population may come from several different ancestor human populations that were each more homogeneous in their mtDNA compositions. In general, if P_1 and P_2 are two human populations with set of mtDNAs S_1 and S_2, respectively, such that the condition

$$S_1 \subseteq S_2 \tag{1}$$

holds, then P_1 can be assumed to be an ancestor of P_2. However, the reverse is not true. In other words, P_1 may be an ancestor of P_2 but the above condition may not hold because either not all mtDNAs were transferred from P_1 to P_2 or some of the transferred mtDNAs have evolved to a different form.

This paper is organized as follows. Section II presents a computational model of the overall similarity between two populations based on mitochondrial DNA haplogroup samples from the two populations. Section III describes experimental results based on five different ancient Neolithic and Bronze Age European populations. These populations are identified by and associated with different cultural artifacts and were not considered related. However, the material culture can change over time to a degree that the relationships among various cultures become unrecognizable. Our experimental study reveals which populations are closer or more distantly related with each other. Finally, Section IV gives some conclusions and directions for future work.

Peter Z. Revesz is with the Department of Computer Science and Engineering, University of Nebraska-Lincoln, Lincoln, NE 68588, USA (revesz@cse.unl.edu).

II. A COMPUTATIONAL MODEL

The degree of relatedness between two individuals can be estimated based on a comparison of their mtDNA haplogroups. In this paper, we use the mtDNA haplogroup classification provided by PhyloTree.org at http://www.phylotree.org.

We say that a *level 1 relationship* exists between two individuals if they belong to the same haplogroup (single capital letter) but do not share further classifications. We say that a *level 2 relationship* exists between two individuals if they belong to the same sub-halogroup (capital letter and number) but do not share further classifications. In general, we say that a *level n relationship* exists between two individuals if their haplogroup classifications share the first *n* elements. For example, H1a2 and H1a5b have a level 3 relationship because they share H1a, that is, three elements, namely the haplogroup H, the sub-haplogroup H1 and the sub-sub-haplogroup H1a.

Note that the largest shared level is a unique number for any pair of haplogroups. This allows us to define the function

$$Level: s_1 \times s_2 \rightarrow N$$

which takes as input two haplogroups s_1 and s_2 and returns the maximum level numbering of the relationship that exists between them. For example,

$$Level(H1a2, H1a5b) = 3.$$

We also define the *weight function*

$$W: N \rightarrow N$$

which takes as input a level number and returns a weight value. For example, W(3) returns the weight of level 3 relationships. The weight is intended to describe the degree of unusualness of the existence of a relationship. Normally we would expect the weights to increase exponentially in value because the mtDNA haplogroup tree has many branches at all levels.

We define the *overall similarity* between two populations P_1 and P_2 with associated mtDNA samples S_1 and S_2, respectively, by the following equation:

$$sim(P_1, P_2) = \frac{\sum_{a \in S_1,\ b \in S_2} W(Level(a, b))}{n \times m} \qquad (2)$$

where *n* and *m* are the number of samples in the two populations. Here P_1 and P_2 are bags (instead of sets) and can contain repetitions. Equation (2) says that the similarity of two populations equals to the weighted sum of the relationships between pairs of individuals from the two populations divided by the total number of possible pairs. Overall similarity measures closely related to Equation (2) were previously studied also in arbitration theory [7], [8] and cancer research [6]. Equation (2) defines a symmetric relation. Hence

$$sim(P_1, P_2) = sim(P_2, P_1) \qquad (3)$$

Equation (2) could be further refined if we would know precisely the probabilities of each haplogroup because then we could select the weigh function to return for each level a value that is in inverse proportion to the probability that two random haplogroup samples have the given level of relationship.

Although Equation (2) could be improved with more statistical information, it is a good first approximation of the overall similarity between two populations. For simplicity, in this paper we assume that the weight function contains the following:

$$W(1) = 0$$
$$W(2) = 0$$
$$W(3) = 1$$
$$W(4) = 5$$
$$W(5) = 25.$$

III. EXPERIMENTAL RESULTS

A. The mtDNA Database

We obtained mtDNA data from five ancient populations from the website http://suyun.info/index.php?p=ancientdna which lists the source and age of the samples and classifies them according to cultural groupings. From that database, we selected the following five ancient populations:

1. **Andronovo culture**: The Andronovo culture, which is noted for the domestication of horses and burial in kurgans, flourished in the steppe region to the north and the east of the Caspian Sea in today's Kazakhstan and Russia [13]. The database contains nine Andronovo mtDNA samples dated 1800 – 1400 BC.

   Andronovo = {H6, K2b, T1a, T2a1b1, U2e, U4, U4, U5a1, Z1}

2. **Bell Beaker culture**: The Bell Beaker culture is a prehistoric Western European culture that was named after its characteristic bell-shaped pottery [14]. Some megalithic structures, for example, Stonehenge is associated with the Bell Beaker culture [14]. The database contains eighteen Bell Beaker mtDNA samples dated 2600 – 2050 BC.

   Bell_Beaker = {H, H, H1, H1e7, H3, H3b, H4a1, H5a3, H13a1a2c, I1a1, J, K1, T1a, U2e, U4, U5a1, U5a2a, W5a}.

3. **Minoan culture:** The Minoan culture flourished on Crete, Santorini and some other Aegean islands [15]. The Minoan culture is noted for building the ancient palace of Knossos that is associated with the mythical labyrinth where King Minos supposedly hid the Minotaur [15]. The database contains 34 Minoan mtDNA samples dated 2400 – 1700 BC.

ISBN: 978-1-61804-350-4

Minoan = {H, H, H, H, H, H, H5, H7, H13a1a, HV, HV, HV, I5, I5, I5, J2, K, K, K, K, K, K, R0, T, T1, T2, T2, T2, T3, T5, U, U5a, W, X}.

In some cases, the haplogroup classification can be refined based on the http://www.phylotree.org website that gives an mtDNA classification tree based on the known mutations that characterizes each branch. The PhyloTree.org classification tree also changed slightly since the Minoan study was done. For example, in the latest version (February 19, 2014) the classifications T3 and T5 are now placed within the T2 branch. Using the updated classifications, the Minoan group can be refined as follows, where the updated values are highlighted in blue:

$Minoan_2$ = {H, H, H, H, H, H, H5a1b, H7, H13a1a, HV, HV, HV, I5, I5, I5, J2, K, K, K, K, K, K, R0, T2, T1a, T2, T2, T2, T2, T2e, U, U5a1f1/U5a2e, W, X}.

Note that the H5a1b identification is possible because of the mutation 11719A. Note also that U5a can be expanded to either U5a1f1 or U5a2e because both of these contain the 16311C mutation.

4. **Rössen culture**: The Rössen culture is a Neolithic Central European culture that built settlements consisting of trapezoidal or boat-shaped long houses [16]. The database contains ten mtDNA samples dated 4625 – 4250 BC.

Rössen = {H1, H5b, H16, H89, HV0, K, N1a1a, T2, T2e, X2j}

5. **Únětice culture**: The Únětice culture is a Bronze Age culture with sites known from Central Germany, the Czech Republic and Slovakia [17]. The Únětice culture is noted for the Nebra Sky disk and other metal artifacts [17]. The database contains twenty mtDNA samples dated 2200 – 1800 BC.

Únětice = {H11a, H2a1a3, H82a, H4a1a1a5, H3, H7h, I, I1, T1, T2, T2, T2b, U, U2, U5a1, U5a1a, U5b, W, X}

B. Computation of a Similarity Matrix

Using Equation (2), we computed the similarity matrix for the five ancient populations as shown in Table 1. Note that the similarity matrix is symmetric because of Equation (3). Each non-diagonal entry of the similarity matrix contains the overall similarity value between two different populations described in the corresponding row and column.

Table 1 Similarity matrix among five different ancient populations.

	Andro	Bell-B	Minoan	Rössen	Únětice
Andro.		.0432	.0196	0	.0556
Bell-B.	.0432		.0523	0	.0417
Minoan	.0196	.0523		.0029	.0074
Rössen	0	0	.0029		0
Únětice	.0556	.0417	.0074	0	

Table 2 shows pairs of mtDNA samples that indicate a level 3 or higher distant relationship between the Andronovo and the Bell Beaker populations.

Table 2 Andronovo and Bell Beaker relationships.

Andronovo	**Bell Beaker**	**Relationship Level**
T1a	T1a	3
U2e	U2e	3
U5a1	U5a1	4

Hence by Equation (2) the overall similarity between the Andronovo and the Bell Beaker populations can be calculated to be:

$$sim(Andronovo, Bell\_Beaker) = \frac{1 + 1 + 5}{9 \times 18} = 0.0432$$

Similarly, when we compare the Bell Beaker and the Minoan samples, Table 3 shows the pairs that indicate a level 3 or higher relationship.

Table 3 Bell Beaker and Minoan relationships.

Bell Beaker	**$Minoan_2$**	**Relationship Level**
H5a3	H5a1b	3
H13a1a2c	H13a1a	5
T1a	T1a	3
U5a1	U5a1f1	4

Hence the overall similarity between the Bell Beaker and the $Minoan_2$ populations is:

$$sim(Bell\_Beaker, Minoan_2) = \frac{1 + 25 + 1 + 5}{18 \times 34} = 0.0523$$

As another example, in comparing the $Minoan_2$ and the Rössen populations only one pair indicates a level 3 or higher relationship as shown in Table 4.

Table 4 Bell Beaker and Rössen relationships.

$Minoan_2$	**Rössen**	**Relationship Level**
T2e	T2e	3

Hence the overall similarity between the $Minoan_2$ and the Rössen populations can be calculate to be:

ISBN: 978-1-61804-350-4

$$sim(Minoan_2,\ Rössen) = \frac{1}{34 \times 10} = 0.0029$$

Likewise, we can calculate the following similarity values:

$$sim(Andronovo, Minoan) = \frac{1+5}{9 \times 34} = 0.0196$$

$$sim(Andronovo, \text{Únětice}) = \frac{5+5}{9 \times 20} = 0.0556$$

$$sim(Bell\_Beaker, \text{Únětice}) = \frac{5+5+5}{18 \times 20} = 0.0417$$

$$sim(Minoan, \text{Únětice}) = \frac{5}{34 \times 20} = 0.0074$$

Finally, between the Andronovo and the Rössen, the Bell Beaker and the Rössen, and the Rössen and the Únětice populations, no pair of samples shows a level 3 or higher relationship. Hence

$$sim(Andronovo, Rössen) = 0$$
$$sim(Bell\_Beaker, Rössen) = 0$$
$$sim(Rössen, \text{Únětice}) = 0$$

C. Discussion of the Results

In general, the higher is the similarity value between two populations, the more closely related those two populations are. According to that intuition, the highest similarity (0.0556) is between the Andronovo and the Únětice populations as shown Table 1. There is an almost equally high similarity (0.0523) between the Bell Beaker and the Minoan populations. The results also reveal that the Rössen population is only related with the Minoan populations with a relatively small similarity (0.0029).

Perhaps a deeper insight can be gained from the data if we also consider which haplogroups are the major links (level 3 or higher relationships) between each pair of populations.

Table 5 shows that the Minoan and the Rössen populations are related via the T2e haplogroup, while the Minoan and the Bell Beakers populations are related via H5a, H13a, T1a and U51a haplogroups. The major links between the Andronovo and the Minoan populations are via T1a and U5a1 haplogroups, while the U5a1 haplogroup is the only major link between the Minoan and the Únětice populations.

It needs to be mentioned that in the current database many of the ancient mtDNA samples are only fragments instead of complete mtDNAs. Hopefully, the haplogroup classifications may be further refined with improved testing methods in the future. The results may change slightly as some haplogroup classifications are extended from two to three or more letters. Nevertheless, it seems extremely unlikely that the refinement of some of the mtDNA classifications would change the current clustering picture instead of further strengthening the already existing groupings.

Table 5 The level 3 or higher haplogroup relationships among the five different ancient populations. Only the entries in the upper triangular part of the matrix are shown because the matrix is symmetric.

	Andro	Bell-B	Minoan	Rössen	Únětice
Andro.		T1a U2e U5a1	T1a U5a1		U5a1
Bell-B.			H5a H13a1a T1a U5a1		H4a1 U5a1
Minoan				T2e	U5a1
Rössen					
Únětice					

The experimental results suggest either T2e gene flows between or a common origin of the Minoan and the Rössen populations. Similarly, the results suggest either H5a and H13a1a gene flows between or a common origin of the Bell Beaker and the Minoan populations. The origin and the dispersal of the T1a and U5a1 haplogroups are less clear because they are shared more widely among the five studied populations.

Unfortunately, the mtDNA data does not allow drawing conclusions regarding the language associated with each of the five sample populations in this study. However, either gene flows or common origin between pairs of populations raises the chance of similarity of language. Hence some language similarity is plausible between Minoan and Rössen and between Bell Beaker and Minoan.

IV. CONCLUSIONS AND FUTURE WORK

In this paper, we defined a measure for the overall similarity between two populations with mtDNA haplogroup samples. Our study is not merely the study of the dispersal of various mtDNA haplogroups but the dispersal of various populations that are already heterogeneous in terms of their mtDNA compositions.

Our mtDNA haplogroup-based population similarity measure could be extended easily to a Y-DNA haplogroup-based population similarity measure. It would be interesting to perform a similar analysis on Y-DNA data for the populations studied in this paper and compare the similarity matrices generated by the mtDNA and the Y-DNA haplogroup-based data. However, ancient Y-DNA data is much harder to obtain than ancient mtDNA data with current technology. Hence such a Y-DNA study may have to wait until further DNA extraction technology improvements.

Another way to extend the research is to study a larger number of populations. We intend to study more ancient populations as well as currently living populations to gain more insight into the origin and dispersal of various populations. The five populations were all ancient Neolithic or Bronze Age European cultures. Considering populations that encompass a broader time scale and a larger geographic area may give a deeper insight into human pre-history.

ISBN: 978-1-61804-350-4

REFERENCES

[1] D. Baum and S. Smith, *Tree Thinking: An Introduction to Phylogenetic Biology*, Roberts and Company Publishers. 2012.

[2] B. G. Hall, *Phylogenetic Trees Made Easy: A How to Manual*, 4th edition, Sinauer Associates, 2011.

[3] P. Lerney, M. Salemi, and A.-M Vandamme, editors. *The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing*, 2nd edition, Cambridge University Press, 2009.

[4] P. Z. Revesz, *Introduction to Databases: From Biological to Spatio-Temporal*, Springer, New York, 2010.

[5] P. Z. Revesz, "An algorithm for constructing hypothetical evolutionary trees using common mutations similarity matrices," *Proc. 4th ACM International Conference on Bioinformatics and Computational Biology*, ACM Press, Bethesda, MD, USA, September 2013, pp. 731-734.

[6] P. Z. Revesz and C. J.-L. Assi, "Data mining the functional characterizations of proteins to predict their cancer relatedness," *International Journal of Biology and Biomedical Engineering*, 7 (1), 2013, pp. 7-14.

[7] P. Z. Revesz, "On the semantics of arbitration," *International Journal of Algebra and Computation* , 7 (2), 1997, pp. 133-160.

[8] P. Z. Revesz, "Arbitration solutions to bargaining and game theory problems," *Annales Universitatis Scientarium Budapestinensis, Sect. Comp.* , 43, 2014, pp. 21-38.

[9] N. Saitou and M. Nei, "The neighbor-joining method: A new method for reconstructing phylogenetic trees," *Molecular Biological Evolution*, 4, 1987, pp. 406-425.

[10] M. Shortridge, T. Triplet, P. Z. Revesz, M. Griep, and R. Powers, "Bacterial protein structures reveal phylum dependent divergence," *Computational Biology and Chemistry*, 35 (1), 2011, pp. 24-33.

[11] R. R. Sokal, and C. D. Michener, "A statistical method for evaluating systematic relationships," *University of Kansas Science Bulletin*, 38, 1958, pp. 1409-1438.

[12] T. Triplet, M. Shortridge, M. Griep, J. Stark, R. Powers, and P. Z. Revesz, "PROFESS: A protein function, evolution, structure and sequence database," *Database -- The Journal of Biological Databases and Curation*, 2010, Available: http://database.oxfordjournals.org/content/2010/baq011.full.pdf+html

[13] Wikipedia, "Andronovo culture," downloaded August 19, 2015. Available: https://en.wikipedia.org/wiki/Andronovo_culture

[14] Wikipedia, "Beaker culture," downloaded August 19, 2015. Available: https://en.wikipedia.org/wiki/Beaker_culture

[15] Wikipedia, "Minoan civilization," downloaded August 19, 2015. Available: https://en.wikipedia.org/wiki/Minoan_civilization

[16] Wikipedia, "Rössen culture," downloaded August 19, 2015. Available: https://en.wikipedia.org/wiki/Rössen_culture

[17] Wikipedia, "Unetice culture," downloaded August 19, 2015. Available: https://en.wikipedia.org/wiki/Unetice_culture

Peter Z. Revesz holds a Ph.D. degree in Computer Science from Brown University. He was a postdoctoral fellow at the University of Toronto before joining the University of Nebraska-Lincoln, where he is a professor in the Department of Computer Science and Engineering. Dr. Revesz is an expert in bioinformatics, databases, data mining, and data analytics. He is the author of *Introduction to Databases: From Biological to Spatio-Temporal* (Springer, 2010) and *Introduction to Constraint Databases* (Springer, 2002). Dr. Revesz held visiting appointments at the IBM T. J. Watson Research Center, INRIA, the Max Planck Institute for Computer Science, the University of Athens, the University of Hasselt, the U.S. Air Force Office of Scientific Research and the U.S. Department of State. He is a recipient of an AAAS Science & Technology Policy Fellowship, a J. William Fulbright Scholarship, an Alexander von Humboldt Research Fellowship, a Jefferson Science Fellowship, a National Science Foundation CAREER award, and a "Faculty International Scholar of the Year" award by *Phi Beta Delta*, the Honor Society for International Scholars.