A Last Genetic Contact Tree Generation Algorithm for a Set of Human Populations

Peter Z. Revesz
Computer Science and Engineering
University of Nebraska-Lincoln
Lincoln, NE 68588-0115
1 402 472 3488
revesz@cse.unl.edu

ABSTRACT

The rapidly growing amount of ancient human genetic data enables the tracking of the spread of human populations over time. However, several challenges that need new solutions in order to most effectively mine the available data. We introduce an efficient algorithm that generates a last genetic contact tree for a set of populations. The computation complexity of the algorithm is shown to be $O(n^3)$ where n is the number of populations. The algorithm requires a preprocessing time to set up a similarity matrix.

Categories and Subject Descriptors

J.3 [**Computer Applications**]: Life and Medical Sciences – *biology and genetics*

General Terms

Algorithms; theory.

Keywords

Clustering; last genetic contact; phylogenetic tree; population genetics; similarity matrix.

1. INTRODUCTION

Phylogenetic tree construction algorithms [1, 3, 4], including UPGMA [12], Neighbor Joining [10] and our CMSM (*Common Mutations Similarity Matrix*) [6] phylogenetic tree algorithm, are frequently used to analyze the evolutionary relationships among a set of genomes based on a growing number of gene and protein databases [2, 5, 11, 13]. However, these phylogenetic algorithms cannot be applied to cases when we need to consider the evolutionary relationship of a set of populations, where each population is a set of heterogeneous genomes. That is because in a phylogenetic tree the leaves are individual genomes instead of a set of genomes.

It is tempting to try to reduce the problem of dealing with a set of populations to the problem of dealing with a set of genomes by simply selecting a single representative genome from each population. However, the task of finding representatives of different human populations is hard because human populations are genetically heterogeneous.

Hence we cannot find a good representative for each population and then compare just those representatives to see the relatedness of the various populations. To efficiently analyze the evolutionary relationship among a set of heterogeneous populations requires a more sophistical algorithmic approach than the ones provided by the current phylogenetic tree generation algorithms.

In a preliminary work on this subject, we proposed comparing mitochondrial DNA (mtDNA) population samples by taking the pairs that have the most number of matching level [14]. For example, H4a and H4b2 have two levels in common, that is, H4 is the common part. However, in this paper we use a more natural similarity of two haplogroups by considering the number of mutations that they share. In addition, we develop in this paper a novel last genetic contact algorithm that can generate a tree for a set of human populations. In this tree, each leaf represents a population, and closer leaves have had more recent genetic contact than leaves that are further apart.

This paper is organized as follows. Section 2 describes the last-genetic-contact algorithm and analyzes its computational complexity. Section 3 gives some conclusions and future work.

2. THE LAST GENETIC CONTACT TREE GENERATION ALGORITHM

Behind our algorithm the main intuition, which we explain by choosing mtDNAs, is the following. Suppose that some ancestor human population A at some point of time contains mtDNA haplogroups h_1, h_2, … , h_n, and splits into two subpopulations B and C. Then populations B and C can be expected to originally contain the same set of haplogroups. However, these haplogroups will gradually diverge from each other due to random mutations and evolutionary change. The more time passes, the more distant even the closest pair of haplogroups in B and C become. This can be measured as follows.

Let the *Hamming Distance,* or *HD* for short, be the number of mutations by which two mtDNA haplogroups differ. These mutations can be random changes of nucleotides or deletions or insertions. Then we define the overall distance between two populations B and C as follows:

$$Dist(B,C) = Min\,\{HD(h_i,h_j):\ h_i \in B,\ h_j \in C\}$$

BCB '16, October 2–5, 2016, Seattle, WA, USA
Copyright 2016 ACM 978-1-4503-4225-4/16/10.
http://dx.doi.org/10.1145/2975167.2985654

where each h_k is some mtDNA haplogroup of some sample.

If B and C move to different areas, then they incorporate new haplogroups. We can expect these new haplogroups to pay no role in defining Dist(B,C) because the newly incorporated haplogroups are likely to be greatly different from each other.

For example, suppose that the original population A is the British population of the past, population B is the present population of a former British colony in Africa, and population C is the present population of a former British colony in Asia. Then in any small sample of the populations, Dist(B, C) can be expected to depend on the closeness of some pair of British-origin genes in B and C rather than a pair of indigenous African and Asian genes. Moreover, Dist(B, C) gives an estimate when the colonization ended because since then relatively few people moved from the UK to the former colonial territories. Moreover, the relative percentage of the indigenous African and the indigenous Asian population to the percentage of British-origin population does not effect the Dist(B, C) measure as long as the samples of B and C contain enough British-origin genes.

While distance matrices are used by UPGMA, our last genetic contact (LGC) algorithm uses instead a type of similarity matrix. We set up the similarity matrix as follows. Suppose we have a common origin for all mtDNA samples in all the populations that we consider. For instance, haplogroup N* may be the common origin of all the samples we use. We count from this common ancestor the number of mutations on which two mtDNA samples agree. For example, a K1b1a1 and a K1b1a2 haplogroup share the K1b1a haplogroup, which has 27 common mutations from the root N*. We denote this as *ComMut(K1b1a1, K1b1a2) = 27*. We define the similarity between two different populations B and C as follows:

$$Sim(B,C) = Max\left\{ComMut\left(h_i, h_j\right) :\ h_i \in B,\ h_j \in C\right\}$$

Hence for each pair of populations we find the highest number of mutations on which any pair of haplogroup samples from both populations agrees. The algorithm takes as input S and a similarity matrix M, where $M[i,j] = Sim\left(P_i, P_j\right)$ when $i \neq j$ and $M[i,j] = 0$ otherwise. The algorithm is as follows:

```
ALGORITHM Last-Genetic-Contact(S, M)
Create an independent node $N_i$ for each population Si in S.
Let $N = \{ N_k : 1 \leq k \leq n \}$.
While ($N$ is not empty) Do
    Find $P_i$ and $P_j$ in S such that $Sim(P_i, P_j)$ is maximum.
    Merge the nodes $N_i$ and $N_j$ associated with $P_i$ and $P_j$,
    creating a parent node $N_{ij}$ with children $N_i$ and $N_j$.
    $N = \left(N - \{N_i, N_j\}\right) \cup N_{ij}$
    Update the similarity matrix by deleting the rows and the
    columns associated with $P_i$ and $P_j$ and creating new row for $P_{ij}$
    and a new column for $P_{ij}$, which represent the merge of $P_i$ and $P_j$.
    Let the merged entry be in column c. Then for all remaining rows
    and columns indexed by k we have:
        $M[c,k] = Max(M[i,k], M[j,k])$ and
        $M[k,c] = Max(M[k,i], M[k,j])$
End-While
Return the tree generated
```

Theorem 1 The Last-Genetic-Contact algorithm runs in O(n^3) computational time complexity for an input matrix of size n × n with n number of populations that were sampled.

3. CONCLUSIONS AND FUTURE WORK

In the future we plan to apply our last genetic contact algorithm to the study of the spread of ancient human populations using samples in: http://www.ancestraljourneys.org/ancientdna.shtml.

4. REFERENCES

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

[2] Billa, S., Griep, M., and Revesz, P. Z. 2011. Approximate search on protein structures for identification of horizontal gene transfer in bacteria, *Proc. 9 International Symposium on Abstraction, Reformulation and Approximation*, AAAI Press, pp. 18-25, Cardona, Spain.

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

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

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

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

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

[8] Revesz, P. Z. and Triplet, T. 2010. Classification integration and reclassification using constraint databases. *Artificial Intelligence in Medicine,* 49 (2), 79-91.

[9] Revesz, P. Z. and Triplet, T. 2011.Temporal data classification using linear classifiers, *Information Systems*, 36 (1), 30-41.

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

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

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

[13] Triplet, T., Shortridge, M., Griep, M., Stark, J., Powers, R., Revesz, P. Z.. 2010. PROFESS: A protein function, evolution, structure and sequence database. *Database -- The Journal of Biological Databases and Curation*, doi:10.1093/baq011

[14] Revesz, P. Z. 2016. A mitochondrial DNA-based model of the spread of human populations, *International Journal of Biology and Biomedical Engineering*, 10 (1), 124-133.