Selective pressure on DNA physical properties: A computational approach

Matthew Cooke

McGill University

December, 2016

DNA mutations are essential to evolution. They are the cause of genetic variation. DNA structure however, between the mutations, is hypothesized to be mostly conserved between close ancestors, as to not greatly affect the function of the gene. This was tested by statistically comparing the effect of mutations on secondary structure MFE, minimum free energy, from a human-primate ancestor (labelled HP) to human, and comparing it with a fictional mutation of human-primate ancestor to a sample data set with 1 random mutation (labelled Random). The hypothesis was that the mean difference in MFE would be significantly lower in the real evolutionary data compared to the random sample data. I conclude that, although the means differed in the direction predicted, the difference was not significant enough to discount type I error. Future applications of the algorithm could help refine the algorithm and retest the hypothesis.

1. Introduction

DNA is ubiquitous in the cell and is extremely important. DNA effects are determined by its structure, the DNA folded onto itself, which is in turn affected by its sequence. (Alberts et al, 2002). This folding process generates a secondary structure. This secondary structure of DNA therefore predicts the effect the gene has on the species.

Mistakes are made during DNA replication. These errors affect the folding of the secondary structure and in turn affect the function of the gene. These mistakes are the ultimate source of genetic variation. Mutations occur throughout the genome at a rate of ~1x10^-8 per site per generation in humans (Nachman et al, 2000). These rates vary from species to species and even at different sites within the same species.

Mutations in DNA, and in turn RNA, are what cause variability among all existing species. Most gene-coding regions in an organism do not allow mutations with large fitness effects, as these effects prohibit the organism's ability to procreate and survive (Eyre-Walker, 2006). Because DNA secondary structure is extremely important for the function of DNA, I predict that mutations that have minimal effects on secondary structure would be more likely to be found in close ancestors. Evolution desires to preserve function and secondary structure predicts the function of the gene.

Today's different species can be traced back to their evolutionary origins (Cole et al., 2005). Phylogenetic trees inferring evolutionary relationships among various biological species or other entities, based upon similarities and differences in their physical or genetic characteristics, have been developed (Letunic et al, 2006). In this study, this evolutionary assumption forms the basis for the hypothesis that because of the effect secondary structure has on the function of DNA segments, close evolutionary ancestors' DNA mutations should affect their secondary structure less than that of random mutations.

This study makes many simplifying assumptions about mutation specificity, as well as mutation rate. It assumes that in our gene-segment there will always be one and only one mutation. It also assumes that this mutation has an equal probability of mutating any given nucleotide to any other. These two assumptions are not true to nature. However, if it can be shown that there is a significant difference between these pseudo-real random mutations and our evolutionary data, generalizations to fit real random mutations can be made.

Other studies have used phylogeny to help improve secondary structure prediction accuracy (Zuker et. al., 1991). This method differs from previous studies in that it examines the relationship between secondary structure preservation and ancestral relation. To the best of my knowledge, this has never been previously examined.

2. Methodology

I am interested in comparing the secondary structure of DNA between close evolutionary ancestors (HP-Human) and minimally mutated random data (HP-Random). I've hypothesized that the ancestral data should have less effect on DNA secondary structure than random point mutations.

I began with a multiple sequence alignment of human chromosome 22 to other genetically related sequences. For this paper, I was only interested in secondary structure differences between a Human-Primate ancestor (HP) and Human. To compare this data to random point mutations, single nucleotide mutations on the HP data were generated to create a Random dataset. Using mutation data given in .maf format, I extracted all of the data pertaining to these two sets and stored them in a .fasta file using Biopython, an open source bioinformatics tool library for python. This allowed the date to be quickly parsed into a format which could be efficiently handled by the viennaRNA package.

I randomly selected a selected a sample of ~700 pairs of HP-Human mutation data and ~700 from HP-Random. ViennaRNA's RNAfold was used to compute the secondary structure and MFE values of both data sets (see Figure 1). Once MFE values were computed for all the 2800 secondary structures, I looked at differences between the MFE values within their respective sets (HP_MFE-Human_MFE and HP_MFE-Random_MFE). The means of the data and summed totals were calculated, and a t-statistic was used to determine if there were significant differences in the values.

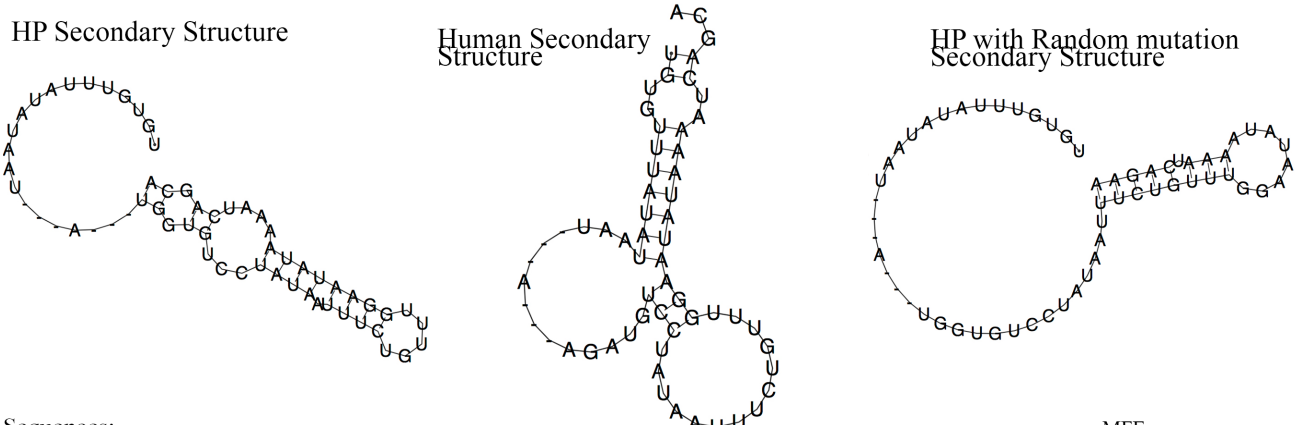

Figure 1: Illustrates an example of DNA secondary structure data from both data sets.

3. Results

After running the algorithm, as well as viennaRNA's RNAfold on the data, the results were examined. The hypothesis was that the Human data, because it was closely evolutionarily related to the Human-Primate (HP) ancestor data, would be more similar in δ-mean, and have a lower summed minimum free energy.

$$\delta - mean = \frac{\sum_N MFE_1 - MFE_2}{N}$$

While this was true, with a δ-mean of -3.08 and summed MFE of -207.03 for HP-Human and a δ-mean of 0.93 and summed mean of 22.42 for HP-Random, the difference was not significant. An independent samples t-test was computed on the data. The results showed that these deviations were not significantly different from random using $p<0.05$, with $t=0.093 < t\text{-critical}=1.962$.

$$t = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{S_1^2}{N} + \frac{S_2^2}{N}}}$$

I have graphed the data to illustrate the minimal differences in the δ-means of the two datasets.

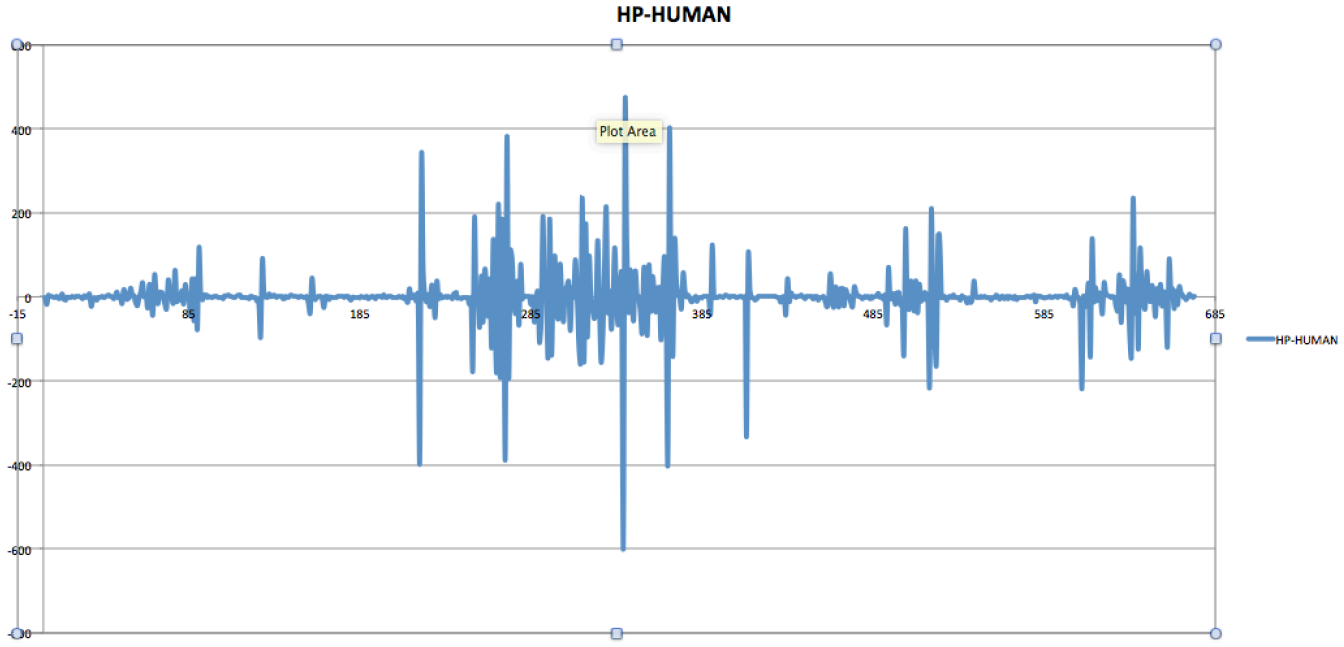

Figure 2: HP-HUMAN δ-mean values plotted along the X-axis.

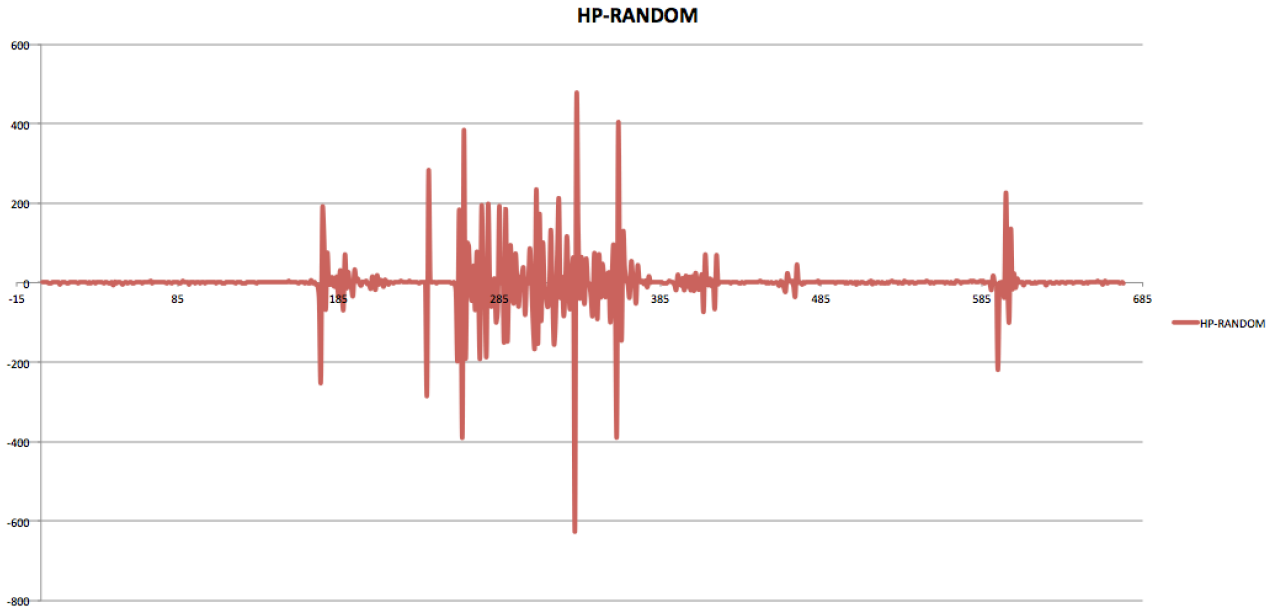

Figure 3: HP-RANDOM δ-mean values plotted along the X-axis.

Figures 2, 3 and 4 illustrate the minimal difference in values between the two datasets when comparing δ-means.

It is important to understand that although the data is not statistically significant, the results suggest that there may be a better way to conduct the analyses and construct the datasets in order to improve the significance of the results. Ideas about how to alter the methodology and analysis of the study will be discussed in the following section.

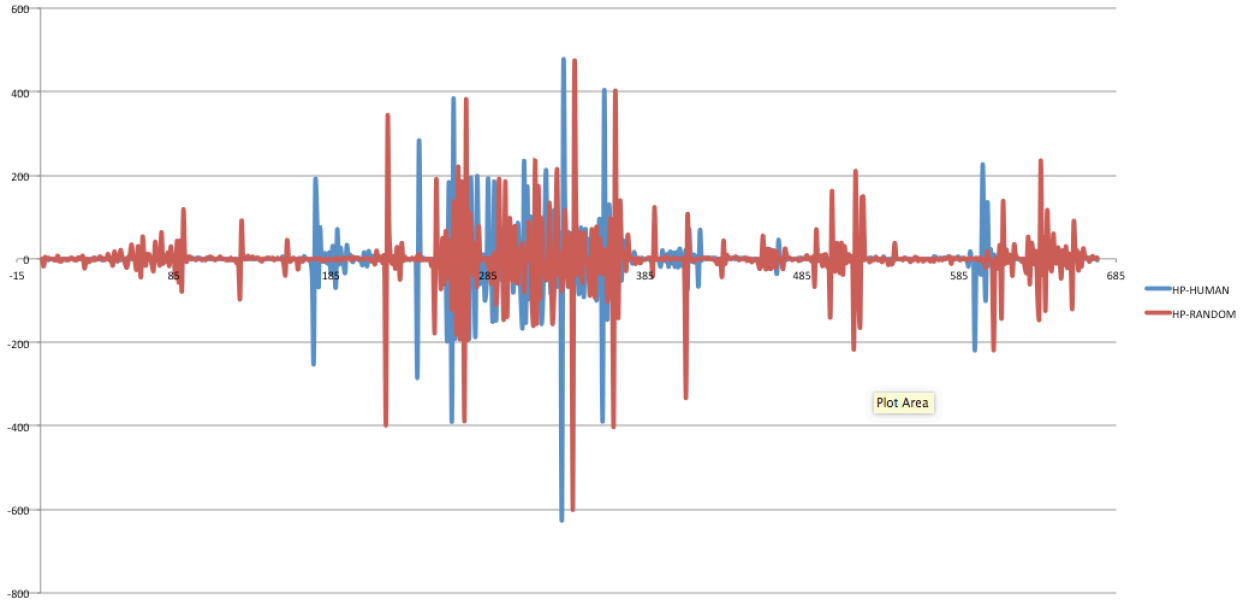

Figure 4: HP-HUMAN (Blue) and HP-RANDOM (Red) data overlapped.

4. Discussion and future work

While the results obtained in this study were not statistically significant, given that this area of investigation is in its infancy and there were possible sampling limitations, as outlined below, I believe that further investigation into the hypothesized relationship is warranted. Methods for improving the algorithm and the datasets will be discussed, as well as directions for future research.

1. **Improving the Algorithm**

The current version of the algorithm makes several assumptions about random mutations to the DNA. Firstly, it assumes only one nucleotide is mutated per comparison. This assumption was made because it is the strictest version of the algorithm that can be hypothesized. If the results were significant, which they were not, this would have meant significant results could have been generalized to a large number of mutations. The second assumption that was made about the data is that the single random mutation was equally likely anywhere along the gene segment, and was likely to mutate to any other nucleotide with equal probability ($p = ¼$). Again, this assumption was made to test the strictest case in an initial study. Changing the probabilities to more evolution-like styles should in turn narrow the effect of these mutations because of the interactions with neighbour nucleotides.

In a consequent study, either of these two corrections could be applied with relative ease. It would also make sense to compare HP-Human and HP-Random to some far ancestor of HP. The further we go up the phylogenetic tree, the closer to the HP-Random dataset and further from HP-Human our results should become. The more two genetic segments are separated from an evolutionary perspective, the less similar their sequences, and therefore structure, should be. Follow up studies should be done to examine these avenues and test their significance.

I also ran into issues sending large datasets into the viennaRNA package and had to reduce my data to a random sample of 683 code-segments. Improving this portion of the algorithm may be more representative of the population, and increase the validity of the results.

2. **Improvements on the Data**

There are multiple ways in which the data utilized in the study could be improved. Firstly, testing could be narrowed to consist of only a range of gene-segment lengths. This would help to reduce variance in the samples and generate clearer results. Some of the gene-segments used were as short as 2 nucleotides, where a mutation would completely change the sequence. If a range of gene-segments of 20-50 nucleotides was examined, an improved comparison of the algorithm could be generated.

Another limitation of the current data is the method used to compare the secondary structure. Although MFE values allow for easy comparison of the secondary structure, they do not always accurately depict the structure. Methods of simultaneous folding and alignment can be used to calculate the best singular structure between 3 sets of data: HP only, HP-Human, and HP-Random. We can then look at the change in secondary structure from HP to HP-Human and from HP to HP-Random, seeing if the latter cause a larger change (see Turner et. al., 2002). Multivariate data analysis would provide a better statistical test of the hypothesis.

Running Time

It may also prove beneficial to run these computations on more powerful hardware. To parse the current algorithms, manipulate and perform computations on the large datasets took upwards of 0.5 hrs per dataset on a 2.6GHz core i7. Parallelization of the computations may exponentially decrease running time, as well as access to better hardware. The viennaRNA package can process 100 x 100 nucleotides samples in 0.01 seconds. (Lorenz et al. 2011). Utilizing these methods would allow the testing of large data samples.

Future work

In the future, I would like to retest my algorithm after making the modifications stated above. I am confident that the hypothesis promises statistically significant results if the data and algorithm more accurately represent the question at hand. The improvements to both the sample-data and the algorithm are relatively minor and could be easily accomplished in a follow-up study.

Acknowledgements

I would like to acknowledge the teams behind viennaRNA, Biopython, Numpy, and Scipy for the use of their tools in this study. I would also like to thank Dr. Blanchette and Dr. Waldispühl for their guidance and datasets.

5. Bibliography

Alberts, Bruce, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. Molecular Biology of the Cell. New York: Garland Science, 2002. Print.

Anfinsen, C.B. 1973. Principles that govern the folding of protein chains. Science 181: 223–230.

Capra JA, Paeschke K, Singh M, Zakian VA. G-quadruplex DNA sequences are evolutionarily conserved and associated with distinct genomic features in *Saccharomyces cerevisiae*. PLoS Comput Biol. 2010;6:e1000861.

Chapman BA and Chang JT (2000). Biopython: Python tools for computational biology.

Cock PA, Antao T, Chang JT, Bradman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B and de Hoon MJL (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics.

Cole JR, Wang Q, Cardenas E, et al. The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. Nucleic Acids Research. 2009.

D. Sankoff, J.B. Kruskal, S. Mainville, R.J. Cedergren. (1983) Fast algorithms to determine RNA secondary structures containing multiple loops., chapter 3, pages 93-120.

Drake, J.W., Charlesworth, B, CHarlesworth, D. & Crow, J.F. Rates of Spontaneous Mutation. Genetics, April 1, 1998, Vol 148, NO. 4, 1667-1686.

Eyre-Walker, Adam (October 2006). "The genomic rate of adaptive evolution". Trends in Ecology & Evolution. Cambridge, MA: Cell Press

Frequently Asked Questions: Data File Formats. UCSC Genome Bioinformatics. U.S. Department of Health and Human Services.

Huppert JL, Balasubramanian S. Prevalence of quadruplexes in the human genome. Nucleic Acids Res. 2005;33:2908–2916.

Kent WJ, Hsu F, Karolchik D, Kuhn RM, Clawson H, Trumbower H, Haussler D. Exploring relationships and mining data with the UCSC Gene Sorter. Genome Res. 2005 May;15(5): 737-41.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. The human genome browser at UCSC. Genome Res. 2002 Jun;12(6): 996-1006.

Letunic, I and Bork, P. October 18, 2006. Interactive Tree Of Life (iTOL): An Online Tool for Phylogenetic Tree

Lorenz R, Bernhart SH, Höner zu Siederdissen C, et al. ViennaRNA Package 2.0. Algorithms for Molecular Biology.

Mathews DH, Turner DH. Dynalign: An algorithm for finding the secondary structure common to two RNA sequences. J Mol Biol. 2002;317:191–203.

Nachman MW, Crowell SL (September 2000). "Estimate of the mutation rate per nucleotide in humans". Genetics.

Polyatail. "Polyatail/biopython." GitHub. N.p., 23 Mar. 2012. Web.

Robinson, J.T, Thorvaldsdóttir, Helga, Winckler, Wendy, Guttman, Mitchell, Lander, Eric, Getz, Gad, Mesirov, J. Integrative Genomics Viewer. Nature Biotechnology 29, 24–26 (2011)

YE DING, CHI YU CHAN, and CHARLES E. LAWRENCE. RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble. RNA August 2005 11: 1157-1166; doi:10.1261/rna.2500605

UCSC Genome Browser Gateway. UCSC Genome Browser Gateway. N.p., n.d. Web.

Zuker, Jaeger, Turner. (1991) A comparison of optimal and suboptimal RNA secondary structures predicted by free energy minimization with structures determined by phylogenetic comparison. Nucleic Acids Res., 19, 2707-2714.

Zuker, M., Mathews, D.H., & Turner, D.H. Algorithms & Thermodynamics for RNA Secondary Structure prediction, A Practical Guide, pg 1-33.