HOW SELECTION FOR LANGUAGE COULD DISTORT THE DYNAMICS OF HUMAN EVOLUTION

W. GARRETT MITCHENER

Mathematics Department, College of Charleston
Charleston, SC, USA
mitchenerg@cofc.edu

The human language faculty is supported by a gene regulatory network that influences the development and operation of the motor, sensory, and cognitive systems. Consequently, this network must be fairly large and complex, and probably includes genes scattered throughout the genome. It is under selective pressure to continue functioning. Many gene loci are therefore in linkage disequilibrium with some element of this network, potentially disrupting their evolutionary dynamics. In the interest of exploring how significant this effect could be, we consider artificial life simulations, in which agents are required to perform an information coding task analogous to the replay of a memorized gesture. The task requires a network of interacting genes. The population is then branched, and phylogenetic trees are constructed based on genetic distances between leaf populations. Distances are determined by comparing genes for a simple task unrelated to the coding task. The process is then repeated without the coding task. Sample runs from these two simulations are compared to data from a neutral drift model. Both a-life simulations result in lower edge weights than the neutral model, giving the appearance of a slower molecular clock. Furthermore, the simulation with the coding task shows even lower edge weights, even though the mutation rate is the same. Therefore, the presence of a large gene network such as the one for language could distort the evolutionary trajectories of unrelated genes.

1. Introduction

The neutral theory of genetic drift is based on the assumption that most variations of a gene within a species are neutral mutants (Kimura, 1984; Ohta, 1992; Nowak, 2006). Beneficial mutations should spread and reach fixation in a process called a selective sweep. Deleterious mutations should die out quickly. Mathematically, each neutrally drifting site in a gene mutates according to a Poisson process. This model is the basis of the *molecular clock*, which asserts that the genetic distance between species should be related in a straightforward way to the time since they diverged. However, biological data is often inconsistent with this model. Similar genes in different species seem to have very different mutation rates, as do different genes within the same species, and sometimes even different sites within the same gene. Often there is no clear explanation for these discrepancies (Ayala, 1999; Cutler, 2000). Suggested explanations include the possibilities that the natural mutation rate varies in unexpected ways, that finite population size effects are surprisingly significant, and that weak selection may be complicating the dynamics. This third option is the subject of the computational experiments described here.

Selection acting on one set of genes can disturb the dynamics of others, even those that perform unrelated functions. Selective sweeps and hitchhiking (a form of linkage disequilibrium) are known to create such complications, but so could the tangled nature of gene regulatory networks. For example, selective breeding for friendliness during domestication of animals affects many other characteristics, apparently because it places pressure on the timing of the entire development process (Trut, 1999).

Once a selectively favored genetic network is established, purifying selection acts on it, meaning that any mutation causing it to malfunction will die out quickly, which systematically destroys a small amount of genetic diversity. If the network is sufficiently large and performs a highly favored function, there are many possible deleterious mutations. Consequently, purifying selection could reduce genetic diversity enough to give the illusion that the mutation rate is lower than it actually is, even in an unrelated gene, thereby distorting the molecular clock.

The human language faculty combines motor, sensory, and cognitive machinery, which suggests that it is generated by a large gene regulatory network. Supporting molecular evidence comes from the well known speech-related gene FOXP2, which regulates many other genes (Enard et al., 2002). Pressures on other language-related genes may therefore have affected evolutionary dynamics throughout the human genome.

There are statistical tests for selective pressure at the molecular level, usually expressed as whether the data is sufficient to reject a null hypothesis built from the neutral theory, but they do not indicate which aspect of the neutral theory fails to apply and are generally controversial (Nielsen, 2005). They are also not the appropriate tool here because the phenomenon of interest is not whether a specific gene has recently experienced selection, but generally how selection on a large network might affect the evolution of an unrelated gene as a side effect.

In the interest of developing new tools for investigating how selection acting on a large network affects the rest of the genome, consider the following experiment. Begin with a root population of a single species, and place it under selective pressure to evolve and maintain a large gene regulatory network that performs some information processing task analogous to language, as well as genes that perform simpler mundane tasks. Allow the population to evolve a solution to the mundane task, and at least a partial solution to the information processing task. Then make multiple independent copies of this population, and allow them to continue evolving under the same conditions. After some time, branch the populations again. Then construct phylogenetic trees from the resulting leaf populations (subspecies) and compare their statistics with the known process by which they were created. Such trees are commonly reconstructed by examining genes for mundane functions, such as dehydrogenases, assuming they fall under the neutral model.

Although this experiment is very difficult to conduct physically, it is straightforward to conduct computationally using an artificial life or *a-life* simulation. Essential details of the simulation are presented in Section 2. More details are available in (Mitchener, 2014). The source code and configuration will be available at the author's web site `mitchenerg.people.cofc.edu`. Section 3 describes the phylogenetic experiment and key results. These are discussed in Section 4.

2. Artificial life simulation

2.1. *Digital organisms*

Each organism in the a-life simulation, called an agent, has a genome, common to all of its cells, consisting of bit strings that serve as chromosomes. Substrings of 78 bits are interpreted as instructions analogous to biochemical reactions. Instructions are executed in parallel in discrete time steps. Each cell within an agent has an internal state consisting of counts $A[p]$ of how many molecules of each type p are present. Molecules and reactions are abstract, and no attempt is made to simulate molecular structure or chemical bonds. Instead, each type of molecule is an integer called a *pattern*, and each instruction states that if the number of units $A[s]$ of a particular pattern s exceeds some threshold θ, then some units of pattern p are created by adding to $A[p]$, and some units of q are destroyed by subtracting from $A[q]$. A bit of input is provided to a cell by adding to $A[j]$ for a certain designated pattern j on time steps when the input bit is 1, but not when it is 0. Likewise, for each designated output pattern r, a cell generates an output bit of 1 if $A[r]$ exceeds some threshold and 0 otherwise. A cell can have many inputs and outputs. A synapse can be created by connecting an output bit from one cell to an input bit of another cell, thus forming a cellular network. A population of these agents subjected to a selection-mutation process can evolve the ability to solve computational problems.

2.2. *Sequential coding and other tasks*

The *information processing task* for this experiment is sequential coding and decoding. Each agent consists of two cells. One is the sender, and is given an input word consisting of two bits. The sender can send synaptic spikes to the second cell in the agent, the receiver. Two of the receiver's outputs are interpreted as a two-bit word, and the goal is for these to recreate the original input word. The receiver must also set an output called the *stop signal* to indicate when the calculation is complete. The sending cell is given input all at once, but can only transmit information through a narrow synapse, so it must do so over time. This task is analogous to replaying a gesture, such as speaking a particular sound. During speech, entire words and phrases are present in the mind all at once, but must be replayed as muscle movements over time.

The *mundane task* is for the receiving cell to generate an additional output called a *beacon*. It indicates that the agent is basically alive, even if it isn't processing any information.

Each agent is given a rating based on how well it performs the above tasks. It earns a very large number of points by setting the beacon output. Thus, a mutation that damages the beacon is almost surely fatal. The agent is given the opportunity to transmit each possible input word, and earns many points for each bit in each input word that it correctly transmits. It earns extra points by stopping after fewer steps, which is the *timing task*. There is also a tiny penalty for using too many reactions, so overly complex mechanisms tend to get simplified.

2.3. *Evolutionary dynamics*

From each generation of 500 agents, 300 pairs of parents are chosen using tournament selection to produce 300 offspring. They are combined with the top-rated 200 agents of the previous generation to form a new generation of 500. Genomes are diploid, consisting of two pairs of chromosomes. When an agent reproduces, each pair of chromosome is aligned, a random crossover point is selected between genes, and a single recombined chromosome is produced, thus forming a gamete. Two gametes form a new agent. Chromosomes are subject to single-bit substitutions, whole-gene deletions, and whole-gene duplications.

The simulated population typically evolves a single-gene solution to the beacon task right away. Likewise, a single-gene timing mechanism usually develops early on, and enables agents to earn the extra points for stopping early. Solutions to the sequential coding problem evolve in steps. First, a mutation in some gene results in an instruction that is sensitive to one input bit and also activates the synapse in the sender. Another gene must be discovered that reads the synapse and links to the correct output bits. Once that basic connection is in place, gene duplications and other mutations form a link from the other input bit to the synapse in the sender, and from the synapse to the other output bit in the receiver. Typically, the receiver is able to determine which bits it should set from the time at which the sender begins spiking. The sender activates the synapse early if one bit is set, late if the other is set, and very early if both are set. This is usually enough information for the receiver to reconstruct everything except for one bit in one input word. To save time, sample runs used in this article were stopped at this point. (Reaching this stage took 13,000 to 100,000 generations over many hours of computer time. Given much more time, an additional mechanism evolves to handle that one last bit.) The result is a network of 10 to 20 genes that solves all but one part of the coding problem. Once a population reaches this stage, selective pressures on it are dominated by maintaining and possibly simplifying the encoding, decoding, timing, and beacon mechanisms.

There are good reasons to work with such a complicated simulation. The question at hand is whether a large, selectively favored gene network scattered throughout the genome (such as the one supporting the language faculty) could experience sufficient purifying selection to distort the evolutionary trajectories of unrelated genes. Any model that can address this question must account for how a subset of a genome specifies a network, how that network benefits the organism, and which mutations are beneficial, neutral, and deleterious. An a-life simulation of the complexity described here is a reasonable place to start.

3. Phylogenetic experiment

3.1. *Coding problem samples*

Begin with a population of 500 agents with random genomes. Let them evolve as described until the first agent capable of transmitting all but one bit dies, which means that a successful mechanism has saturated the population. The population is allowed to continue for 1000 more generations to ensure that the selective sweep has run its course. This is the *root population*. The genomes from this population are copied into four separate populations which are then run independently for 400 more generations each. Those populations are each branched into four and run for 400 more generations. The result is a population tree with four groups of four (4×4) *leaf populations*.

3.2. *Mundane problem samples*

A second set of leaf populations is created in the same way, except that no points are given for the coding task. Only the timing and beacon tasks affect these populations. Since these mechanisms evolve quickly, these populations are continued for 30,000 more generations before branching so that their history is of the same order of magnitude as that of the coding-problem runs. This extra time is necessary, otherwise some of the phylogenetic tree statistics described below come out artificially different between the coding-problem runs and the mundane-problem runs.

3.3. *Genetic distance between a-life populations*

The raw distance between two leaf populations from the a-life simulation is the average genetic distance between all maximally rated individuals in each population. Most agents in a leaf population achieve the same high rating, but a few have a lower rating due to deleterious mutations and are ignored. The genetic distance between two individuals is the minimum Hamming distance between beacon genes in each. The genetic distance is divided by 78, the length of a gene, to give a fractional distance p, which is converted to a final distance $-\ln(1-p)$. This final step is called the Poisson correction (Nei & Kumar, 2000).

3.4. *Neutral model samples*

As an additional control, a third set of population trees is constructed using a purely neutral process. Each of 1000 individuals consists of a string of 48 bits. To build the next generation, 600 individuals are selected and subject to single bit mutations. The 400 youngest individuals are kept for the next generation. An initial population of strings of all zeros is run for 30,000 generations to match the time required by the coding-problem sample runs, then branched into four copies and run for 400 generations. Each of those is branched into four copies and run for 400 generations, yielding the same population tree structure as the a-life simulation. The distance between pairs of these populations is the average Hamming distance between the bit strings therein divided by 78, then Poisson corrected.

Some explanation is in order for the configuration choices in the neutral simulation. Genes in the a-life simulation consist of 78 bits, 18 of which specify the pattern p such that $A[p]$ is increased, and 12 of which specify the instruction's threshold. That leaves up to 48 bits in the beacon gene that could potentially experience neutral drift. If the neutral simulation is run with strings of 78 unconstrained bits, the distances between populations are proportionally larger, which turns out to be undesirable. Also, the neutral simulation is run on a population of 1000 strings, which is one for each allele in the diploid a-life simulations.

3.5. *Phylogenetic reconstruction*

Each set of 4×4 leaf populations is analyzed as follows. Each leaf population is paired with each other leaf population that shares its parent. Each such pair is paired with each other pair with the same root. This gives a total of 216 quartet problems, that is, four leaf populations that must be assembled into an un-rooted tree with two internal nodes. The correct branching structure is known. Each branch is then given a non-negative weight so that the total weight along the path from one leaf to another is approximately their genetic distance. Specifically, weight assignments minimize the sum of squares of differences between tree path distance and genetic distance. This method of assigning weights is a standard technique in phylogenetics (Nei & Kumar, 2000).

According to the neutral theory, the distance between two leaf populations should be roughly proportional to twice the time since they branched. The actual time between branches is always 400 generations, and the mutation rate is fixed across all runs. Thus, all phylogenetic trees from all three simulations should ideally have weight w on the edges to the leaves, and $2w$ on the central edge, and all three simulations should result in similar values of w. The central edge should be weighted twice as much because the tree does not include a node for the root population. We therefore consider the following statistics. Given a quartet tree with weights as shown in Figure 1, the *discrepancy* from an ideal tree with branches weighted w is

$$R(w)^2 = (w_{\mathrm{AX}} - w)^2 + (w_{\mathrm{BX}} - w)^2 + (w_{\mathrm{CY}} - w)^2 + (w_{\mathrm{DY}} - w)^2 + (w_{\mathrm{XY}} - 2w)^2$$

The optimal w and corresponding R are

$$\begin{aligned} w^* &= \operatorname*{argmin}_w R(w) = \frac{1}{8}\left(w_{\mathrm{AX}} + w_{\mathrm{BX}} + 2w_{\mathrm{XY}} + w_{\mathrm{CY}} + w_{\mathrm{DY}}\right) \\ R^* &= R(w^*) \end{aligned}$$

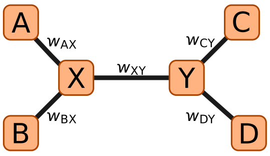

Figure 1. General un-rooted phylogenetic quartet tree for four leaf populations (A, B, C, D) with two ancestral populations (X, Y).

Twenty samples of each of the three models (neutral, mundane, coding) were run, all 216 weighted trees were constructed, and w^* and R^* were calculated (Figure 2). The charts show quartiles for the numbers gathered from each sample run, but it should be noted that numbers from the same run (within each bar) are not entirely independent because they are calculated from phylogenetic trees that can share leaves. Statistics calculated from different runs are independent, however. Weighted trees constructed from biological and simulation data are often not in ideal proportions (Nei & Kumar, 2000, p. 79), so the range of discrepancy values seen in Figure 2 is expected.

It is clear that w^* and R^* are distinctly larger for the neutral sample runs than the a-life sample runs. There is also a distinction in w^* between the a-life models. The Mann-Whitney-Wilcoxon rank sum test applied to sets of median w^* values from each of the sample runs from the mundane problem and coding problem simulations yields a p-value of 0.0154. That is, the difference in the median value of w^* derived from the mundane problem of 0.0245 and the value derived from the coding problem of 0.0201 is statistically significant. Consequently, the weak selectional forces acting directly and indirectly on the mundane task gene in the mundane-only and coding task simulations are not negligible when assigning weights to phylogenetic trees.

4. Discussion and conclusion

When reconstructing weighted phylogenetic trees of the great apes, the human-specific data has several odd features. When mitochondrial DNA is used to estimate genetic distances, a frequently seen anomaly is that the path from root to

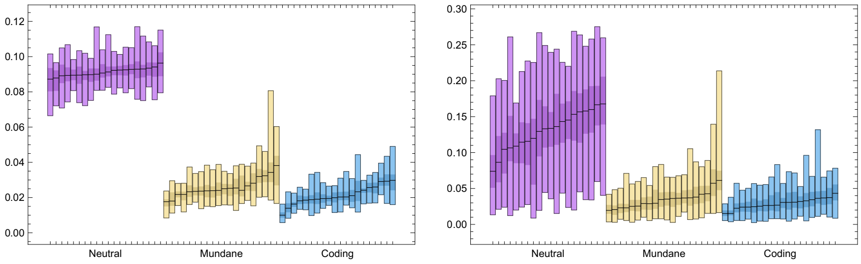

Figure 2. Charts for all sample runs under each model. Each bar ranges over all values from all weighted trees from one sample run. Stripes indicate quartiles. Black lines indicate medians. Runs from each model are ordered by median for easier comparison, and are not in the same order in each chart. Left: w^*; Right: R^*.

human is slightly shorter (smaller total weight) than paths to the other species, though it is not necessarily statistically significant (Nei & Kumar, 2000; Glazko & Nei, 2003). A related puzzle is that compared to the other great ape species, humans show a distinctly low level of genetic diversity (Deinard & Kidd, 1999). Many forces potentially underlie these anomalies: random noise, differences between mitochondrial and nuclear DNA, natural polymorphism, the fact that speciation is not instantaneous, founder effects due a bottleneck in the history of the human population (Deinard & Kidd, 1999), and a gradual increase in the time between generations (Scally et al., 2012).

The simulation presented here shows that even if the mutation rate, population size, and branch points are fixed and idealized, the phylogenetic trees can be distorted if there is a relatively large and important gene network. Using the neutral model as a baseline, the low edge weights on the mundane-problem and coding-problem trees would underestimate the mutation rate or the time between leaf populations and the first branches. (There is also evidence that phylogenetic methods can significantly overestimate mutation rates (Scally & Durbin, 2012).) Coding-problem trees had even lower edge weights overall than mundane-problem trees. Thus, the selective pressure maintaining a large gene network suffices to distort the molecular clock of unrelated genes.

The relatively low genetic diversity of humans may be partly due to this force. Mitochondrial DNA would be likewise affected. The initial formation of a large, selectively favored gene network, followed by a selective sweep, would contribute to the hypothetical bottleneck and associated founder effects. The language faculty is an obvious candidate for the phenotype of such a network. It is not yet clear how to determine the extent to which this large-network effect might be responsible for anomalies in human genetics, compared to disease, climate, and other non-genetic forces. However, the simulations described here imply that selective pressure on the language faculty, even just to maintain its function, could have significantly altered the evolutionary dynamics of many other human genes.

References

Ayala, F. J. (1999). Molecular clock mirages. *BioEssays*, *21*(1), 71–75.

Cutler, D. J. (2000). The Index of dispersion of molecular evolution: slow fluctuations. *Theoretical Population Biology*, *57*(2), 177–186.

Deinard, A., & Kidd, K. (1999). Evolution of a HOXB6 intergenic region within the great apes and humans. *Journal of Human Evolution*, *36*(6), 687–703.

Enard, W., Przeworski, M., Fisher, S. E., Lai, C. S. L., Wiebe, V., Kitano, T., Monaco, A. P., & Pääbo, S. (2002). Molecular evolution of FOXP2, a gene involved in speech and language. *Nature*, *418*(6900), 869–872.

Glazko, G. V., & Nei, M. (2003). Estimation of Divergence Times for Major Lineages of Primate Species. *Molecular Biology and Evolution*, *20*(3), 424–434.

Kimura, M. (1984). *The Neutral Theory of Molecular Evolution.* Cambridge University Press.

Mitchener, W. G. (2014). Evolution of communication protocols using an artificial regulatory network. *Artificial Life*, *20*(4), 491–530.

Nei, M., & Kumar, S. (2000). *Molecular evolution and phylogenetics.* Oxford: Oxford University Press.

Nielsen, R. (2005). Molecular signatures of natural selection. *Annual Review Of Genetics*, *39*, 197–218.

Nowak, M. A. (2006). *Evolutionary Dynamics: Exploring the equations of life.* Cambridge, MA: Harvard University Press.

Ohta, T. (1992). The Nearly neutral theory of molecular evolution. *Annual Review of Ecology and Systematics*, *23*, 263–286.

Scally, A., & Durbin, R. (2012). Revising the human mutation rate: implications for understanding human evolution. *Nature Reviews Genetics*, *13*(10), 745–753.

Scally, A., Dutheil, J. Y., Hillier, L. W., Jordan, G. E., Goodhead, I., Herrero, J., Hobolth, A., Lappalainen, T., Mailund, T., Marques-Bonet, T., McCarthy, S., Montgomery, S. H., Schwalie, P. C., Tang, Y. A., Ward, M. C., Xue, Y., Yngvadottir, B., Alkan, C., Andersen, L. N., Ayub, Q., Ball, E. V., Beal, K., Bradley, B. J., Chen, Y., Clee, C. M., Fitzgerald, S., Graves, T. A., Gu, Y., Heath, P., Heger, A., Karakoc, E., Kolb-Kokocinski, A., Laird, G. K., Lunter, G., Meader, S., Mort, M., Mullikin, J. C., Munch, K., O'Connor, T. D., Phillips, A. D., Prado-Martinez, J., Rogers, A. S., Sajjadian, S., Schmidt, D., Shaw, K., Simpson, J. T., Stenson, P. D., Turner, D. J., Vigilant, L., Vilella, A. J., Whitener, W., Zhu, B., Cooper, D. N., Jong, P. de, Dermitzakis, E. T., Eichler, E. E., Flicek, P., Goldman, N., Mundy, N. I., Ning, Z., Odom, D. T., Ponting, C. P., Quail, M. A., Ryder, O. A., Searle, S. M., Warren, W. C., Wilson, R. K., Schierup, M. H., Rogers, J., Tyler-Smith, C., & Durbin, R. (2012). Insights into hominid evolution from the gorilla genome sequence. *Nature*, *483*(7388), 169–175.

Trut, L. (1999). Early Canid Domestication: The Farm-Fox Experiment. *American Scientist*, *87*(2), 160.