Correlating Genetic and Geographic Distance in Sample Genomes

Shawn O'Neil and Nate Garrison – Bioinformatics Project

December 14, 2008

1 Introduction

As the costs of genome sequencing and assembly continue to decline, new opportunities and methodologies for research are becoming available. On such opportunity is the ability to study large-scale genetic variation across many individuals in a given population. Studies such as this have been loosely categorized under the label "population genomics." An important goal of population genomics is to identify outlier loci which exhibit "locus specific effects" in contrast to "genome wide effects [1]." For example, after sampling many individuals in a population, we would expect to see genome-wide genetic drift between individuals. However, some interesting subset of the population may be exhibiting strong selection at a few loci—separating this locus-specific "wheat" from the genome "chaff" can help direct the course of interesting study.

Luikart et al. summarize population genomics as follows:

> The two main principles of population genomics are that neutral loci across the genome will be similarly affected by demography and the evolutionary history of populations, and that loci under selection will often behave differently and therefore reveal 'outlier' patterns of variation[6].

Identifying such outliers can be useful for two reasons. First, by identifying outlier loci, we can also identify neutral loci—those which are not experiencing locus-specific effects in a population and hence are more useful for comparative studies such as phylogenetic tree construction. Second, and more obviously, outlier loci are likely to be interesting regions of the genome, with previously unknown outlier loci signaling areas of the genome which warrant further investigation.

Example uses of outlier loci are clustering a population into subpopulations [4] and testing for correlations between genetic distance and geographic distance between individuals [6].

It is this last example, correlating genetic and geographic (or some other measure) distance, which we seek to explore in this project. Specifically, we develop an exploratory data analysis tool to help find regions of a genome which are changing in correlation with a distance metric. Such areas might be interesting, for example, if the reason for this correlation is positive selection which gets stronger as the geographic distance gets larger.

2 Approach

While we presume that the data gathered represents **i)** a finished reference genome and **ii)** finished sample isolate genomes annotated with the distance from the reference, in reality the data could represent similar draft genomes or even comparable collections of expressed sequence tags.

Suppose we align each sample isolate sequence to the reference sequence. For any such alignment we can consider a window of size K starting at position i in the reference, and create a two-dimensional data point for this window with the first dimension specifying the annotated distance and the second dimension specifying the genetic distance. With n alignments/sample isolates, we would get n such data points. Figure 1 illustrates this logical construction.

Given this data, how "interesting" is this region of

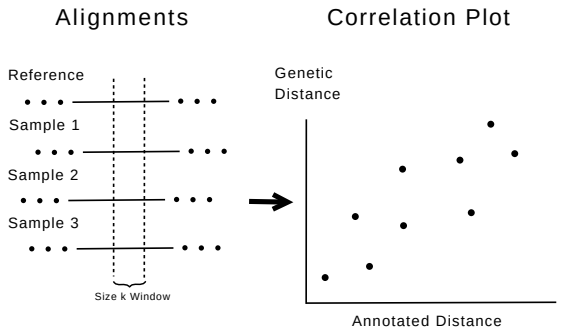

Figure 1: Hypothetical plot of genetic distance vs. annotated distance, for a given window of size K. If the slope of a line fit to the plot on the right and the correlation coefficient are both large, this area of the genome is changing relevant to the annotated sample distance.

the genome in terms of annotated and genetic distance correlation? Consider a line fitted to the plot in Figure 1. If the slope of the line is small, genetic distance isn't strongly related to annotated distance, even if the correlation coefficient of the data is large. If the slope is relatively large, genetic distance appears to be related to annotation distance, but this relationship is only strong if the correlation coefficient is also large.

Thus, for a particular window, we use $(slope\ of\ fit\ line) * (correlation\ coefficient)$ as a combined measure of "interestingness." By plotting this $slope * correlation$ statistic for each window at $i = 1, 2, \ldots, l - k + 1$ (where l is the length of the reference genome), we can quickly and visually find areas of interest in the genome.

3 Methods

As a first pre-processing step, each sample genome is individually aligned to the reference genome using the `nucmer` tool from the Mummer package [5] with the default parameters. A single global alignment is produced using the `delta-filter` tool with the `-r` and `-q` options. When reconstructing the alignments for the analysis, we disregard sections of the alignments where there are gaps in the reference genome. This ensures that all n alignments are exactly l bases long, where l is the length of the reference genome.

After a window size K is decided on, the linear regression line slope and correlation coefficient are computed for each window beginning at $i = 1$ to $l - k + 1$ as follows: For each alignment, a point $< annotateddistance, geneticdistance >$ is included in the calculations *if* **i)** the number of gaps in the sample genome is less than half the window size (since we treat gaps as missing data, and don't want to include erroneous points) and **ii)** the genetic distance is defined. Our tool currently implements Jukes-Cantor as well as Kimura two-parameter genetic distance measures, both of which are undefined if the genetic distance is greater than "random." Finally, while the correlation coefficient is technically undefined if all sample points have the same genetic distance, in this case we define the correlation to be 1 and the slope to be 0.

The data is displayed for analysis using the program `kst` [7]. Figure 2 shows an example. The top panel plots correlation, slope, and the $correlation * slope$ statistics for each window indexed by window start position in the reference genome. (Actually, $correlation * |slope|$ is displayed, to preserve the sign of the statistic.) For ease of viewing, all three plots are scaled so that the maximum of their absolute value is 1.0.

The bottom panel of the plot shows the number of sample points used in computing the statistics. (Recall that large gaps in a sample alignment can lead to that alignment's window not contributing.) The bottom panel also shows where in the genome the top 5 percentile $slope * correlation$ statistics appear.

3.1 Runtime

Because we align each sample genome to a reference we avoid needing to do a multiple alignment between what could be a large number of sample genomes. Further, because we use streaming algorithms to compute the linear regression line slope and correlation in constant space (for each window), the method uses only $O(n + l)$ space. Linear space algorithms are important in the field of population ge-

Name of Isolate	Acc. Number	Days
GZ01	AY278489	0
ZA-A	AY394997	10
ZS-C	AY395004	19
GZ-B	AY394978	39
HZS-2A	AY394983	46
GZ-50	AY304495	64
CUHK-WI	AY278554	67
Urbani	AY278741	72
Tor2	AY274119	73
Sin2500	AY283794	75
TWI	AY291451	82
CUHK-AG01	AY345986	93
CUHK-L	AY394999	150

Table 1: SARS sample isolate names, GenBank accession numbers, and number of days past December 16, 2002 each isolate was sampled. For this dataset, we use the number of days as the annotated distance, allowing us to find areas of the genome where temporal distance is highly correlated with genetic distance.

nomics, as the number of samples n is often large [6]. The method runs in $O(nl)$ time for both the Jukes-Cantor and the Kimura distance measures. (Neither of the time and space bounds include the preprocessing step of aligning the genomes.)

4 Example: SARS Dataset

SARS is an example of a coronavirus which was first seen in humans in the second half of 2002. Because of its lethality, the entire genome for the virus was sequenced from isolates varying in time during the outbreak (December 2002 to May 2003). Cristianini and Hahn provide a list of 13 of these, annotated with sample location and collection date. We reproduce this list of isolates, including name, GenBank accession number, and time of sample collection in days from December 16, 2002 in Table 1.

This dataset provides an interesting test case for our tools. Note that in this case, the annotated distance isn't geographic, but temporal. As such, loci where the fit line has a high slope would be mutating faster than other regions. Where the correlation is high, we would expect simple genetic drift to be the only factor present in mutation. Thus, we posit that areas with a high $correlation{*}slope$ statistic would be good candidates for areas which model a "molecular clock," and hence would be good candidate regions for use in phylogenetic analysis.

Figure 2 shows the output of our tool on the SARS dataset, using the Kimura genetic distance measure and a 1,000 base pair window size.

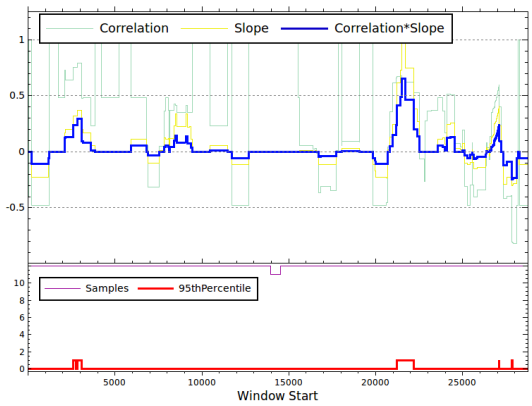

Figure 2: Temporal distance vs. genetic distance $correlation * slope$ plot for the SARS dataset. The largest peak in this plot occurs at position 21,507, which coincides with the start of the spike protein.

The largest peak in Figure 2 occurs at the 1,000 base pair window which starts at position 21,507 in the reference genome. This position corresponds to the start of the spike protein, which runs from positions 21,476 to 25,243 in the reference genome [8]. It appears to be no coincidence, then, that the spike protein has been used to reconstruct the phylogenetic history of the SARS virus, since it is already known to be prone to high mutation rates [2, 9]. Our analysis suggests that the first 1,000 to 1,200 bases of the spike protein would be best for this use.

A second interesting feature of Figure 2 is the sign of the slope of the fit line. Since in general we expect genetic distance to increase with time, we would also expect the plots in Figure 2 to be non-negative. However, this is not always the case. These negative slope

regions may be a statistical coincidence or artifact of our methodologies; thus far we haven't investigated these features.

5 Simulation

We simulated a diverse dataset in order to test the capability of our method. In order to see how our method would detect various types of mutation, we created samples by varying open reading frames from a reference *Mycoplasma Genitalium* genome according to both a variety of region "types" as well as annotated distance [3]. For example, we modified some ORFs as though they were highly conserved, and others as though the amount of mutation was correlated to the annotated distance.

For each sample, the simulation did two major things to the *Mycoplasma Genitalium* genome. It first labeled every open reading frame in terms of the possibilities of how genes could be altered due to an annotated distance. Three factors were considered in determining the likeliness of a single base change in the region. First, an initial base rate of change (low, 0.03%; medium, 0.1%; high, 0.5%) determines the basic amount of "conservation." Second, a slope factor in terms of additional percent chance of change per mile (low, 0.0%; medium; 0.05%, high, 0.1%) determines how much of a factor the annotated distance is in the amount of mutation. Finally, a random amount of percentage chance is added with varying standard deviation σ (low, 0.05%; medium, 0.1%; high, 0.5%) which affects the strength of the correlation. The formula for determining the mutation rate for an ORF i on simulated sample j is given by

$$\begin{aligned}\text{Mutation Rate}_{i,j} = &\ \text{Base Rate}_i \\ &+ \text{Distance}_j * \text{Slope}_i \\ &+ N(0, \sigma_i)\ ,\end{aligned}$$

where $N(\mu, \sigma)$ is the normal distribution with mean μ and standard deviation σ. The above is resampled if the normal variable caused the mutation rate to be negative.

Introns were modified in a similar manner; in this case the base rate of change was high (indicating little conservation) and the slope and sigma were both low (for no relation to the annotation distance). Further, we defined "interesting" regions of the genome (those with a large slope and correlation) to be relatively rare. Twenty sample genomes were simulated, with random annotation distances in the range 1–100. Table 2 gives the ORF types we simulated and the percentages of each in the genome.

Like most things in biology, some of the simulated regions are more interesting than others. In particular, ORF IDs 2 and 3 are quite interesting, as they simulate a change where there is strong correlation and slope between genetic distance and annotation distance. ORF IDs 5 and 6 could be moderately interesting as well. However, the rest of the IDs as well as introns are by design not meaningful. In practice we expect interesting types such as 2 and 3 to be rare, therefore we have simulated them as such.

5.1 Results

The simulation produced expected results. The full genome plot is shown in Figure 3. Figure 4 shows a small portion. The peaks near bases 193,225 through 193,659 and 240,032 through 242,224 in Figure 4 correspond to "interesting" type 3 ORFs. The peak around 198,000 corresponds to a type 2 ORF which occurs from bases 198,517 to 199,152 in the reference genome. The fourth peak corresponds to a "medium interest" type 6 ORF from bases 233,038 to 235,458. Overall it seems the method does a good job separating noise from interesting regions. However, due to the small size of the type 2 ORF, the type 6 ORF looks equally as important in the plot.

6 Conclusion and Future Work

In this paper we presented an exploratory data analysis tool, designed to analyze sample genomes (or other DNA/amino acid sequences) with an eye toward finding regions of the genome which are changing in strong correlation with an annotated distance. Through simulation of genome mutation as well as studying a temporally labeled SARS dataset, we believe that such a tool can be effectively used to find previously unknown interesting locus specific effects.

ID	Description	Percent of ORFs	Base	Slope	Sigma
0	Intron	N/A	High	Low	Low
1	Life Gene	25%	Low	Low	Low
2	Interesting	2.5%	Medium	High	Low
3	Interesting	2.5%	Low	High	Low
4	Noise	30%	Medium	Low	High
5	Noise	30%	Low	Low	High
6	Maybe Interesting	5%	Medium	Medium	Medium
7	Maybe Interesting	5%	Low	Medium	Medium

Table 2: ORF types simulated and the relative percentages of each in the genome. The majority of the ORFs (as well as intron regions) are of no interest in correlating genetic and annotated distance. Those with higher slope and lower Sigma are considered to be of higher interest.

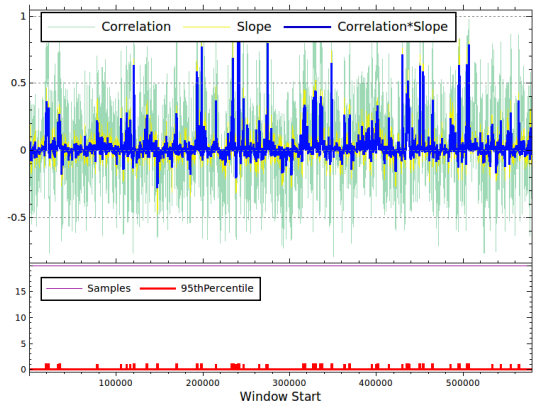

Figure 3: Full genome plot for the simulated results. The tool `kst` allows for easy exploration of the data.

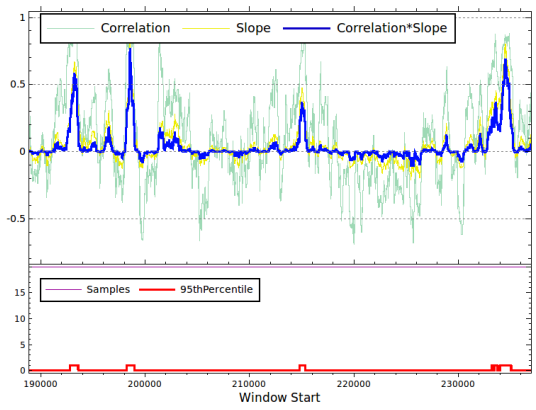

Figure 4: Detail of a portion of the simulation results. The four tallest peaks in the plot correspond to ORF types 3, 3, 2, and 6, respectively.

Here we note that the method is not restricted to Jukes-Cantor or Kimura genetic distance. If labeled open reading frames are available for each sample in the dataset, a measure of selective pressure such as the K_a/K_s ratio can be used as well. Further, for any given window the statistic of $correlation * slope$ is but one of many which could be measured. Another possibility is that of Information Gain, a data mining concept which utilizes entropy to determine if a dataset can be split into discrete groups with high intra-group similarity. While regions of the genome may be changing in accordance with an annotated distance, they may not be doing so in a linear, continuous fashion. In cases such as this, $correlation*slope$ may be less useful than statistics such as Information Gain. Unfortunately, when using other statistics, it may not be possible to retain the $O(n+l)$ space bound and the $O(nl)$ time bound.

Finally, while we have assumed that the samples will be sequences, it shouldn't be much trouble to modify the technique if all that is given is a list of differences between the sample and the reference. (Certainly, at the very worst the sample genomes can be reconstructed from the difference information given in linear time.) We wouldn't expect the runtime

or the required space to improve, however, for full genome plots as we have been discussing.

7 Contributions

Shawn O'Neil focused on tool development and integration with the `kst` scientific plotting tool, while Nate Garrison handled all things simulation. Integral ideas were contributed by both parties.

References

[1] W.C. Black, C.F. Baer, M.F. Antolin, and N.M. DuTeau. Population genomics: Genome-wide sampling of insect populations. *Annual Review of Entomology*, 46(1):441–469, January 2001.

[2] Nello Cristianini and Matthew W. Hahn. *Introduction to Computational Genomics: A Case Studies Approach*. Cambridge University Press, New York, NY, USA, 2007.

[3] C.M. Fraser, J.D. Gocayne, O. White, M.D. Adams, R.A. Clayton, R.D. Fleischmann, C.J. Bult, A.R. Kerlavage, G.G. Sutton, J.M. Kelley, J.L. Fritchman, J.F. Weidman, K.V. Small, M. Sandusky, J.L. Fuhrmann, D.T. Nguyen, T. Utterback, D.M. Saudek, C.A. Phillips, J.M. Merrick, J. Tomb, B.A. Dougherty, K.F. Bott, P.C. Hu, T.S. Lucier, S.N. Peterson, H.O. Smith, and J.C. Venter. Mycoplasma genitalium g37, complete genome. GenBank Accession NC000908, 2008.

[4] O.J. Hardy and X. Vekemans. SPAGEDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. *MOLECULAR ECOLOGY NOTES*, 2(4):618–620, DEC 2002.

[5] S. Kurtz, A. Phillippy, A. L. Delcher, M. Smoot, M. Shumway, C. Antonescu, and S. L. Salzberg. Versatile and open software for comparing large genomes. *Genome Biol*, 5(2), 2004.

[6] Gordon Luikart, Phillip R. England, David Tallmon, Steve Jordan, and Pierre Taberlet. The power and promise of population genomics: From genotyping to genome typing. *Nat Rev Genet*, 4(12):981–994, December 2003.

[7] Barth Netterfield, Staikos Computing Services Inc., Sumus Technology Limited, Rick Chern, Nicolas Brisset, Matthew Truch, Theodore Kisner, Duncan Hanson, and Yiwen Mao. Kst - plots scientific data. `http://kst.kde.org`.

[8] E. Qin, Q. Zhu, M. Yu, B. Fan, G. Chang, B. Si, B. Yang, W. Peng, T. Jiang, B. Liu, Y. Deng, H. Liu, Y. Zhang, C. Wang, Y. Li, Y. Gan, X. Li, F. Lu, G. Tan, R. Yang, W. Cao, J. Wang, W. Chen, L. Cong, Y. Deng, W. Dong, Y. Han, W. Hu, M. Lei, C. Li, G. Li, G. Li, H. Li, S. Li, S. Li, W. Li, W. Li, W. Lin, J. Liu, Z. Liu, H. Lu, P. Ni, Q. Qi, Y. Sun, L. Tang, Z. Tong, J. Wang, X. Wang, Q. Wu, Y. Xi, Z. Xu, L. Yang, C. Ye, J. Ye, B. Zhang, F. Zhang, J. Zhang, X. Zhang, J. Zhou, and H. Yang. Sars coronavirus gd01 isolate genome sequence. GenBank Accession AY278489, 2003.

[9] Y. J. Ruan, C. L. Wei, A. L. Ee, V. B. Vega, H. Thoreau, S. T. Su, J. M. Chia, P. Ng, K. P. Chiu, L. Lim, T. Zhang, C. K. Peng, E. O. Lin, N. M. Lee, S. L. Yee, L. F. Ng, R. E. Chee, L. W. Stanton, P. M. Long, and E. T. Liu. Comparative full-length genome sequence analysis of 14 sars coronavirus isolates and common mutations associated with putative origins of infection. *Lancet*, 361(9371):1779–1785, May 2003.