Journal of Humanistic Mathematics

Volume 9 | Issue 2 July 2019

Using Hidden Markov Modeling for Biogeographical Ancestry Analysis

Melvin R. Currie

Recommended Citation

Melvin R. Currie, "Using Hidden Markov Modeling for Biogeographical Ancestry Analysis," *Journal of Humanistic Mathematics*, Volume 9 Issue 2 (July 2019), pages 60-77. DOI: 10.5642/jhummath.201902.06. Available at: https://scholarship.claremont.edu/jhm/vol9/iss2/6

Using Hidden Markov Modeling for Biogeographical Ancestry Analysis

Melvin R. Currie

Baltimore, Maryland, USA
currie.incoming@gmail.com

Abstract

This paper describes a methodology for analyzing X chromosome data to establish biogeographical contributions to the author's X chromosome. We present an exposition of how Hidden Markov Modeling (HMM) can be used as a black box for ancestry analysis and focus on a set of conditions that are not universal but fairly common. The first condition is that the ancestral populations are drawn from regions that have had very little or no contact with each other since prehistoric times. The second condition is that the number of possible ancestral populations is small. In this analysis, we assume that the ancestral populations are Native North American, Northwestern European, and West African. We compare the result of our analysis with the analyses carried out by the companies 23andMe and deCODEme for the same data. Finally, we point to a mechanism for reducing noise by adjusting the data before applying HMM.

This paper describing the author's analysis of his X chromosome is the result of a marriage between two spheres. The author is a mathematician and an avid genealogist. His formal education is in pure mathematics, having written a PhD dissertation in that domain, which was followed by a period in academia conducting related research. However, he spent the last 25 years of his career before retirement applying mathematics to cryptanalysis and cryptographic design at the National Security Agency. The year before his retirement he wrote an in-depth paper on Hidden Markov modeling (HMM) that covered in gory detail, with all the derivations and proofs, everything from the alpha-pass to the Baum-Welch convergence.[1]

[1]This was an internal NSA paper but is available upon request from the author.

The current article is a byproduct of that paper and introduces a more sophisticated approach to handling the population data when applying HMM to ancestry analysis.

Initially, the author's pursuit of family history employed conventional genealogical tools, and he was able to trace the presence of his ancestors in North America to the 1700s in the colonies (eventually states) of Virginia and North Carolina, specifically the Piedmont regions of those states. When direct-to-the-consumer DNA analysis became available, he had been doing conventional genealogy for two decades. For the past ten years, DNA-related tools have been a welcome addition to oral history and document sleuthing.

Let's set the stage. Among the products that these companies provide is a biogeographical breakdown of contributions to the customer's genome from various regions around the globe. This paper will apply HMM to that problem, specifically the author's X chromosome, the analysis of which showed, among other things, a large contribution from a Native American ancestor or ancestors. Further, the X chromosome for men does not require any effort to tease apart the father's contribution from the mother's contribution. That is to say that no phasing is required, which removes a difficult challenge.

The X chromosome differs from the autosomal chromosomes (chromosomes 1 through 22) in that a rather restricted set of ancestors are potential contributors. Males receive an X chromosome only from their mothers. This is in contrast to females, who receive an X chromosome from each parent. What is just as interesting is that the male passes his X chromosome to his offspring effectively without recombination being a part of the process. As a result, distant ancestors on the X chromosome "glide path" have a greater chance of keeping in play a contribution to a descendant on the X chromosome than they have on the autosomal chromosomes. Figure 1 displays the X chromosome contributors for four ancestral generations. Note that although we have sixteen 2nd great grandparents, only five of them could possibly be contributors to our X chromosome. In the author's case, Native American ancestry is disproportionately represented on his X chromosome by an order of magnitude, with well over 40% of the chromosome reflecting Native American contribution versus just a bit more than 3% overall contribution of Native American ancestry to his genome. We will eventually reach the conclusion that there is no European contribution to the author's X chromosome, despite the author having a far larger European contribution across the genome than Native American contribution.

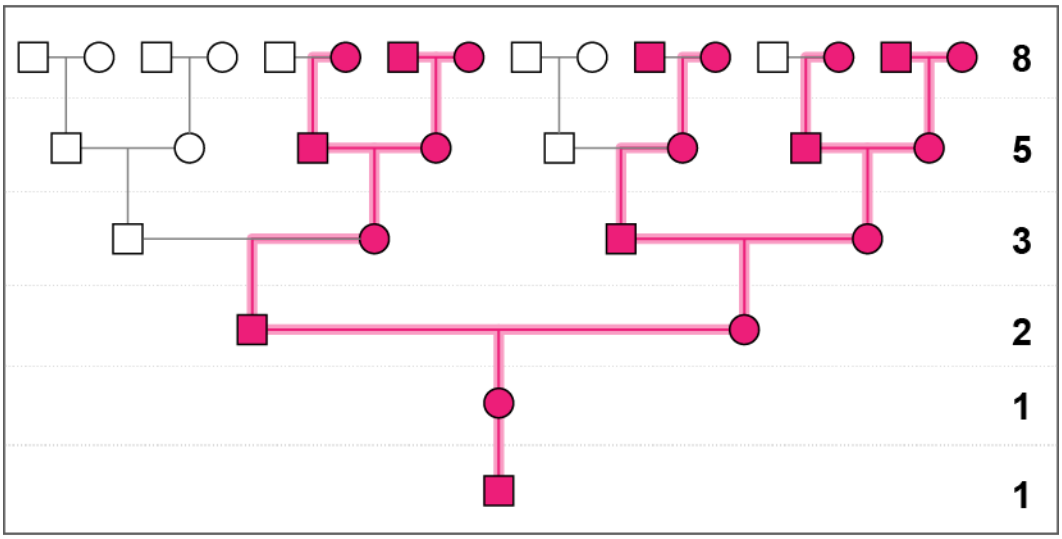

Figure 1: Six generations of the X chromosome inheritance tree (in red) for a male. The letters in black represent the maternal ancestors that cannot contribute. No paternal ancestors contribute to a male's X chromosome and for sake of simplicity the paternal branch is not shown. (Circles-female, Squares-male).

By the way, many readers will recognize that we have the first few terms of the Fibonacci sequence in the column on the right: 1, 1, 2, 3, 5, 8. The number of "X chromosome" 3rd great grandparents is $8 = 5 + 3$. There are 13 (8+5) 4th great grandparents in this category.

Figure 2 shows two analyses of the author's X chromosome, one at the testing company 23andMe (Mountain View, California, USA) and one using the same raw data file processed by the deCODEme service (deCODE genetics, Reykjavik, Iceland; the deCODEme feature was discontinued in 2013). The centromere is a region of the chromosome for which data were unavailable.

Of course, one notices immediately that the results are not the same. There is a significant segment of European ancestry found by deCODEme, while Europe was a no-show in the 23andMe results. We should note that in the graphic representing the deCODEme analysis, the green segments were actually assigned to an umbrella category that the company simply called "Asian." This category included the indigenous North American population. In an earlier incarnation of the 23andMe analysis, half of what the graphic shows as Native American was given the designation "Mongolian.'

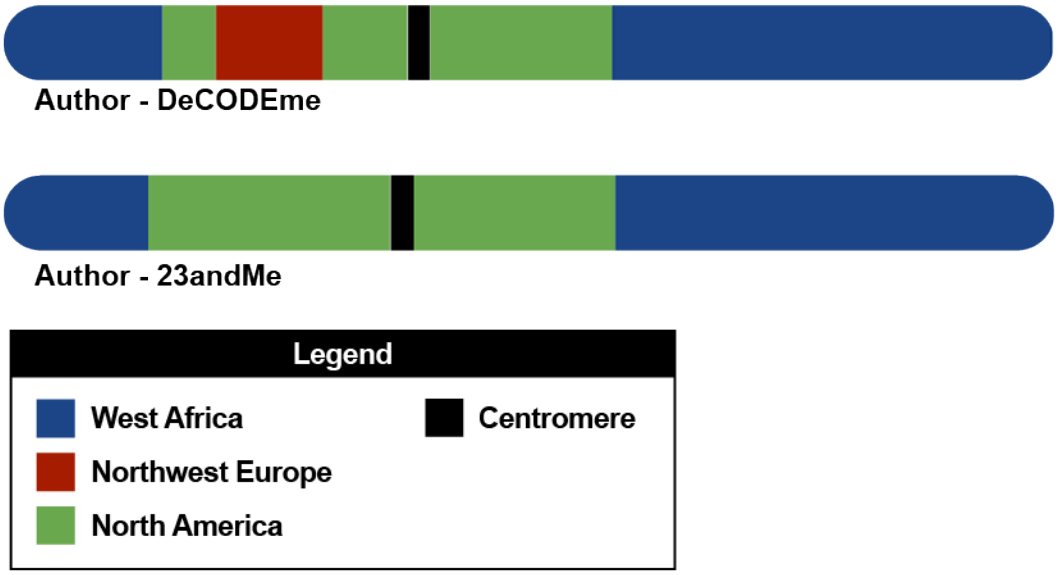

Figure 2: X chromosome biogeographical analyses by 23andMe and deCODEme. Color coding indicates the biogeographical origins of the segments.

Distinguishing Native American contributions from Asian contributions has been quite a challenge for every company the author has dealt with. Recently, 23andMe improved their approach and this segment on the X chromosome is now entirely Native American according to their analysis.

The author was motivated to take on the problem himself using 23andMe's raw data. Only three reference populations were used: Amerind, Northwest European, and West African. We saw no reason to burden the model with populations that with near certainty did not contribute to the author's genome. We suspect that deCODEme used too few populations, and that Asian was an inadequate proxy for Native American. On the other hand, it would appear that 23andMe includes so many populations in their modeling that their analysis suffers under the weight. The population data sets used for this paper were available in Stanford University and University of Michigan databases.

So, what are the components of the model? As stated above, the underlying states are Amerind, Northwestern European, and West African. We need to introduce some language at this point. A single nucleotide polymorphism (SNP) is a base-pair position that is known to exist in the human population in at least two of the four possible nucleotides (adenine, cytosine, guanine, or thymine; A, C, G, T).

At each sampled position of the X chromosome, we know what version of the SNP the author had and the frequencies at which this version occurred in each of the three populations. Table 1 displays a sampled segment of the author's observed (O) SNP data aligned with the frequencies at which the version occurs in each sample of the three populations. The leftmost column provides the position number on the X chromosome for each SNP. The notation $\mathbb{P}(A|B)$ denotes the probability of A given B. For example, we write $\mathbb{P}(\text{O}|\text{Amer})$ to represent the frequency with which an observed version was found in the sample population of Native Americans.

Position	**Author**	$\mathbb{P}(\mathbf{O}\|\mathbf{Amer})$	$\mathbb{P}(\mathbf{O}\|\mathbf{Afr})$	$\mathbb{P}(\mathbf{O}\|\mathbf{Eur})$
31443128	A	0.43	0.151	0.028
31444181	G	0.961	0.663	0.959
31450161	C	0.469	0.488	0.096
31453286	T	0.969	0.919	0.932
31464466	A	0.477	0.442	0.11
31470405	C	0.438	0.14	0.041
31482315	A	0.477	0.369	0.097
31487383	A	0.477	0.302	0.151
31511118	A	1	0.953	1
31512758	A	0.422	0.209	0.068
31520996	T	0.883	1	0.795
31531590	G	0.961	0.791	0.904
31544013	A	0.484	0.488	0.137
31566212	G	0.484	0.407	0.164
31574736	C	0.484	0.581	0.164
31579369	G	0.453	0.209	0.151
31580472	T	1	0.942	0.932
31586017	G	0.883	0.57	0.575
31591291	A	0.969	0.942	0.918
31600900	T	0.477	0.558	0.164
31602710	A	0.469	0.267	0.164
31606813	A	0.422	0.081	0.068
31611420	G	1	1	0.965
31624915	T	0.938	0.872	0.904
31642017	A	0.891	1	0.836
31647484	C	0.563	0.686	0.205
31648155	A	0.492	0.105	0.112
31652167	T	0.492	0.128	0.11

31656866	A	0.391	0.767	0.247
31656915	A	0.977	0.826	0.918
31662790	T	0.445	0.442	0.301
31663741	T	0.875	0.628	0.616

Table 1: Data for a small segment of the author's X chromosome with position numbers, author's version and the frequencies at which the author's version occurred in each of the populations under consideration.

We had a total of 12491 sampled positions ranging from one end of the X chromosome to the other. The task was to divine the probability of the underlying biogeographical ancestry (state) at each observed position given all 12491 observations. Call this probability $\mathbb{P}_{\text{all}}$(State.Position). When this was done, the above section was determined to be Amerind. The value of $\mathbb{P}_{\text{all}}$(Amerind.Position) ranges from 0.90 to 0.999 in this section. See Table 2 below.

Position	**SNP**	**$\mathbb{P}$(Amer)**	**$\mathbb{P}$(Afr)**	**$\mathbb{P}$(Eur)**
31443128	A	0.903177	0.081039	0.015784
31444181	G	0.92649	0.057352	0.016158
31450161	C	0.936347	0.06031	0.003343
31453286	T	0.93939	0.057384	0.003225
31464466	A	0.945719	0.053532	0.000749
31470405	C	0.982157	0.01777	0.000073
31482315	A	0.986096	0.013802	0.000102
31487383	A	0.991059	0.008782	0.000159
31511118	A	0.99113	0.00837	0.0005
31512758	A	0.995754	0.004165	0.000081
31520996	T	0.994838	0.004712	0.00045
31531590	G	0.995648	0.003882	0.000470
31544013	A	0.995943	0.003915	0.000142
31566212	G	0.996536	0.003294	0.00017
31574736	C	0.995879	0.003952	0.000169
31579369	G	0.998006	0.001827	0.000167
31580472	T	0.997813	0.001721	0.000466
31586017	G	0.998563	0.001112	0.000325
31591291	A	0.998446	0.001081	0.000473
31600900	T	0.998564	0.001264	0.000172

Position	SNP	ℙ(Amer)	ℙ(Afr)	ℙ(Eur)
31602710	A	0.999105	0.00072	0.000175
31606813	A	0.999781	0.00171	0.00048
31611420	G	0.999018	0.000200	0.000782
31624915	T	0.999053	0.000183	0.000764
31642017	A	0.99897	0.000561	0.000469
31647484	C	0.999134	0.000684	0.000182
31648155	A	0.99974	0.000183	0.000077
31652167	T	0.999758	0.000160	0.000082
31656866	A	0.998705	0.00098	0.000315
31656915	A	0.998702	0.000828	0.00047
31662790	T	0.998839	0.000823	0.000338
31663741	T	0.999057	0.000591	0.000353

Table 2: Results of the HMM probability analysis for the segment in Table 1.

We shall view as close to unassailable the claim that as you traverse the chromosome, once you move into a segment inherited from a given biogeographical region, you will stay there "for a while." This is our first tenet of faith. The section in the example that we just discussed represents less than one third of one percent of the chromosome. Now we add another tenet of faith: All models are wrong, but, despite this uncomfortable truth, some models are useful. We believe, in particular, that HMM is.

It would be foolhardy to assume that all readers have a knowledge of the ins and outs of Hidden Markov Modeling. We will attempt to use the technique as a black box. Nonetheless, it would be helpful to provide some insight into the mechanisms that are at play. Below we take a stab at describing a "model of the model." (See [1, 2, 3] for more on HMM.)

At each sampled position on the author's X chromosome, we effectively have for the SNP that is present at that position, the probability of that SNP given the biogeographical region. What we want to know is the probability of the biogeographical region being the contributor, given the SNP. The reader who is familiar with Bayes' theorem will not be surprised that this theorem is a foundational piece in the rather elaborate machinery of HMM. In its design, the process moves from one end of the chromosome to the other assessing the probability of the biogeographical region at each sampled position, given all

that has been observed prior to that position. The machinery then moves from the opposite end back to the beginning, computing at each position the probability of each biogeographical region being the contributor, given *everything* that has been seen before *and* after each position.

One can imagine a passenger on a highway moving from east to west across the country, attempting to establish the position of boundaries that occur between rural, urban, and suburban regions, based on observations. Perhaps the presence of a silo indicates that he is now moving through a rural area, but where did the rural area begin? Perhaps the silo is simply a remnant that indicates that the area was once rural, but now the silo stands on a small vacant lot in suburbia. He then reverses his direction and moves east, making observations and reconciling them with what was seen when moving west. When the observer begins his round trip, he must make a guess as to which type of region he finds himself in initially, with essentially no data. Knowing a priori the probability of making a transition from one regional classification to another would also be helpful. An initial guess of these transition probabilities is all the observer has. At the end of the roundtrip, re-estimations of transition probabilities and the initial state are made based on all that has been observed during the round trip. Then a second roundtrip is made. If we can imagine HMM being implemented in this process, it is an important feature of the methodology that each successive roundtrip is guaranteed to give us better results than the previous one, the improvement being based on a measure called the *score*. The score is a probability, so it is bounded above by 1. That forces a convergence to what will at least be a local maximum. This somewhat tortured "model" of Hidden Markov Modeling might be useful for the reader as we move through the analysis.

What we initially have is simply the data described above. We do not know the probability of making a transition on the chromosome from a segment contributed from one biogeographical region to a segment contributed by another. This gets codified in what is called the *state-transition matrix*. Further, we do not know what biogeographical region contributed the first segment that we encounter. One of the strengths of HMM is that we can make an initial estimate of these parameters and then allow the re-estimation machinery of the model to modify them for the better. The values that we have when we converge to equilibrium are the ones that we will have to live with. We are at equilibrium when the score no longer improves, the score being a measure of the probability of the sequence of 12491 observations. For computational reasons, we actually use the logarithm (base 10) of the score, instead of the score itself.

Since the scores are less than 1, the logarithm of the score is negative. The closer the score is to zero, the better it is. For instance, -1000.5 is better than -2103.2. Figure 3 represents the result that we get using HMM.

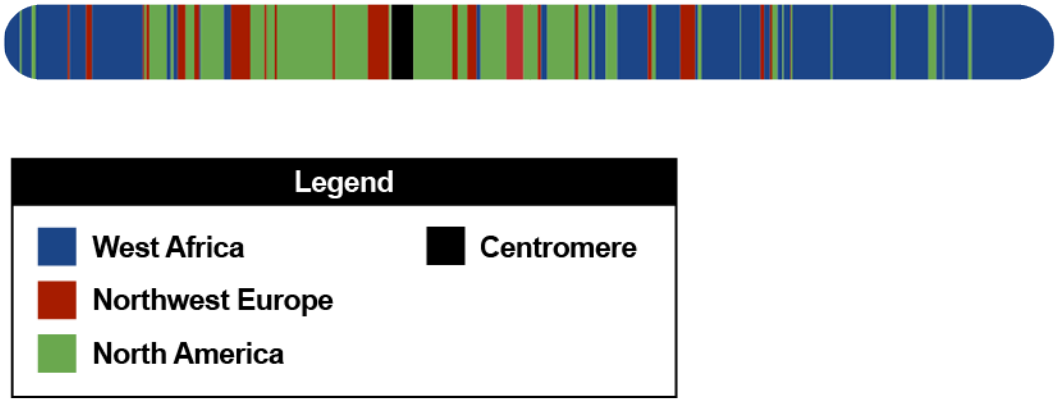

Figure 3: Result of X chromosome biogeographical analysis using HMM for author's X chromosome. Color coding indicates the biogeographical origins of the segments.

We list below some of the choices that we made for the initial state-transition matrices along with the number of iterations to equilibrium. To provide orientation, the entry in the *kth* row and the *jth* column is the probability of transitioning from the *jth* state at position n to the *kth* state at position $n+1$. The means that the matrices are necessarily column stochastic (columns add to 1).

Each initiation with an arbitrarily chosen state-transition matrix led to the same score, $\mathbb{P}_{\text{all}}$(State.Position) values and final state-transition at equilibrium. In all cases the initial probability for each population was taken to be $1/3$. (We do not show the 12491 probability assignments, but most assignments were made based on the $\mathbb{P}_{\text{all}}$(State.Position) being greater than 0.90. See Table 2.) A matrix is said to be symmetric if the value in row j and column k is always the same as the value in row k and column j. The iterations given in Tables 3 and 4 represent the number required to converge to equilibrium. Our symmetric matrices are column stochastic and so the column values must sum to 1. As a result, the symmetric matrices we use are of the form:

$$\begin{pmatrix} a & \frac{1-a}{2} & \frac{1-a}{2} \\ \frac{1-a}{2} & a & \frac{1-a}{2} \\ \frac{1-a}{2} & \frac{1-a}{2} & a \end{pmatrix},$$

where $0 < a < 1$.

Most of the initial states that we ran with our model were with symmetric matrices, because that allows us to initiate without a bias toward any region. However, all the initiations that we ran with non-symmetric matrices converged to the same result.

The model was tested under the following conditions.

a	iterations
0.990	46
0.340	74
0.100	109
0.005	184

Table 3: Model trials.

The model with a non-symmetric matrix such as

$$\begin{pmatrix} 0.6 & 0.5 & 0.7 \\ 0.3 & 0.2 & 0.1 \\ 0.1 & 0.3 & 0.2 \end{pmatrix}$$

was also run with 75 iterations, with similar results.

Although there is nothing in the methodology of HMM that guarantees that we have found the maximum score, choosing such a disparate set of initial conditions and finding that they all result in the same score after convergence to equilibrium provides a level of confidence that we have in fact succeeded in doing precisely that.

Author	**Amerind**	**West African**	**NW European**
Amerind	0.9792	0.006653	0.019388
West African	0.01304	0.99029	0.011955
NW European	0.007412	0.003061	0.96866

Table 4: Consensus for all initiations (12,491 observations), Score: 2403.80695324, state-transition matrix at equilibrium.

Since this analysis is not time-dependent, one might wonder about the arbitrariness of starting the analysis at one end of the chromosome versus the other. The result in Table 4 shows the case when we start from the short-arm end of the chromosome. Table 5 shows what we get when we feed in the data starting from the long-arm end of the chromosome.

Author	**Amerind**	**West African**	**NW European**
Amerind	0.9792	0.00697	0.01801
West African	0.01245	0.99029	0.01333
NW European	.00836	0.002744	0.96866

Table 5: Starts on the long arm of the chromosome. Score: -2403.8076461, initial state-transition matrix was symmetric with 0.99 on the diagonal; 47 iterations.

Note that the diagonals and the score are essentially identical to what we got when we fed in the data starting from the other end of the chromosome. Most importantly, the $\mathbb{P}_{\text{all}}$(State.Position) values that we found are also identical at every position, so the biogeographical assignments are unchanged. What about the off-diagonal values in the transition matrix? After a moment's reflection, we realize that we should expect these values to be different, because the probability of transitioning from one state to another is naturally order-dependent for the object that we are analyzing. The values on the diagonal represent something like a "thickness" measure, the tendency to stay in a biogeographical ancestry. Thickness is not order-dependent. The diagonal values support our tenet of faith. The probability that a given position is not on the boundary of segments contributed from two different biogeographical regions is high. In this particular analysis it was at least 0.96866.

We should return to the graphic that represents our results, because it's the bottom line. With its high degree of fragmentation, the graphic we produced looks nothing like the two produced by the companies. There is no apparent way to determine which of the three analyses (theirs and the author's) is closest to the truth.

To move beyond this impasse, it would be helpful to analyze the X chromosome of someone who can say with confidence that their ancestry is only from one of the regions. Figure 4 shows the results of the analysis for a person whose ancestry has been thoroughly researched and is believed to have genetic contributions only from northwestern Europe.

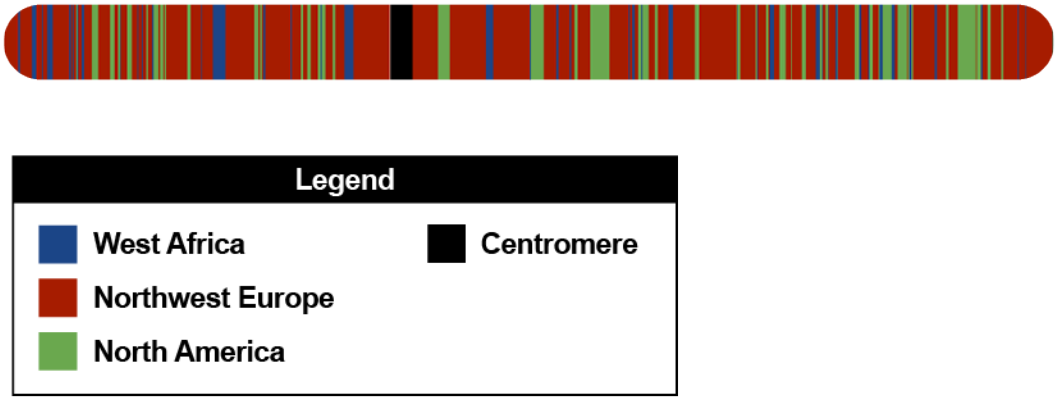

Figure 4: The result of X chromosome biogeographical analysis using HMM for non-admixed European man's X chromosome. Color coding indicates the biogeographical origins of the segments.

It would appear that not only is the model wrong; it is terribly wrong and not useful. However, the title of this paper is not "The Folly of Using Hidden Markov Modeling for Biogeographical Ancestry Analysis." So this is not the end of the exercise.

Smoothing

At this juncture it is pretty clear that the methodology is producing results that are noise-ridden. Our approach is in need of modification and our thinking turned to reducing what we choose to call "volatility."

Instead of using the population frequencies at each sampled position, we use the arithmetic mean of the observed frequencies in non-overlapping blocks of n consecutive samples, where we hope to establish what value of n is optimal. Rest assured, the HMM does not know that we're "cheating." It certainly does not know that we are using average frequencies instead of individual frequencies. The block size, over which we will average, is so small that the entire block almost certainly came from only one of our recent ancestors. Admixture persisted in the analysis for our person of strictly European ancestry for block sizes up to eight sampled positions. It vanished at block-size 9. Here Tables 6-7 represent the converged block-size-9 state-transition matrices, for the non-admixed European man and the author, respectively.

European Man	Amerind	West African	NW European
Amerind	0.03299	$3.8555*10^{-6}$	$3.7519*10^{-7}$
West African	$1.72*10^{-20}$	$4.663*10^{-15}$	$3.754*10^{-22}$
NW European	0.96700	0.99999	0.99999

Table 6: Converged block-size-9 state-transition matrix for European man.

Author	Amerind	West African	NW European
Amerind	0.9979	0.0012	0.994
West African	0.0021	0.9988	$8*10^{-142}$
NW European	10^{-117}	$3*10^{-13}$	0.006

Table 7: Converged block-size-9 state-transition matrix for the author.

Is what we have at this point right? Figure 5 for the author is now essentially identical to the one produced by 23andMe, but it lacks the European ancestry asserted by deCODEme.

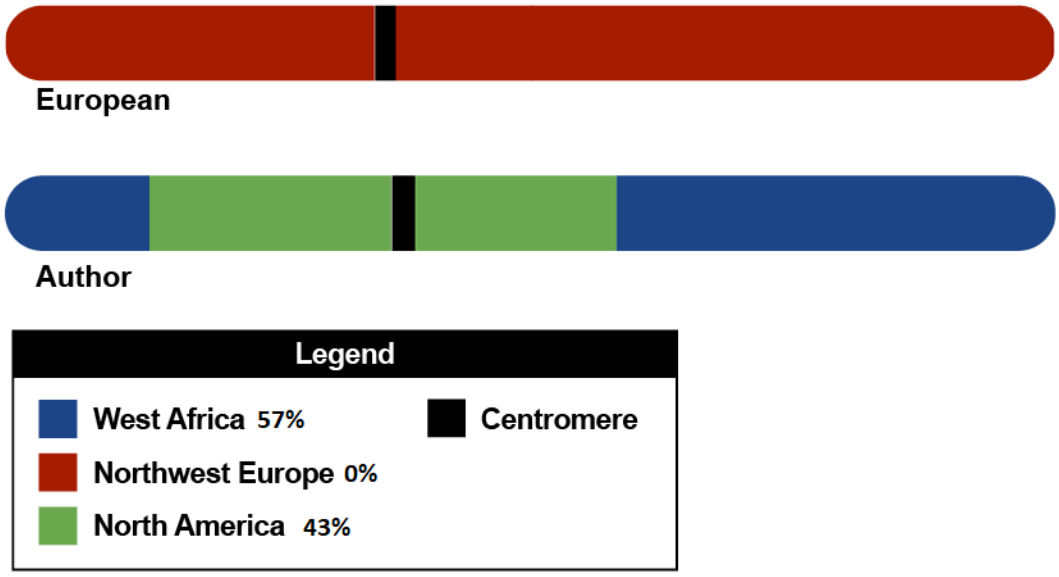

Figure 5: These are X chromosome graphics corresponding to Tables 6-7. The legend provides the percentages for the author.

We've used the same smoothing window for both subjects. Perhaps we should look at the score that the non-admixed European test case had when the "noise" disappeared and then select a block size that gives a comparable score for the author. We remind the reader that the scores are negative and for that reason scores that are smaller in absolute value are larger.

Block Size	Author	non-admixed European
1	-2403.8	-2462.7
3	-2171.0	-2277.0
7	-2097.4	-2223.2
8	-2091.1	-2214.4
9	-2081.0	-2209.9

Table 8: Adjusted Scores

The adjusted scores in Table 8 are found by multiplying the actual score by the size of the block. We multiply by the block size because averaging over a block size of n reduces the number of observations the model "sees" to the original number of observations divided by n. When we consider how probabilities are computed in this context (multiplication), together with the fact that we are taking the logarithm of the result, it seems reasonable to adjust in this way to make comparisons. As an example, for $n = 2$, we would have half as many observations and the score could be expected to be about the square root of the score for the original sequence of unaveraged frequencies. The logarithm of that score would be roughly one half the logarithm of the original score.

Using these crude adjustments, it appears that block-size 3 produces a better score for the author's X chromosome analysis than block-size 9 does for the control (European) case. (Block-size 2 does not.) If we use the score as the criterion for getting the correct amount of noise reduction, while limiting the risk of erasure of real contributions by over-smoothing, we get the results for the author shown in Figure 6 (also see Table 9).

Note that with these adjustments the state-transition matrix for the author has a dominant diagonal, while a glance back at the European control case reveals a dominant European row.

On the basis of the adjusted score, the author suspects that averaging using block-size 9 is than better than using the block-size-3 results. Further, observe that for the author's X chromosome the adjusted scores at block-size 10 begin to decrease, after having steadily improved up to block-size 9. See Table 10.

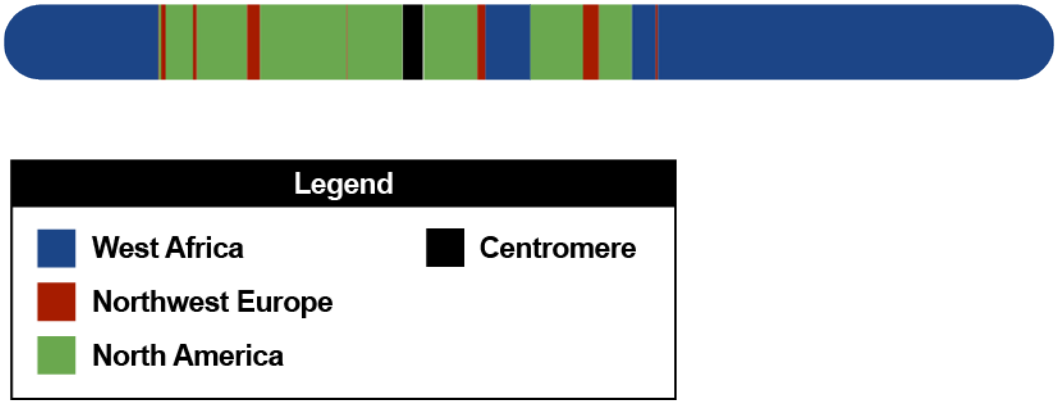

Figure 6: The HMM results for the author using block-size 3.

Author	Amerind	West African	NW European
Amerind	0.991426	$3.8555*10^{-6}$	0.050368
West African	0.001412	0.998596	0.008261
NW European	0.007162	0.001182	0.941371

Table 9: Converged block-size-3 state-transition matrices for the author.

Block Size	Adjusted Score
1	-2462.7
3	-2277.0
7	-2223.2
8	-2214.4
9	-2209.9
10	-2084.77
20	-2082.02
80	-2160.76

Table 10: Adjusted score for various block sizes.

For the European control case, the score continued to increase after block-size 9, but the result did not change. It remained 100 percent European.

Throwing in this last bit of evidence leads to choosing the block-size-9 result, which we show again below (Figure 7) as the "winner," along with the corresponding state transition matrix (Table 11). Note that the diagonal values of the state transition matrix are large only in the positions that correspond to the Amerind and West African populations.

Author	**Amerind**	**West African**	**NW European**
Amerind	0.9979	0.0012	0.994
West African	0.0021	0.9988	$8 \cdot 10^{-142}$
NW European	10^{-117}	$3 \cdot 10^{-13}$	0.006

Table 11: The converged state-transition matrix for the author.

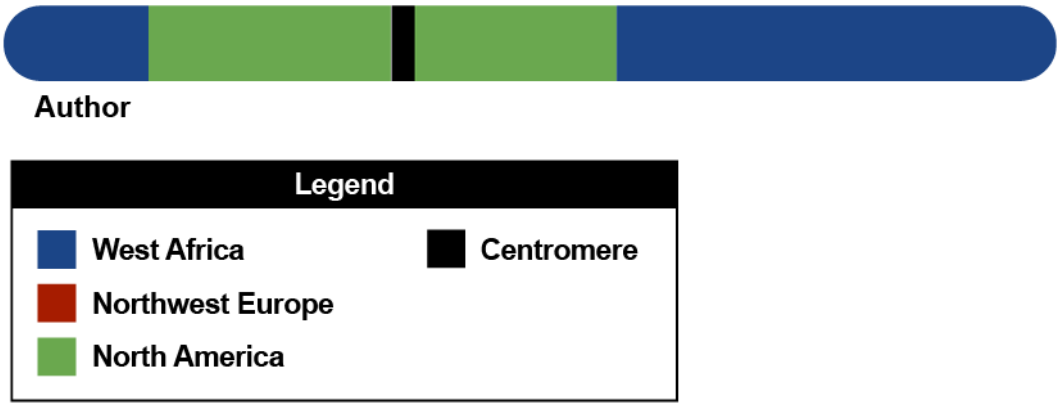

Figure 7: The "winning" graphic for the author.

In Table 11, the value 0.994 in the topmost row might be puzzling. This is best interpreted as an indication of the affinity of the European SNP data characteristics with those of the North American data and sheds light on why there was so much European "noise" for block sizes smaller than 9.

Just for fun, we recall that of thirty-two 3rd great grandparents, only eight of them are on the "glide path" to my X chromosome. Suppose that the Native American ancestor who made the contribution was a 3rd great grandmother.

She would have been one of eight (sixth Fibonacci number) ancestors from that generation who could have contributed to my X chromosome. Figure 8 displays in green a possible path for her contribution.

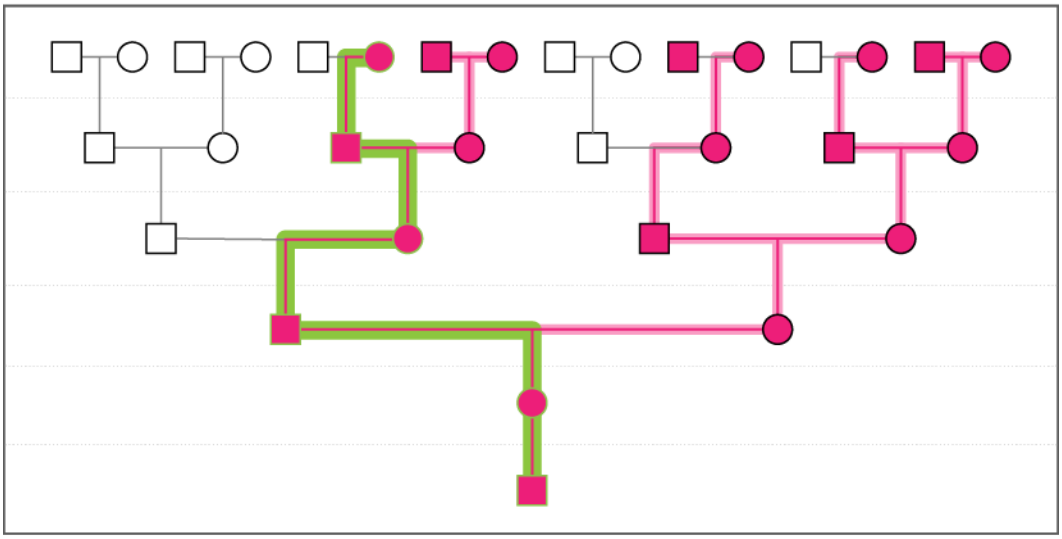

Figure 8: X chromosome contribution path (in green) from the 3rd great grandparent that minimizes recombination events. No paternal ancestors contribute to a male's X chromosome and, for sake of simplicity, the paternal branch is not shown. (Circles are female ancestors, Squares are male.)

Along the green path there are only two recombination events, since the males pass the X to their daughters virtually unchanged. If this scenario corresponds to fact, it is not at all surprising that a large contribution from this 3rd great grandmother might remain intact all the way from her, a woman who was likely born in the 1700s, to the author.

Conclusions

The observations in this study lead the author to conjecture that perhaps the adjusted score is the right mechanism to use as an indication that the correct block size has been chosen. The rule would be: If the analysis shows contributions from more than one biogeographical region, continue to increase the block size until there is only one biogeographical region represented in the analysis or until the adjusted scores level off or begin to decrease.

Of course, we would need to do this kind of analysis for a large sample of people, whose ancestry is well-known, before trying to draw firm conclusions. We have analyzed three other non-admixed Europeans with results consistent with the case shown here. We have also done the analysis for three people whose ancestry is not so well-established and happen to be closely related to the author. The approach also holds up in those cases.

Again, we would need to handle a lot more cases, before trying to sell this as being sound analysis. There is also the question of how many different populations the HMM approach will support at once. Here the analysis was limited to three populations because I was confident that I knew what populations could possibly have contributed to my genome. This is obviously not the case in general.

As for the East Asian versus Amerind difficulties that researchers have encountered, we've addressed this in a straightforward way in our analysis by taking segments that have been assigned as Amerind and holding a bake-off on the segments. When doing this, we take West Africa and Northwest Europe out of the model. Instead, we use only the Amerind population and introduce one East Asian population on the segments that had been designated Amerind in the original three-way analysis. Amerind prevailed, overwhelmingly, for each choice of East Asian population, Yakut, Han, etc.

References

[1] Jeff A. Bilmes, *A Gentle Tutorial of EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models*, U.C. Berkeley, Technical Report TR-97-021, 1997. Available at http://melodi.ee.washington.edu/people/bilmes/mypapers/em.pdf, last accessed on July 17, 2019.

[2] Alan B. Poritz, "Hidden Markov Models: A Guided Tour", *Proceedings of ICASSP-88 (International Conference on Acoustics, Speech, and Signal Processing)*, (Institute for Defense Analyses, 1988), pages 7–13.

[3] Lawrence R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", *Proceedings of the IEEE*, Volume **77** Number 2 (February 1989), pages 257–286.