Evaluation of the Product Rule

Charles H. Brenner
DNA-VIEW
2300 Grand Canal
Venice, California 90291
USA

A couple of years ago a couple of population geneticists generated a stir with an opinion piece in Science [1]. This is somewhat remarkable considering that the article was nothing but speculation, extrapolation, and opinions. It contained no new data, no new ideas, and no science. Repercussions continue in American courts to this day.

Their fundamental concern was with the question of statistical independence of genetic traits. Traits A_1 at locus A and B_2 at locus B are statistically independent if A_1 occurs with the same probability in the presence of B_2 as in the absence. If independence holds, then the "product rule" holds: the proportion of e.g. sperm cells with the combination is simply the product of the proportions with the individual traits. Loosely speaking, we can say that the loci are independent if all their combinations of traits are independent. If the loci are not independent then of course the product rule fails, and on average (average over all combinations weighted by their chance of occurrence) always fails in the awkward direction of underestimating the probability of a combination.

Loci will obviously be dependent if they are linked on the same chromosome or if some issue of selectivity arises e.g. because some combinations are more or less hardy. These mechanisms are not of realistic concern when dealing with RFLP's on different chromosomes so sometimes Europeans and others living in countries more homogeneous than the United States falter at the post and fail to give Lewontin and Hartl's argument even what consideration it deserves. But the population geneticists are concerned about the possibility of selective breeding -- maybe people with trait A_1 preferentially seek partners with B_2. Of course they can't do it on purpose because B_2 is not expressed, but it could still happen. For example perhaps A_1 and B_2 are mostly prevalent among Blacks and most people mate according to skin color.

Selective breeding by race is not a problem because everybody knows (or at least assumes!) that statistics and calculations should be made separately race by race for this very reason and for other reasons. But might there not be some less obvious manifestations of "population substructuring" -- "ethnic substructuring" such as Poles mating with Poles, as Lewontin and Hartl suggest?

Perhaps. Then it is worth asking, how big would the effect be? It's not important that there may be some specific examples of combinations of traits that are non-independent between two loci. Specific examples are just anecdotes -- of course if they exist and you know what they are, then you should watch for them and make adjustments when they arise. But the real concern is the average impact of such combinations -- i.e. what is the expected error of the product rule?

A hypothetical population

Here's an example of the possible situation: At some mythical locus **A**, California population data shows 10 equally frequent alleles, A_1, A_2, ..., A_{10}. However, we find that F_{st}=2%, which in this case just means a 2% excess of homozygotes. Perhaps spurred on by this anomaly we analyze further and eventually hit on the winning idea of graphing the data according to political affiliation. California Democrats and Republicans are equinumerous and alike with respect to most alleles, but there is one allele, A_1, on the far left that is exclusive to Democrats (frequency=20% among Democratic alleles), and one far right allele, A_{10}, that is an invariable indicator of Republicanness (frequency=20% among Republican alleles). F_{st}=2% is a high figure compared to substructuring values suggested in the literature [e.g. 2].

To continue the thought experiment, imagine that there are several other similar loci, with alleles B_1, ..., B_{10} and C_1, ..., C_{10}, etc, where B_1, C_1, are purely Democratic, etc. Every allele has frequency 1/10 but 2-locus combinations ("haplotypes") like A_1B_1 have frequencies greater than 1/100 -- namely 1/50. It thus turns out that the chance that two sperm selected at random have the same haplotype is 1.04/100 -- a 4% error for the product rule.

It is interesting to consider the error across 3 or more loci. This example is easy to calculate, and with 3 loci the percentage error using the product rule is 3 times as great (the correct matching chance is 1.12/1000), with 4 loci about 6 times as great, and so on. The reason is easy to see: with three loci A, B, and C there is statistical dependence between triple the number of combinations: e.g. not just $A_1B_1C_x$ but also $A_1B_xC_1$ and $A_xB_1C_1$. Similar calculations with slightly more complicated models give the same sort of behavior, namely:

> To the extent that the product rule fails because of substructuring, as the number of factors increases the relative error grows, but rather slowly -- approximately as the square of the logarithm of the matching odds.

But this is a digression. The question remains as to whether the magnitude of substructuring built into the "California" model is relevant to amounts that one might be confronted with in the real world.

A real substructured population

To this end we consider the real DNA typing data from 3235 Virginians all typed in each of the 4 probes YNH24, TBQ7, EFD52, and CMM101 (D2S44, D10S28, D17S26, D14S13). The Virginians are about half Black and half Caucasian. Examining the Caucasian and Black data separately makes it clear that the Virginians are a substructured population. The Nei genetic distances between these races are 0.24, 0.12, 0.29, and 0.27 respectively for the four probes, compared to 0.4 for Republicans vs Democrats. F_{st} values for the mixed race population are slightly lower than in the California example.

A single test of the product rule is defined as follows. A person is selected, and a 2, 3, or 4 locus haplotype is selected from the person by choosing an allele at random from each of several loci. An observed PI (so called because of the relationship of this experiment to a paternity case) **obs** is then obtained by observing the reciprocal of the fraction of other people who are potential donors of approximately the same

haplotype. This is compared with an expected PI calculated by counting the reciprocal matching allele fraction locus by locus, then using the product rule to predict **obs**. Approximately 50,000 haplotypes were sampled in this way. Grouping the data according to ranges of values of **exp**, **exp** and **obs** were compared by log-linear regressions. With **exp** < 2500 there was no discernable difference. Above that the number of observed multiple-locus matches was usually 0, so the statistical result depends on the convention chosen for the reciprocal of 0. Put otherwise, despite a conspicuous amount of built-in substructure, the product rule proves accurate to the limits of the available data.

Now, Lowentin claimed [3] that inter-ethnic differences are about 50% greater than inter-racial. Others have disputed this using different statistics or different data. However from our present point of view the dispute is splitting hairs -- within a factor of 2 one way or the other the magnitudes are the same [4]. The experiment with the mixed-race Virginians revealed no perceptible error in the product rule. Hence it is extremely unlikely that a mixed-ethnic population would give any different result.

Conclusions

A priori computations show that the product rule cannot fail significantly with any remotely believable amount of substructuring. An examination of live data confirms that with a plausible amount of substructing there is no discernable error in the product rule for product frequencies down to 1/2500. By computation on simple models it seems that product rule errors due to substructuring can only creep in gradually. Therefore, the conclusion that the product rule is a reasonable approximation can be extrapolated to even very much smaller products.

It is also worth noting in passing that the usual practice of computing a mixed-race frequency race by race is needlessly punctilious.

Acknowledgements

Thanks to Dan Demers and the Fairfax Identity Lab in Fairfax, Virginia for frequency data, and to Jeff Morris for useful ideas.

References

[1] RC Lowentin & Daniel L Hartl, Population genetics in forensic DNA typing, Science 254:1735-1739 (1991)

[2] NE Morton, Genetic structure of forensic populations, Proc Natl Acad Sci USA, 89:2556-2560 (1992)

[3] RC Lowentin, Ethnic diversity, Evol Biol 6:381 (1972)

[4] RC Lowentin, letter, Science 255:1054-1055 (1992)