Monte Carlo Sampling: The Probability Distribution of Identity States

Danni Sun
Shuxin Yu

University of Illinois at Urbana-Champaign

November, 2013

Abstract

In genetic epidemiology, disease association is a popular examination designed to test whether the alleles are correlated with the disease identified. However, operating a pedigree of enormous data size like more than 1000 individuals with the disease association studies is challenging. In this paper, we discuss the methods which could be used to enable the disease association on the very large pedigrees. We choose the diploid Wright-Fisher Model to simulate the probability distribution of identity states through the Monte Carlo Sampling. Since kinship coefficient is the expectation of the identity states probability, so then the kinship coefficients would be an effective and efficient tool to analyze the relatedness in a case-control association study.

1 Introduction

A pedigree is a graph designed to represents the family relationship in genealogy which is widely used in genetic epidemiology. In disease association studies, tracing all alleles in the pedigree, genotypes can be determined if they have characteristics of the disease. A comparison between two groups - the case group and the control group - indicates the correlation of the disease to a genotype in a family studies.

To analyze the disease association, linkage studies would be considered as the most accurate method. These track transmissions of genes in a known pedigree so linkage studies can provide strong evidence of genetic inheritance, however, it would be challenging to apply linkage analysis to a large pedigree with very large size. The running time for linkage analysis is exponential in the number of individuals in a pedigree. Considering the genome-wide association studies (GWAS), it investigates many common genetic variants in a population, however, without the pedigree, the information is not accurate enough in presence of the population stratification or cryptic relatedness .

In this paper, we research on the the method that focus on the strength of the linkage analysis and the GWAS that would be faster than the linkage analysis and more accurate than the GWAS. We focus on the part of the pedigree which is related to the pair-wise individuals.

Through the Monte Carlo Sampling to simulate the probability distribution of pair-wise identity states in different pedigrees and then use the kinship coefficient to explain the disease association. Without doubt, this method would be more accurate than GWAS. Therefore, we hope this research would contribute a lot in the fields of forensics or relationship inference and pedigree inference studies.

2 Background

In this research, we started on the Wright-Fisher Model to create the pedigree and then simulate thousands of different inheritance paths for each pedigree to result the probability distribution of pair-wise identity states.

2.1 Inheritance Path

Inheritance Path is the map of inheritance routines of all alleles in a pedigree. Although the inheritance path is constrained by the relationship within the pedigree, it is not determined by them.

2.2 Identity States

In our improved W-F model, each individual has 2 alleles: one allele inherited from the mother is called maternal allele and the other from the father is called paternal allele. For two randomly chosen alleles, if they are inherited from the same ancestor, then we connect them by an edge in a graph named the identity state. In this graph, four alleles of two selected individuals A and B are marked by A_1, A_2, B_1, B_2 and are ordered as a square. Therefore, each four alleles should have 15 possible identity states that are indexed as $i = 1, 2, 3, ..., 15$ (as in Figure 3). Meanwhile, we can group identity states into two categories, inbreed and outbreed states. By definition, an individual is outbreed if 2 parents are not related and an individual is inbreed if not outbreed.

To figure out the probability distribution of these 15 identity states, we use two methods, forward-in-time and backward-in-time, to sample in the improved Wright-Fisher model.

2.3 Kinship Coefficient

The identity states are used to compute the kinship coefficients. The kinship coefficient for a pair of individuals of interests is defined as the probability that a randomly chosen pair of alleles from each individual of interest is identical by descent. Let the matrix Φ

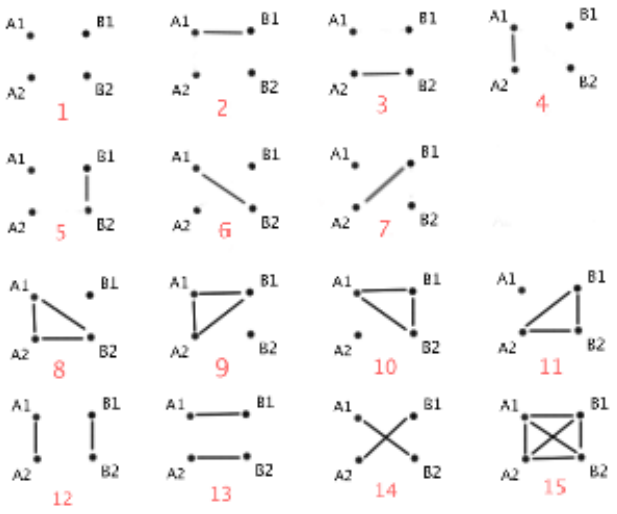


Figure 1: Index of Identity States

contain all the pair-wise kinship coefficients in a pedigree. Entries $\Phi_{ij}, i \neq j$ are kinship coefficients between two individuals and entries Φ_{ii} being inbreeding coefficient, which are the coefficient between an individual's two alleles or the kinship coefficient between the individual's two parents. Label the alleles of individuals i and j with distinct labels: (A_1, A_2) and (B_1, B_2). By the definition of kinship,

$$\Phi_{i,j} = (1/4)Pr[A_1 \sim B_1] + (1/4)Pr[A_1 \sim B_2] + (1/4)Pr[A_2 \sim B_1] + (1/4)Pr[A_2 \sim B_2]$$

where $\sim$ indicates identity by descent.

2.4 Markov Chains

A Markov process is a random process in which the future state depends only on the current state, not on the sequence of the previous events. In another words, given a set of states, $S = \{s_1.s_2, ..., s_r\}$, each move from one state s_i to another state s_j has a probability p_{ij}. Then, a Markov Chain is a collection of random variables $\{X_t\}, t \in \mathbb{N}$

$$P(X_t = j | X_0 = i_0, X_1 - i_1, ..., X_{t-1} = i_{t-1}) = P(X_t = j | X_{t-1} = i_{t-1}) = p_{i_{t-1}j}$$

2.4.1 Wright-Fisher Model

The reason we choose the Wright-Fisher Model rather than other genetic drift models is the idealized assumptions that simplify the inheritance but it is still reasonable to be applied in a large pedigree. In the Wright-Fisher Model, we assume there is a population of N diploid asexual individuals (which means there are $2N$ genes at each generation). Since the population is fixed, some external environmental factors could be ignored. Also, the W-F model supposes that generations do not overlap and no mutation happens during an inheritance. At the next generation, each individual

receives two genes that each one is selected randomly and with replacement from the previous generation. To express the process mathematically as follows:

- N asexual individuals, $2N$ genes
- At generation $n = 0, n = 0, 1, ..., x$ of these genes are the same type(they are inherited from the same gene), $(0 \leqslant x \leqslant 2N)$
- X_n is the random variable for the number of one type of same genes at time n
- $(X_n|X_{n-1} = x_{n-1}) \backsim \text{Bin}(2N, x_{n-1}/2N)$
- $P(X_n = x_n|X_{n-1} = x_{n-1}) = \dfrac{2N!}{x_n!(2N - x_n)!} \left(\dfrac{x_{n-1}}{2N}\right)^{x_n} \left(1 - \dfrac{x_{n-1}}{2N}\right)^{2N-x_n}$

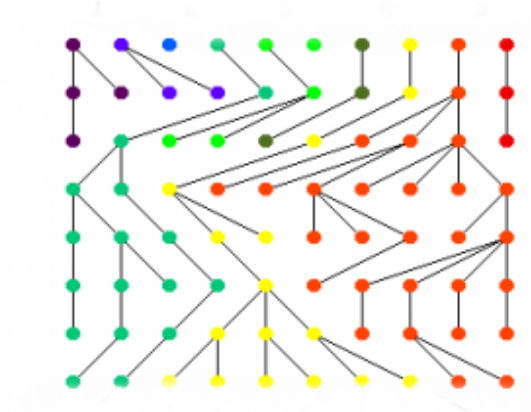


Figure 2: Wright-Fisher Model

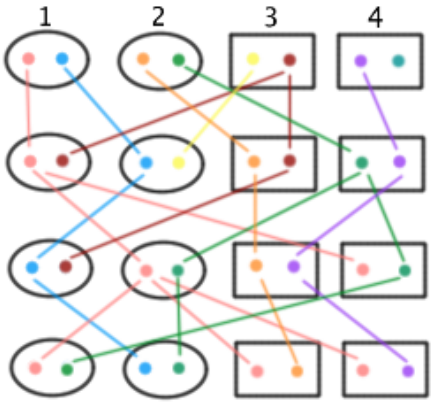


Figure 3: Improved Wright-Fisher Model

2.4.2 Improved Wright-Fisher Model

However the Wright-Fisher model ignores gender difference, we improve this model by adding the effect of the gender to make it more realistic. In our model, we keep some assumptions in the W-F model that there is no overlapping and no mutation. Then there are $2N$ diploid individuals with gender at each generation that the half is females and the other half is males. Similarly, for the next generation, each gene is derived from the previous generation by sampling randomly with replacement, but for each individual's two genes, one is selected from the father's genes and the other is from the mother's genes. We can imagine as each individual have two genes in different genders, one male and one female.

3 Methodology

3.1 Forward-In-Time Sampling

Our improved W-F model (N, G) consists of N males and N females per generation and the number of G generations. We call individuals at the first generation founders. The steps are as follows.

- To begin with, we generate the pedigree for these $2NG$ individuals following the rule: each individual uniformly picks his/her parents, one male indexed as m, $m \in \{1, 2, ..., N\}$ and one female indexed as f, $f \in \{N+1, N+2, ..., 2N\}$.
- Then, at the latest generation, we can choose two individuals uniformly without replacement. Since we focus on the distribution of identity states which are obtained from inheritance path, we move on to build up the inheritance within the pedgiree .
- To generate the inheritance path, we set a binary variable $X_e \in \{0, 1\}$ since each one has two alleles. So $X_e = 0$ means that the allele inherited from the parent is the grand-paternal allele, otherwise, if $X_e = 1$, the allele inherited is the grand-maternal allele.
- The next step is to trace each alleles' inheritance path from the top to the bottom generation. Through a data structure of Depth First Search, we mark all alleles if they are inherited from the same ancestor, and then we can find out the genome map based on its pedigree and inheritance path.

If we fix a pedigree and generate a number of inheritance paths, like sampling multiple positions in the genome of the fixed pair of individuals, we can get a series of statistics which may fit the realistic whole-genome simulations.

In order to understand this procedure of the forward-in-time sampling, we take a simple example to explain a bit more.

Example 1 *As in the Figure 4, we set $G = 3, N = 1$ which means there are three generations and one male and one female per generation. Although, with this setting, there is only one pedigree, it could have different inheritance paths.*

With this inheritance path in the figure, we index four alleles of the founders and set each allele's path a unique color. If some alleles have the same color then mark them with the same index.Then we can get a genome map represented as a matrix.

$$\begin{bmatrix} 1 & 2 & 3 & 4 \\ 1 & 4 & 2 & 3 \\ 1 & 3 & 1 & 3 \end{bmatrix}$$

Through this matrix, we found $A_1 = 1, A_2 = 3, B_1 = 1, B_2 = 3$. Hence, we get $A_1 = B_1, A_2 = B_2$ and the corresponding identity state should be 13 as the picture follows.

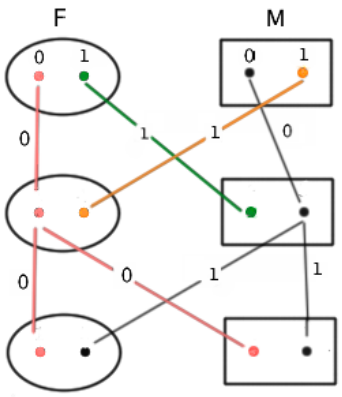
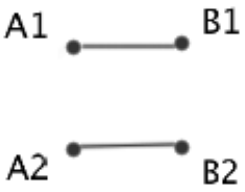


Figure 4: $(3, 1), G = 3, N = 1$



3.2 Backward-In-Time Sampling

To save running space and running time, we switched to another method called backward-in-time sampling. Unlike in forward-in-time sampling that generates the whole pedigree and inheritance paths for $2NG$ individuals, we only take care of the persons who are connected to the selected two individuals at the latest generation. The steps are as follows.

- Repeat the same procedure as in the forward-in-time to generate the pedigree: for each individual in the current generation, choose an index uniformly from $\{1, 2, ..., N\}$ to represent the male parent and uniformly from $\{N+1, N+2, ..., 2N\}$ to represent the female parent.
- In the graph data structure, add edges to the children until reaching the founders. Then we can get a family tree.
- Then focusing on the four concerned alleles, we trace along the pedigree from the bottom to the top generation until there exists a node (i.e. some pair of children have the same parent).
- Then check if they are inherited the same gene from their same parent by randomly flipping the binary variable X_e. If two binary variables are the same, then it means these two alleles are inherited from the same allele so that they are identical. If not, then keep upward until reaching the founders.

By this way, there are at most 2^G individuals in the pedigree instead of $2NG$ and we only concentrate on the inheritance path of four alleles rather than all individuals. Also we can fix a pedigree and generate a number of different inheritance paths to get a series of data.

Similarly, we would like to use an example here to explain this method.

Example 2 *In this example, we set* $G = 3, N = 3$ *(i.e. there are* 3 *generations and* 3 *males and* 3 *females per generation) as in the Figure 5. Since we just generate the pedigree of connected individuals so that there are only ten individuals rather than eighteen in the forward-in-time method.*

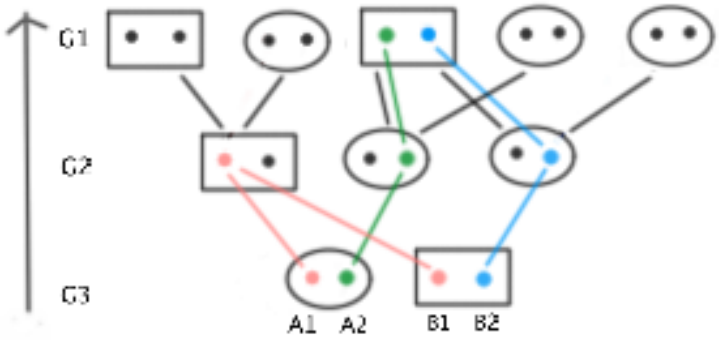


Figure 5: Backward-in-Time: (3, 3), G=3, N=3

Within the pedigree, we trace along the pedigree to see if there is a node. Then we found A_1, B_1 have the same parent, so we flip the binary variable and get that they are both from the same allele, i.e. $A_1 = B_1$. Also there is another node at the first generation so that A_2 and B_2 have the same grandfather. However, the binary variable gives us different values, which means they are inherited from different alleles.

At the end, we get $A_1 = B_1$ and the identity state is 2 as the picture follows.

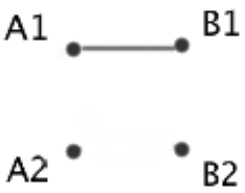


Figure 6: Identity State 2

3.3 Results of Forward and Backward

Running the forward-in-time and backward-in-time sampling, in order to make the model more realistic and stable, we let $G = 5, N = 5$, and run 300 different pedigrees that each runs 5000 different inheritance paths, 500 different pedigrees that each runs 3000 different inheritance paths, and 500 different pedigrees that each runs 5000 different inheritance paths, which all result in a decreasing probability distribution as Figure 7, Figure 8 and Figure 9.

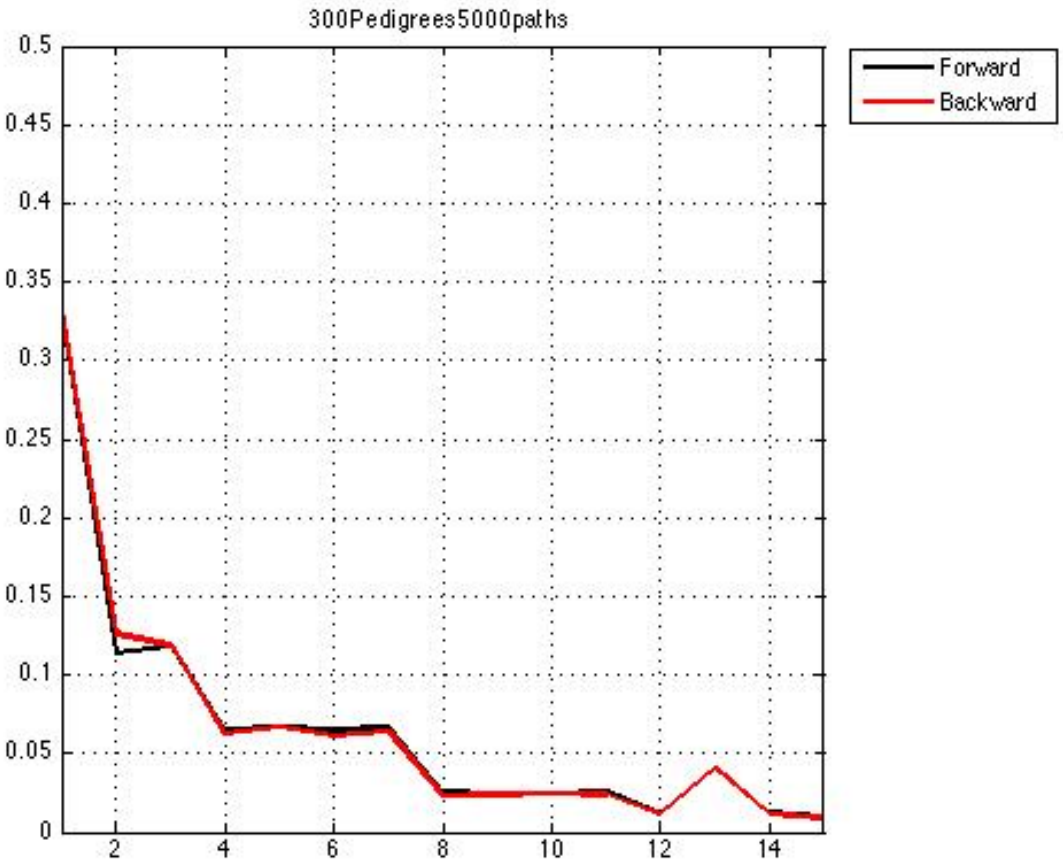


Figure 7: Probability distribution of identity states through Forward-in-Time Sampling and Backward-inTime Sampling: $G = 5, N = 5$, 300 pedigrees and each with 5000 inheritance paths

Compare the results of the forward and backward graph, we can observe that they almost have the same distribution that the probability of identity states is decreasing as the states are from the least related to the most related.

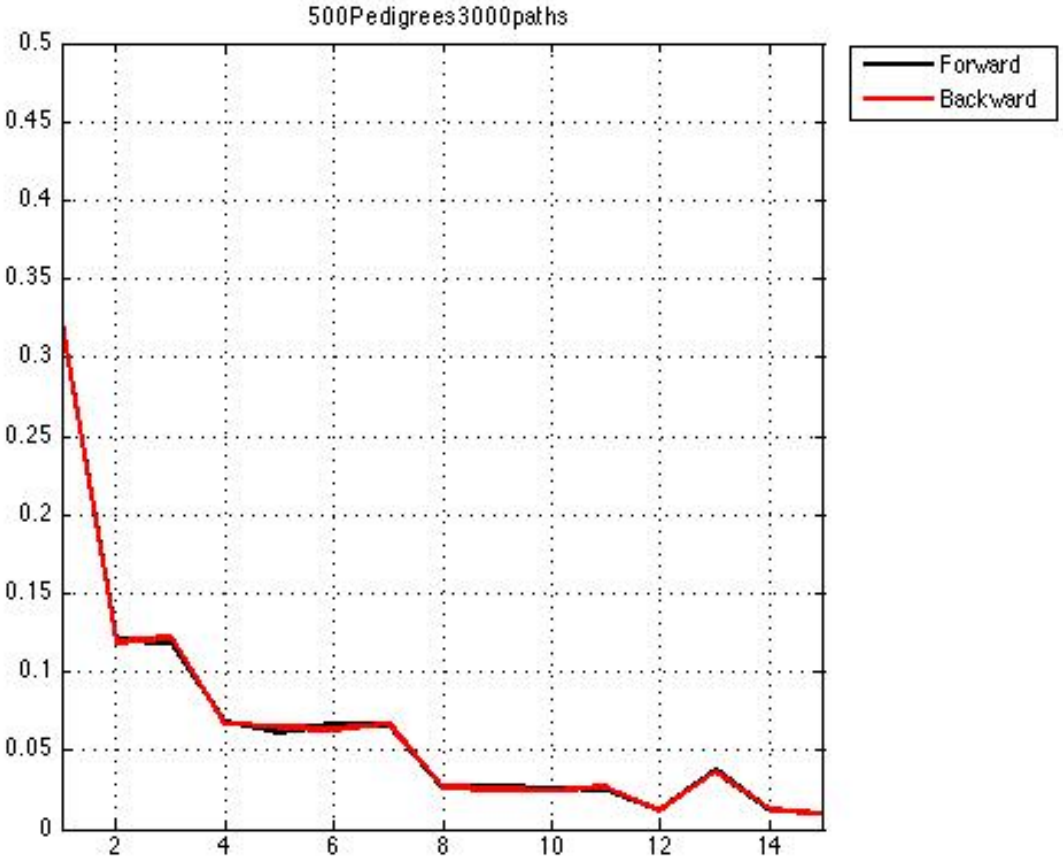


Figure 8: Probability distribution of identity states through Forward-in-Time Sampling and Backward-inTime Sampling: $G = 5, N = 5,500$ pedigrees and each with 3000 inheritance paths

3.4 Kinship Coefficients

Since we have defined the kinship coefficient in terms of an expectation over identity states, the result of probabilities of identity states that we got through forward-in-time and backward-in-time sampling can be used to compute kinship coefficients.

As we described in the introduction, we group identity states into inbreed and outbreed categories. Here we can define that each identity state, s, on the alleles of two individuals i, j has a vector $n^s = (n_1^s, n_2^s, n_3^s)$, where n_1^s is the number of outbreeding edges between i and j, n_2^s is the number of inbreeding edges in the individual i and n_3^s is the number of inbreeding edges in the individual j. Then use the probabilities of

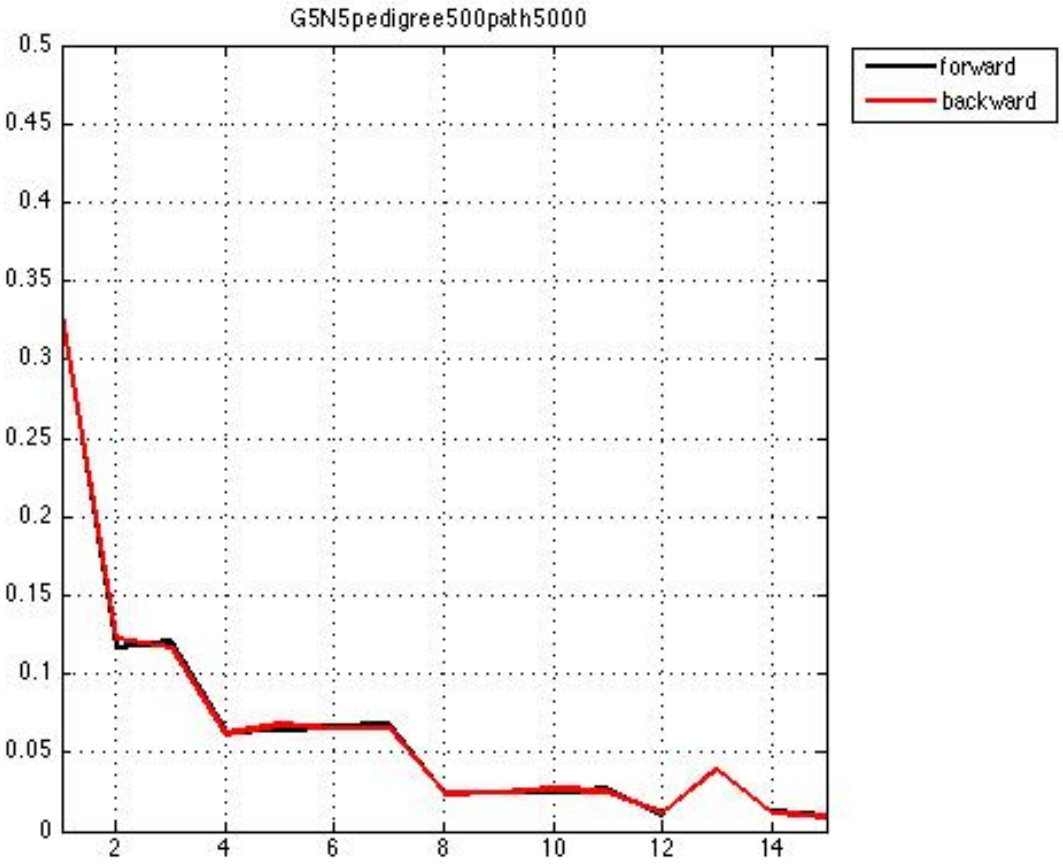


Figure 9: Probability distribution of identity states through Forward-in-Time Sampling and Backward-inTime Sampling: $G = 5, N = 5,500$ pedigrees and each with 5000 inheritance paths

identity states on selected individuals i, j to compute

$$\hat{\Phi}_{ij} = \sum_s \frac{n_1^s}{4} \hat{Pr}(s)$$

$$\hat{\Phi}_{ii} = (1 + \sum_s \frac{n_2^s}{4} \hat{Pr}(s))/2$$

$$\hat{\Phi}_{jj} = (1 + \sum_s \frac{n_3^s}{4} \hat{Pr}(s))/2$$

where we use notation $\hat{\Phi}$ since we use the probabilities through sampling.

In a case that it is too difficult to get probability distributions, people use a dynamic programming recursion to compute the kinship coefficient. The recursive algorithm for computing kinship is developed in detail in Pedigree Analysis in Human Genetics and in a 1981 paper by Karigl. The recursive equations are follows: for all founders f, the

matrix is initialized as

$$\Phi_{ff} = 1/2,$$
$$\Phi_{fj} = 0, \quad \text{for any individual } j \text{ that is not a descendant of } f.$$

Let the mother and after of i be denoted m and p. Then use the kinship coefficient between j and i's parents to compute

$$\Phi_{ij} = (\Phi_{mj} + \Phi_{pj})/2 \quad \text{where } i \text{ is not a ancestor of } j \text{ and } i \neq j$$

Similarly, we compute the inbreeding coefficient for i from the i's parents:

$$\Phi_{ii} = (1 + \Phi_{mp})/2$$

4 Summary

The theme of our review has been the unifying of the Kinship Coefficients Φ defined with dynamic programming recursion above and the estimate Kinship Coefficients $\hat{\Phi}$ through the simulation of identity states in a fixed pedigree. According to the results of forward and backward simulation, each pedigree with 3000 inheritance paths has been a reasonable assumption to get a stable probability distribution, which implies we can use this assumption to get our estimate kinship coefficients. We measure the rectilinear distance between these two relation coefficients and plot them as the change of the size in the pedigree. We can observe from these plots that there is a decreasing trend of the distance between the real kinship coefficients and the estimate coefficients when the size of the pedigree increases.

With the evidence that the small distance between the real kinship coefficients and the estimate kinship coefficients, kinship coefficients is illustrated to be accurate. Therefore besides the methods of GWAS and linkage studies, this method of kinship coefficients is also worth considering to analyze the disease association. Especially for a very large pedigree, the kinship coefficients would contribute a lot because of its efficient running time and relatively reliable results compared to other methods.

Acknowledgements

We would like to thank Professor Kay Kirkpatrick and Dr. Bonnie Kirkpatrick for directing our work and for providing background support. We thank Professor Bruce Reznick for suggestions. This research is supported by NSF Grant DMS-1106770.

References

[1] Donnelly K.P. The probability that related individuals share some section of the genome identical by descent. Theor Popul Biol 23(1983): 3463.

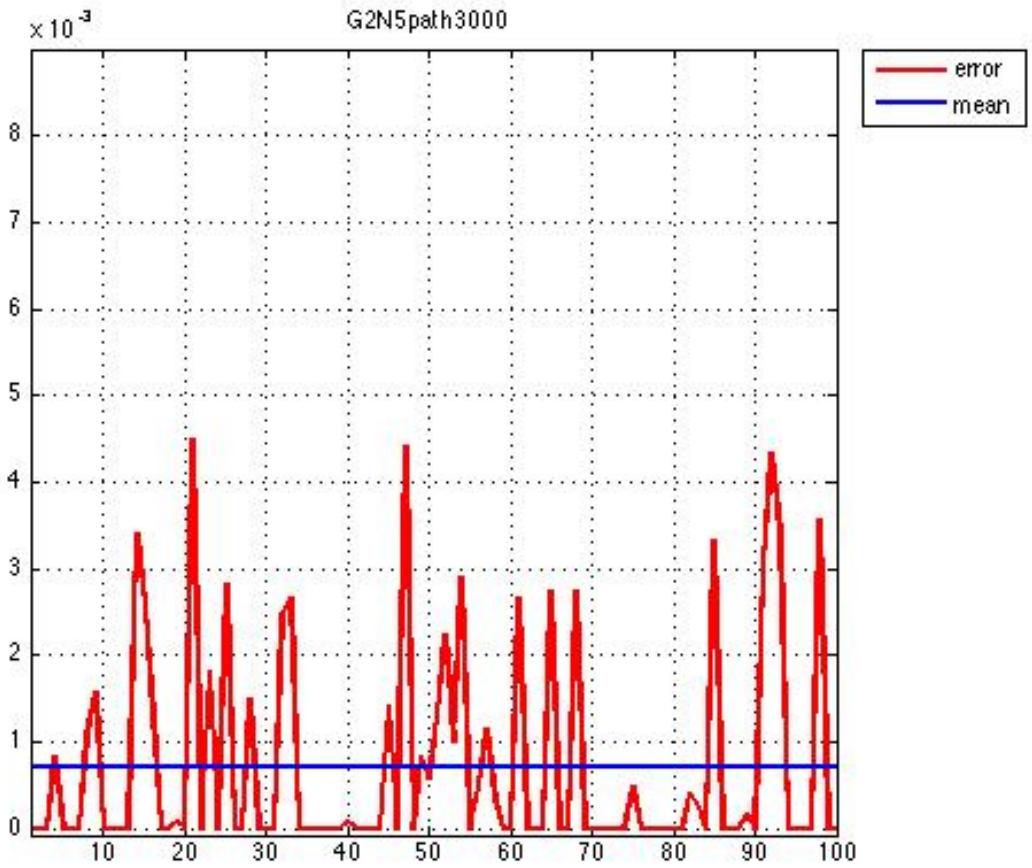


Figure 10: Rectilinear distance between the real kinship coefficients and the estimate kinship coefficients when $G = 2$, $N = 5$

[2] Eric Anderson An Example from Population Genetics: The Wright-Fisher Model. May 3, 1999.

[3] G.Karigl A recursive algorithm for the calculation of identity coefficients. Annals of Human Genetics, 45(3):299, 1981

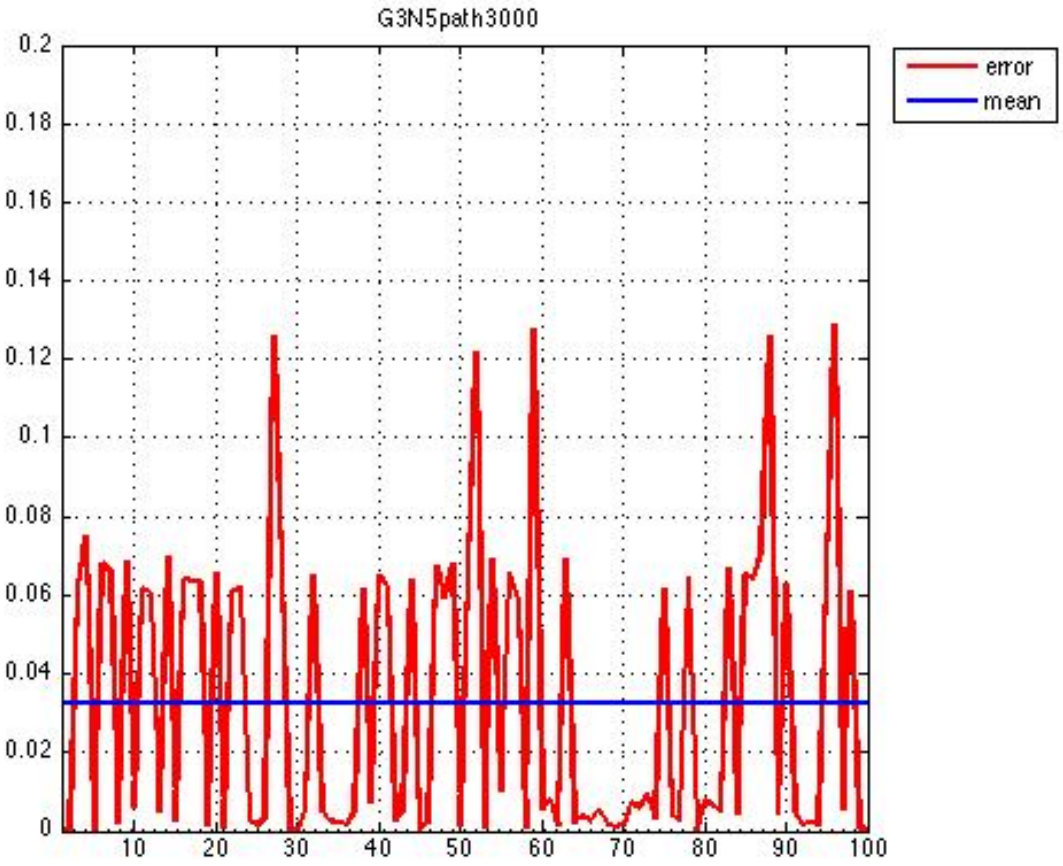


Figure 11: Rectilinear distance between the real kinship coefficients and the estimate kinship coefficients when $G = 3$, $N = 5$

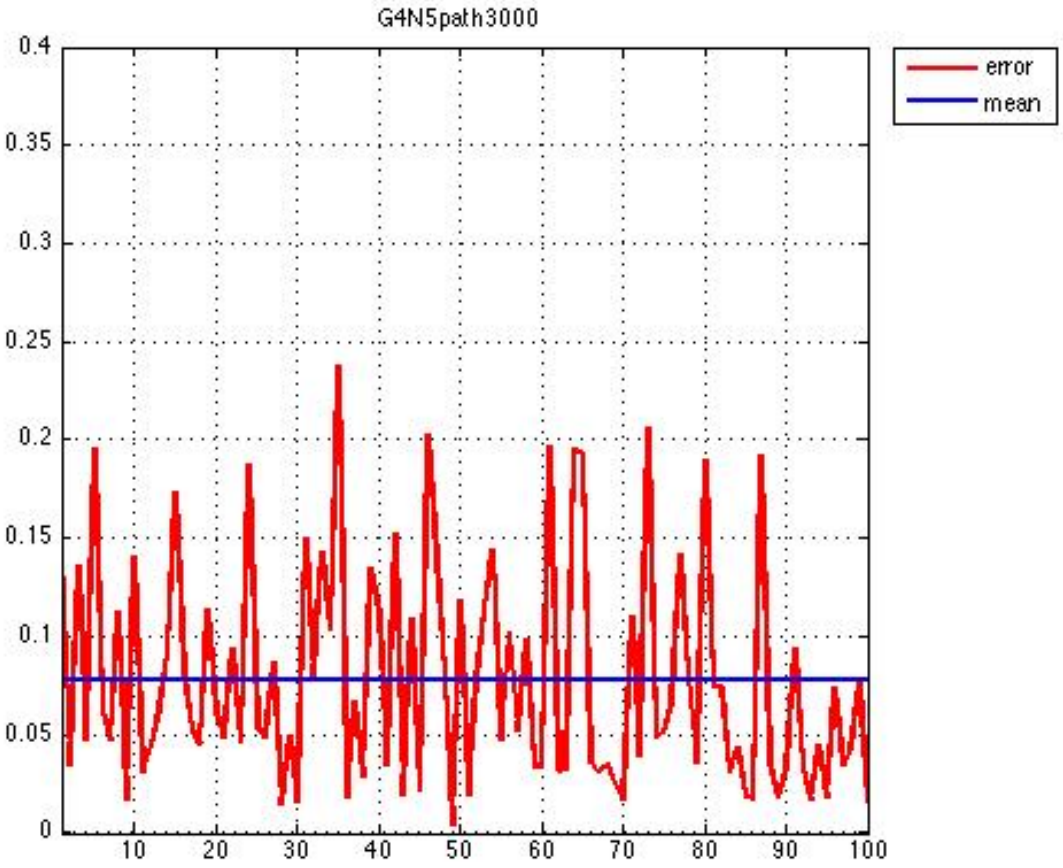


Figure 12: Rectilinear distance between the real kinship coefficients and the estimate kinship coefficients when $G = 4$, $N = 5$