Computational Methods for Analysis of Cryptic Recombination in the Performance of Genomic Recombination Detection Software

Jamal Alhiyafi¹, Cavitha Sabesan¹, Shiyong Lu¹, Melphine Harriott², Deborah Jurczyszyn², Raquel P. Ritchie², Felicitas S. Gonzales², and Jeffrey L. Ram²

¹ Department of Computer Science, Wayne State University
{alhiyafi, at0079, shiyong} @wayne.edu
² Department of Physiology, Wayne State University
{ai3143, ah8887, aj1073, ae4480} @wayne.edu and jeffram@med.wayne.edu

Abstract

The detection of recombination from DNA sequences is relevant to the understanding of evolutionary and molecular genetics. While programs such as GENECONV have been identified as detecting recombination more reliably than others, previous studies have not analyzed how many recombinations they fail to detect. We develop a method for testing how often such programs fail to identify recombinations and how detectability is affected by pairwise differences among the parental sequences. Recombination of sequences having a range of average pairwise differences (APD) is simulated by a stochastic method, and then the history of recombinations is compared to the recombinations identified by GENECONV. With high APD, GENECONV fails to detect ~50% of recombinations; while at a more typical intraspecies APD of 1% to 2%, >70% of recombinations are undetected. Quantitative results suggest corrections for estimating recombination rates more accurately and methods to detect evidence of recombination more consistently.

1. Introduction

The genomic era has produced a plethora of DNA sequences that provide opportunities for computational biologists to discover meaningful information about biological processes. DNA sequences have information with great relevance to the evolution of organisms; to ongoing genetic processes that mediate antibiotic resistance, genetic diseases, and adaptation; and to bioengineering applications for both beneficial (e.g., pharmaceuticals) and nefarious (e.g., bioterror) purposes. Among the most important processes that may play a role in all of these phenomena is horizontal gene transfer, in which a new section of DNA sequence appears in the DNA sequence of an organism, often from exogenous sources. Similarly, new combinations of DNA sequence are produced by sex, which causes an intentional mixing (or “recombination”) of genes from both parents to create a novel daughter sequence. As will be described, computational analysis of DNA sequences can sometimes detect where such “exceptional” insertions of DNA sequences have occurred.

DNA recombination can be categorized into two kinds of processes: homologous and heterogeneous recombination. Homologous recombination occurs between two homologous DNA molecules and can itself be divided into two kinds: gene conversion (replacement), in which one DNA donates part of its genetic information to another DNA (Fig. 1A), or crossing over (exchange), in which both parental DNAs exchange part of their genetic information (Fig. 1B). Gene conversion includes donating larger or smaller pieces of DNA to create a daughter sequence that has portions of sequence from both parental sequences. Gene conversion can occur due to a double-crossing-over event, as occurs during recombination in normal meiosis. Heterogeneous recombination occurs where a completely unrelated sequence is inserted into a sequence from a non-homologous region of DNA. The present paper focuses on the detection of gene conversion events.

A) Gene Conversion

<table>
<thead>
<tr>
<th>1st Parent: P1</th>
<th>2nd Parent: P2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Daughter: P2

B) Crossover

<table>
<thead>
<tr>
<th>1st Parent: P1</th>
<th>2nd Parent: P2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Daughter: P2

Figure 1. Different forms of Homologous Recombination.

The evidence for recombination as a potential mecha-
nism for genetic change leads computational biologists to develop statistical methods to compare existing DNA sequences to identify evidence and locations of prior recombination. Methods work by identifying discontinuities in sequence similarities or genetic distance (e.g., GENECONV [6, 7] and MAXCHI [4, 8]) or by phylogenetic methods that, for example, identify incongruous tree topologies (e.g., RECPARS [1]). Many methods have been evaluated [4, 5, 11] and new ones continue to be developed. In comparisons of recombination analysis software, GENECONV was among the most highly ranked [4] and is included in the RDP2 analysis suite [2].

How well do recombination detection programs work? Since recombination detection programs detect putative recombinant fragments based on identifying significant differences in sequence, and recombination occurs at least as frequently between similar sequences as between divergent ones, how frequently do programs, such as GENECONV, fail to identify recombinations that are known to have occurred? This paper tests the hypothesis that such programs fail to identify a significant number of recombination events and characterizes how the pairwise differences between the parental sequences affect the recombination detection success rate. Pairwise Difference (PD) is the total number of base pairs different between the alleles of a particular gene of two individuals. The method described here involves simulating recombination, so that an explicit history of recombination in a set of DNA sequences is known, and then testing the resultant DNA sequences with GENECONV to see how frequently it failed to detect recombination events known to have occurred to produce the simulated sequences. As a result, this paper identifies the occurrence of “cryptic recombination,” i.e., recombinant events that are known to have occurred but were not identified by a recombination detection program.

The main contributions of this paper are:

1. A program to simulate the DNA recombination process was developed. History of the recombinant events is saved for further comparison and analysis.
2. The known history of recombination occurring in the simulation was compared with the output of putative recombinations detected by a well-known highly ranked recombination detection program (GENECONV).
3. The results show that the recombination detection software fails to identify more than 50% of recombination events, designated by this paper as “cryptic recombinations.”
4. Analysis of factors affecting the frequency of cryptic recombination show that detectability is increased by pairwise differences between parental sequences, particularly of the non-transferred segment. Pairwise differences in parental sequences should be taken into account in estimating recombination rates and in designing recombination detection experiments.

Organization. Section 2 describes the stochastic simulation method used to create populations of DNA sequences of known recombination history. Section 3 presents biological data on bacterial DNA sequences that provide a justification for values of various parameters used in testing the model. Section 4 describes the application of GENECONV to the simulated data and how its output was compared to the recombinations known to be present in the input sequences. Section 5 shows the results of simulation experiments. Section 6 discusses the results and points the direction for future work.

2. Simulation Method

An allele is one of the variant forms of a gene sequence. Recombination rate (RR) is the probability that the alleles of two arbitrary individuals will recombine. In the described experiments, RR was chosen to have a fixed value (RR\text{fixed}) that resulted in the number of alleles surviving after 1500 generations being close to the initial number of alleles at the start of the simulation. In the Discussion (section 6), variable rates of recombination are also considered.

2.1. Simulation Model

Our Recombination Simulation Program (RSP) simulates recombination, replication, and selection through many generations. One of the key features of this algorithm is that a history of all of the recombination events that occurred in each generation is recorded.

Algorithm CreateInitialPopulation, given in Fig. 2, corresponds to the first task in RSP. In line 10, the first initial allele of length L is randomly generated and added to the list of alleles. L is the length of DNA sequence for which recombinations are being simulated. In line 12, another random allele which differs on average from the first allele by the Average Pairwise Difference (APD) is generated and added to the list of initial alleles. APD is the average of all PDs in a population of genes. After adding the most recent allele to the list of initial alleles, an interim APD (called mAPD) of the new list is calculated. If the difference between the targeted APD and the APD among the existing alleles (mAPD) is within a certain range then the counter of alleles is incremented and the program proceeds to line 12; otherwise, the program jumps to line 15 where the last allele generated is dropped and then the program returns to line 12 without incrementing the counter of alleles. In line 18, N/K copies of each allele are duplicated, where N is the number of individuals in the population and K is the final
number of alleles. In this way, an initial population is generated for which the average pairwise difference, APD, for all genes is approximately equal to the target value given at the beginning of the program (line 3). Since there is a random element in the generation of the initial populations, the actual APD of these initial populations is only approximately equal to the target APD values of 1%, 2%, 5%, and 10%, as described in the results section.

01 Algorithm CreateInitialPopulation
02 Input:
03 APD: targeted average pairwise difference
04 K: number of alleles
05 N: number of individuals in the population
06 L: length of an allele
07 Output:
08 The initial population that contains N individuals and K alleles
09 Begin
10 Generate a random allele P[1]
11 For k=2 to K allele do
12 P[k] = generate a random allele where it differs than P[1] by APD
13 mAPD = the average of PD in P[k]
14 If \[ mAPD - APD \leq \epsilon \] Then
15 drop the last allele created without incrementing k
16 End If
17 End For
18 Duplicate N/K copies of each allele
19 End Algorithm

**Figure 2. Algorithm CreateInitialPopulation**

In M generations, recombination events, replication, and selection processes occur. Algorithm Recombination given in Fig. 3 corresponds to the recombination process during each generation. In line 11, the first potential partner gene \( g_i \) of the recombination is chosen and checked to see if it has not previously recombined in that generation. If gene \( g_i \) is in the recombination set, then another \( g_j \) is chosen; otherwise, in line 14, the second partner gene \( g_j \) is randomly chosen from the partner set. In line 15, the recombination rate for these two genes is calculated. In the case illustrated, a fixed recombination rate \( RR_{fixed} \) is used for all recombination events (an alternative program that uses variable recombination rates modifies this step). In line 16, a random floating point number between 0 and 1 is compared to the recombination rate, \( RR_{fixed} \). If this random number is less than or equal to the recombination rate then both genes \( g_i \) and \( g_j \) will recombine. If gene \( g_i \) does not recombine with \( g_j \), the program jumps to line 23 where \( g_j \) is removed from the partner set and the program returns to line 14 to choose another gene \( g_j \) from the partner set.

This iteration continues until a recombination has occurred or until all genes in the partner set have been given a chance to recombine. If recombination occurs between \( g_i \) and \( g_j \), the DNA fragment that is moved from one sequence to another for recombination has a length \( RL_i \), and the following steps occur: in line 17 the starting position of the gene where recombination will happen is randomly chosen. The starting position is a random integer between -(RL) (recombination length) and (L) (length of the sequence).

01 Algorithm Recombination
02 Input:
03 population[N], N number of individuals
04 RL: recombination length
05 L: length of individual
06 Output:
07 population [N] after recombination
08 Begin
09 RecombinationSet = \( \phi \) (record individuals that have engaged in a recombination process
10 For \( i = 0 \) to N do
11 If \( g_i \in \text{RecombinedSet} \) then continue;
12 PartnerSet = genes - \( g_i \) - RecombinedSet \( \cup \) candidate partner set
13 While PartnerSet \( \neq \phi \) do
14 choose a random gene \( g_j \) from PartnerSet
15 \( RR_{g_i, g_j} = RR_{fixed} \)
16 If randf(0.0,1.0) \( \leq RR_{g_i, g_j} \) then
17 startpos = randi(-RL, L)
18 endpos = min(L-1, startpos+RL)
19 Replace substring \( g_j \) \( \text{[startpos, endpos]} \) by \( g_i \) \( \text{[startpos, endpos]} \)
20 RecombinedSet = RecombinedSet \( \cup \) \( \{g_i, g_j\} \)
21 break
22 Else
23 PartnerSet = PartnerSet - \( \{g_j\} \)
24 continue;
25 End If
26 End While
27 End For
28 End Algorithm

**Figure 3. Algorithm Recombination**

In line 18, the end position of the recombinant fragment is calculated as min \((L-1, \text{startpos}+\text{RL})\). In line 19, the substring \( g_j \) \( \text{[startpos, endpos]} \) is replaced by the substring \( g_i \) \( \text{[startpos, endpos]} \). In line 20, both genes are added to the recombination set for that generation.

After all recombinations for that generation are simulated, the resultant population is replicated, simulating potential growth to a population of size \( 2 \times N \). To keep population size constant, as would be the case for a stable biological population, N individuals are randomly chosen from the total \( 2 \times N \) population for the next generation.

The previous process of recombination, replication, and selection counts as one generation and is repeated for \( M \) generations. A history of all recombination events is recorded for future comparisons and analysis.

Once the process of simulating DNA recombination is finished, the system produces a FASTA file containing the surviving alleles. This FASTA file is one of the inputs for the GENECONV detection program.

3. DNA Sequence Data

The choices of L and RL and the range of APD used for the simulations were based on biological data from *Escherichia coli*. For the gene for beta-glucuronidase, 525 bases (bases 331 to 855) were sequenced in 1323 strains of *E. coli* 148 alleles were identified, of which 76 occurred in the population at least twice. The APD was 10.5 bases, or approximately 2%. GENECONV identified significant recombination in the 76 multiply occurring alleles, identify-
ing an average gene conversion fragment size of 243 bp. For the fimH gene, 531 bases (bases 80 to 610) were sequenced in 52 strains known to have different beta-glucuronidase alleles. 40 alleles were identified. The APD was 11.0 bases, or approximately 2.1%. GENECONV identified significant recombination among the 40 alleles, identifying an average gene conversion fragment size of 241 bp.

E. coli sequence data has also been analyzed with GENECONV for ectopic recombination, for which it is hypothesized that intragenomic recombination takes place between genes coding for similar proteins identified by BLASTCLUST [3]. We repeated the analysis to identify additional variables not previously reported. For 4 different E. coli genomes (U00096, AE014075, BA000007, AE005174), the average lengths of aligned DNA analyzed by GENECONV sequences, weighted according to the number of sequences in each cluster, were 965, 828, 871, and 919, respectively. The APD within each cluster, weighted according to the number of sequences in the clusters, averaged 13%, 14%, 10%, and 10%, respectively. The average lengths of gene conversion fragments identified by GENECONV in these clusters were 440, 198, 271, and 315, for an overall average of approximately 300 bp.

Populations simulated in this study had APD values ranging from 1% to 10%, which overlaps with the APD range actually observed and analyzed with GENECONV in the above biological experiments. Similarly, sequence lengths, L, were 531 or 1,000 bp, which overlaps with the biological data, and the models simulated gene conversion fragment lengths of 261 and 300 bp, also in the range of the biological data.

4. GENECONV Detection Method

Given an alignment of DNA sequences, GENECONV finds the most likely candidates for aligned gene conversion events between pairs of sequences in the alignment. The FASTA file of DNA allele sequences present at generation 1500 of each simulation was submitted to GENECONV, which generated lists of pairs of sequences identified as having putative gene conversions. To determine whether the pairwise inner fragments identified by GENECONV were statistically significant, the number of pairwise inner fragments generated from the submitted population was compared to the number of pairwise inner fragments generated with the “randomize_Sites” parameter turned on. The “randomize_Sites” parameter generates the number of pairs that would have been observed to have gene conversions just by chance with the given number of polymorphisms in the submitted population. Since the number of pairwise fragments is expected to be an approximately Poisson distribution, significance was determined from the Z value.

To identify cryptic recombination, the saved history of recombination events generated by the simulation was compared to the list of putative gene conversions identified by GENECONV. Since multiple recombination events could obscure the recombinant products of earlier generations, in this paper, only the recombination products known to have been generated in the last ten generations of the simulation were compared to the GENECONV output list. The effects of pairwise differences of the parent sequences, the start position of the recombed fragment, the pairwise difference of the recombed fragments from the replaced fragments, and the pairwise difference of the non-transferred segments of the parent sequences were analyzed for their effect on the detectability of the resultant gene conversion events by GENECONV.

5. Experimental Results

Two sets of parameters were used in the simulation. One set used length of sequence (L) equal to 531 bp with 263 bp for the recombination length (RL). The other set used 1000 bp for the length of sequence (L) with 300 bp for the recombination length. The recombination rate was fixed and equal to 3.2e-6. Populations of size N = 1000 individuals were simulated. Although all simulations began with 40 alleles, the final number of alleles after 1500 generations changed after 1500 generations, increasing or decreasing depending on the initial APD. With 1%, 2%, 5%, and 10% APD, the average number of alleles after 1500 generations was 28.9, 40.4, 45.6, and 49.7 respectively.

The APD of initial populations created with target APD values of 1%, 2%, 5%, and 10% are illustrated in Fig. 4. After 1500 generations, the APD of the final population decreases to about 50% of the initial population APD.

The number of recombinations occurring in 1500 generations with the above mentioned parameters was approximately 6000 recombination events, an average of 4 recombinations/generation. In the last 10 generations that were compared with GENECONV output, 35-50 recombination events occurred.
5.1. GENECONV Results

After analysis of the final set of alleles by GENECONV, Z values were always much larger than 2.5, indicating that the number of gene conversion fragments identified in the submitted data was significantly greater than would have occurred by chance. For example, the value of Z for 49 alleles resulting from running the simulation for 1500 generations with \( L = 1000 \) and \( RL = 300 \) bp, was 32.42. Another run of GENECONV with 44 alleles resulted in \( Z = 25.05 \).

![Figure 5. Effect of APD % on recombinations detected by GENECONV.](image)

If GENECONV is ideally sensitive, then all recombinations known to have occurred in the last 10 generations should have been identified; however, the program often failed to identify the majority of them. For example, with 2% APD, in the last 10 generations of one run, 41 recombinations occurred and 49 alleles were given to GENECONV. GENECONV detected 257 recombinations, but only 11 of the 41 that actually occurred in the last 10 generations of the simulation process. In another example, where 5% APD was used and 52 alleles were submitted to GENECONV, 39 recombination events occurred in the last 10 generations, and GENECONV detected only 22 of them. Fig. 5A shows that low parental PD reduces recombination detectability. For the last 10 generations of the simulation, the total number detected for various PD were divided by the total number of occurrences in the 10 runs. Figs. 5B, C summarize how APD affects whether GENECONV detects known recombinations at different APDs. We used ANOVA in Figs. 5B & C where \( p < 0.001 \).

The number of non-detected recombinations, i.e., “cryptic recombinations,” decreases as the APD increases and is a more significant problem when APD is low. Subsequent experiments analyzed whether the percent detected could be related to specific characteristics of the recombination event, such as the position of the recombined fragment and the pairwise difference present in the transferred or non-transferred fragments.

Fig. 6 shows that detectability was virtually the same regardless of the position of the recombined segment. The example shown is for sequences analyzed from simulations with initial APD of 2%. Populations with other APD values also failed to show any consistent differences with fragment position, showing the same differences of detectability at all positions as the overall detectability varied between different APD groups (Fig. 4).

![Figure 6. Lack of effect of position of recombinant fragment on detectability.](image)

Analysis of whether the pairwise difference of the recombined fragment affected its detectability also showed no difference across a wide range of values. However, as illustrated in Fig. 7, the pairwise difference of the non-transferred segment has a significant influence on the detectability of the recombination event. For 10 runs of each initial APD %, the total number of recombinations detected in the recombination history of the last 10 generations was divided by the total number of recombinations in the same period for various pairwise differences of the non-transferred segments. At 1% and 2% parental APD, the percent detected increased over the entire range of the pairwise differences of the non-transferred segments, and for 5% and 10% parental APD percent detected increased to a plateau of around 60% for pairwise differences of the non-transferred segment greater than approximately 2.5% of the non-transferred segment.

6. Discussion and Future Work

DNA recombination detection systems underestimate the amount of recombination since they are unable to detect the most frequent recombinations that occur between similar sequences. For DNA sequences with average pairwise differences of 1% to 2%, more than 70% of recombinations are known to have occurred failed to be detected. These figures are based only on comparing sequences from the most recent 10 generations of simulation; however, even considering the entire 1500 generations, the total set of gene conversions identified by GENECONV is often much less than the 6,000 such events known to have been simulated in 1500 generations; viz., the representative example cited below Fig. 5. Potentially, by knowing the APD of a population...
being analyzed, the number of cryptic recombinations can be estimated from the quantitative analysis provided here. These results would yield a higher recombination rate than previous analyses have suggested.

![Figure 7. Effect of pairwise differences of the non-transferred segment of parental sequences on detection of recombinations by GENECONV.](image)

The causes of the low detectability of some recombinations does not seem to be related to the position of the recombined piece (Fig. 6) nor the number of bases different in the transferred piece of DNA. Since the pairwise difference of the parents appears to affect detectability (Fig. 5) but the pairwise difference of the transferred fragment does not, then a key variable mediating detectability may be the pairwise difference of the non-transferred segment, an observation supported by results illustrated in Fig. 7. Future studies should further investigate the causes of non-detectability in order to reduce this problem or estimate its magnitude in future analyses.

The present study demonstrates that the population APD is a key variable in affecting recombination detectability. Pairwise differences could also affect quantitative estimates of recombination in another way: biological studies have determined that recombination is more likely to occur between similar sequences (i.e., those with a lower pairwise difference) than between more distant sequences [9, 10]. Thus, the recombinations that are hardest to detect, according to the present study, also occur most frequently. The stochastic simulation model developed in the present model is an ideal platform for analyzing the impact of pairwise differences on estimates of recombination rates. Preliminary results with simulations in which recombination rates decrease with increasing pairwise differences indicate that an even higher proportion of recombinations fail to be detected than is demonstrated here using a fixed rate of recombination.

As further evidence of the importance of taking cryptic recombination into account, we have begun to examine critically the phenomenon of ectopic recombination in bacterial genomes, as described by Morris and Drouin [3]. While corroborating differences in GENECONV detected gene conversion between pathogenic and non-pathogenic strains of *E. coli*, modifications of the analysis procedures have revealed up to three times more putative recombinant fragments than previously described (unpublished results). The revelation of more recombination confirms that a significant amount of cryptic recombination was present in the previous analysis, and provides further support from biological data for the concept of cryptic recombination.

In future work on detecting gene conversion, realization of the concept of “cryptic recombination” may make possible more realistic estimates of rates of recombination based on the number of gene conversion pairs detected and a knowledge of the APD of the population being analyzed. In addition, these results suggest the need for developing more sensitive recombination detection algorithms that take into account the pairwise differences of parental sequences.

References