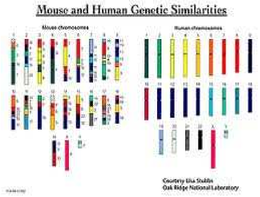
Genome Rearrangements with Insertions and Deletions

Sam Brueggen and Kyle Vogt
Faculty Mentor: Manda Riehl
University of Wisconsin-Eau Claire

1. Genes as Permutations

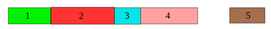
Many mathematical models have been developed to help solve the genome rearrangement problem, whose goal is to find the shortest sequence of mutations for one genome into another. However, few of these consider small segments of DNA in which insertions and deletions are the primary mutations that occur. When developing this model, we chose to consider Neurofibromatosis (NF), which causes uncontrollable neuron growth leading to tumors, blindness, and learning disabilities. NF is one of several autosomal dominant disorders caused primarily by insertion and deletion mutations during DNA replication.
We began by envisioning a gene sequence as a number permutation, where each number represents a genome fragment. We can see that the assumption that a genome is composed of conserved segments (each represented by a number) is justified by looking at the differences between mouse and human genomes. Though the number of chromosomes and arrangement of genetic material is different, most of the information is the same, just presented in a different order.



2. Definitions

Given a genome segment, $a_1, a_2, ...a_n$, and another genome segment, $b_1, ..., b_m$, we define an *insertion* to be the operation yielding the single genome segment obtained by inserting $b_1, ..., b_m$ directly after gene a_i, for $i = 0$ to n. What this means is that a deleted segment of DNA is inserted between two genes of any other segment.

Here a segment of DNA before an insertion:



Here is the resulting segment inserting the 5 segment between the 2 and the 3 segment.

Given a genome segment, $A = a_1, a_2, ...a_n$, we define a *deletion* to be the operation yielding two genome segments obtained by deleting $a_i, ..., a_j$ from A with $i \leq j$, for i, j from 1 to n. This means that a consecutive portion of DNA is removed from a segment.

Here is a segment of DNA before a deletion:

Here is the resulting segment after deleting the 23 segment from the initial segment:

3. An Insertion-Deletion Mutation Example

Here is an example of how a person might try to find the distance:
Suppose we have the genome, 0123456, and we want to know the distance to the mutation, 0642 13 5.

0123456 → 0642 13 5

	0456	123	(delete the 123 segment)
	0456	2 13	(delete the 2)
045	6	2 13	(delete the 6)
	0645	2 13	(reinsert the 6)
064	5	2 13	(delete the 5)
	0642	13 5	(reinsert the 2)

Thus we conjecture that the distance is 6. However, we can't be sure that the sequence of mutations above is optimal without an exhaustive search. Therefore a rigorous algorithmic approach to finding the distance is necessary.

4. Theorem: Distance Formula

The distance, defined as the optimal sequence of mutations for the transformation of one genome into another, can be determined as follows:

$$d(\iota, \sigma) = (\delta - 1) + 2(n - (\delta - 1) - t)$$

Where the first term, $(\delta - 1)$ is the number of genes that take just one move to correctly position and $2(n - (\delta - 1) - t)$ is the number of genes costing exactly 2 moves to position. This is determined by subtracting the number of genes that take just one move to correctly position, and t, the number of genes that do not need to change position, from n, which is the total number of genes in ι and also in σ.
So,

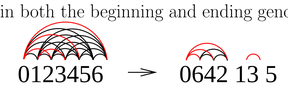
$$\begin{aligned} d(\iota, \sigma) &= \delta - 1 + 2n - 2\delta + 2 - 2t \\ &= \delta - 1 + 2n - 2\delta + 2 - 2t \\ &= 2(n - t) + 1 - \delta \\ &= 2(n - t) - (\delta - 1) \end{aligned}$$

5. Example

Consider the mutation

0123456 → 0642 13 5

In order to find t, we'll identify possible time saving relationships by drawing an arc diagram. Each arc represents an ordered pair of genes, where the order is left to right. The red arcs indicate that the ordered pair is in both the beginning and ending genomes.



Now, we can write an nxn matrix where n is the number of possible time saving relationships. The purpose of the matrix is to identify pairs of possible time savers that do not cross arcs in either permutation. If this is the case, the entry is a 1, if not it is 0.

	0,2	0,4	0,6	1,3
0,2	1	0	0	0
0,4	0	1	0	1
0,6	0	0	1	1
1,3	1	1	1	1

Then, the size of the largest submatrix of 1's is t. So, $n = 6$, $t = 2$, and $\delta_\sigma = 3$. And

$$d(\iota, \sigma) = 2(n - t) - (\delta - 1) = 2(6 - 2) - (3 - 1) = 6.$$

6. Future Research

The next step in our research concerning genomes will be to find an equation for the distance between two permutations in the general case. The only difference between the general case and the special case, for which we've already determined the distance, is that the general case allows for breaks in beginning genome. A similar line of reasoning to determine the distance leads to:

$$d = 2(n - t) - (\delta_\iota - 1) - (\delta_\sigma - 1)$$

where $2(n - t)$ is the number of genes that need two operations performed on them to be placed correctly in the destination permutation, $\delta_\iota - 1$ and $\delta_\sigma - 1$ are defined as the number of *breaks* in the start and end genome. This equation works well almost every time, but fails in one particular case. In this case, the algorithm for finding t fails, rendering the equation useless. We take it as tautologous that the distance can be defined thusly: $d = 2(n - s) - \Delta\delta$ where s is defined as the number of genes that do not need two operations performed on them to be placed correctly, and $\Delta\delta$ is defined as the change in the number of chunks, $\delta_\sigma - \delta_\iota$. Unfortunately we have not yet been able to come up with an algorithm for finding s. We would also like to find the distribution of genes that can be arrived at by performing x mutations on any genome.

References

[1] E. Ars, E. Serra, J. Garcia, H. Kruyer, A. Gaona, C. Lazaro, X. Estivill, *Mutations Affecting mRNA Splicing are the Most Common Molecular Defects in Patients with Neurofibromatosis Type 1*, Sém. Human Molecular Genetics 9, (2000) 237-247.

[2] A. Bergeron, J. Mixtacki, and J. Stoye, *A Unifying View of Genome Rearrangements*, WABI 2006, (2006) 163-173.

[3] G. Fertin, A. Labarre, I. Rusu, . Tannier, and S. Vialette, *Combinatorics of Genome Rearrangements*, Sém. MIT Press, Cambridge, Massachusettes, (2009).

[4] S. Yancopoulos, O. Attie, and R. Friedberg, *Efficient Sorting of Genomic Permutations by Translocation, Inversion and Block Interchange*, Sém. Bioinformatics 21, (2005) 3340-3346.

Acknowledgments

- Office of Research and Sponsored Programs, UW-Eau Claire
- All programs written and executed using Maple and Python.
- Images created and rendered in LaTeX.