**Evolutionary relationships in the *drosophila-*genus;  
a computational analysis for *d. melanogaster* and *d. miranda***

**Introduction**

*Drosophila* is a genus of small flies, most commonly referred to as fruit flies due to the habit of many of its species to circle around rotting fruit. The entire genus contains over 1,500 distinct species with a wide variety in (breeding) behaviour and appearance. One species of this genus, *D. melanogaster*, is used as a model organism in developmental biology and genetics, resulting in an abundance of knowledge about the genus and this species in particular. The entire genus, however, also holds valuable information for evolutionary biology and speciation. The genus can be found on all continents (*source*) and its high number of species and diversity of characteristics in combination with a relatively small genome which is easy to sequence and the available knowledge about breeding the species makes the genus a great candidate for evolutionary studies. This study will analyse the genomes of two related *Drosophila* species, namely: *D. melanogaster* and *D. miranda* by developing a number of algorithms suitable for solving this specific problem*.*

Based on genetics, appearance and behaviour, the *Drosophila-*genus has been divided into two species groups; the *Drosophila melanogaster* and the *Drosophila obscura*. Info about the two species and the groups in which they belong: obscura and melanogaster (?). These two species are genetically alike, since their genetic information is exactly the same, the only difference is that their genes are in a different order.

Evolutionary biology uses several methods, one of which is referred to as maximum parsimony. This method, used for deciding on the evolutionary relationship between species in a phylogenetic tree, focusses on finding the smallest number of evolutionary events to explain the genetic data. Another, more advanced method, is called Bayesian interference. This method relies on assumptions which are made beforehand based on the likelihood of certain mutations or genetic flips to occur, resulting in a more reliable phylogenetic tree.

This study will focus on the algorithms and computation needed to unravel the evolutionary relationship between these two species. As mentioned before the genomes of the two species are really alike and exist of the same genes, but in a different order. Therefore, we will simplify our model into a series of 25 numbers in a different order, as shown in figure 1.1. Furthermore, our model will also be a simplification of the real biology by not taking epigenetic marking, which is found in all eukaryotes (source), into account. In this study we will analyse several evolutionary pathways, first of all we will try to find the pathway consisting of the lowest number of total flips, secondly our aim is to find the pathway with the lowest mutation score. The mutation score is based on the length of a flip, in which a longer flip results in a higher score.

**Mutation scores nog uitleggen….**

**Methods**

**Flip Sorter**

To identify the upper bound of the problem described in the section above, we implemented a basic flip-sorter. This sorter sorts the sequence by moving the small numbers to the beginning of the sequence, in a way comparable to selection sort, but instead of making swaps using flips. Figure 2 shows how this algorithm has sorted the genetic sequence of *D. melanogaster* to the sequence of *D. miranda*. The following results, defined as the upper bound, were obtained through this algorithm, in which n is the length of a flip:

* Maximal number of flips: 18
* Total sum of n: 147
* ½n2: 963.5

**Depth First Search**

After defining our upper bounds, we implemented an algorithm known as a Depth First Search (DFS). DFS algorithms are always constructive searching algorithms, which work by DFS NOG FF UITLEGGEN

Prior to running the algorithm, we define all possible mutations for the sequence of interest. When running the algorithm, we keep track of the mutations through a mutation tracker. This tracker assures that we are always able to retrieve information about all previous flips. Furthermore, the combination of defining all possible mutations and keeping track of them as well allows us to make sure that we never try the same mutation on a genome and helps saving time as the process continues because we are not randomly assigning mutations so we never have to test if our mutation is unique for a genome.

Based on the upper bound defined by our flip sorter, all branches are pruned at length 18. Unfortunately, this algorithm does not function for genome lengths of 25 due to a shortage in memory. To be able to run this algorithm for the required genome length, we implemented an archive. This archive keeps track of all sequences it encounters and the level on which it encounters this sequence. If the algorithm finds a sequence which is already in the archive on a level lower or equal to the level where it encounters the sequence at this point, the branch is pruned. We used this algorithm to find the minimum numbers of flips needed, but also to find the lowest possible solution to the mutation formulas. The results are as follow, in which n is the length of a flip:

* Maximal number of flips:
* Total sum of n:
* ½n2:

Fig 1.1. Genomes sequences of Drosophila Melanogaster and Drosophila Miranda.

**State Space**

**Genome**The 25 genes of the *Drosophila* can be put in any random order. The first place can have any of the 25 genes, the second place can be any of the 24 remaining genes, third place can be any of 23 remaining genes et cetera. So the number of possible states of the genome is 25\*24\*23\*21… \*2\*1 = 25!:

25! = 15511210043330985984000000 = 1.551 \* or 15 septillion possible ways to order the genome of the *Drosophila.*

**Mutations**  
The genome can change by flipping a part of the genome. The length of the flip can be 2 to 25 genes long on gene 1, but if the flip starts at gene 2 the flip can only be 24 genes long at the most. A flip at gene 24 can only switch around gene 24 and 25, and a flip starting at gene 25 can’t happen. So the number of possible different flips in the genome is 24+23+22+ … +2+1 = 300.

When the genome has changed twice there is only a one in 90 000 chance the second change reversed the first change.

**References**

https://en.wikipedia.org/wiki/Computational\_phylogenetics#Maximum\_parsimony

Markov chain Monte Carlo sampling algorithms

The Sankoff-Morel-Cedergren algorithm