**Fruitvliegen**

**Het achterhalen van het evolutietraject tussen twee fruitvliegsoorten   
met behulp van het breakpoint algoritme**

**Team VliegFruit**

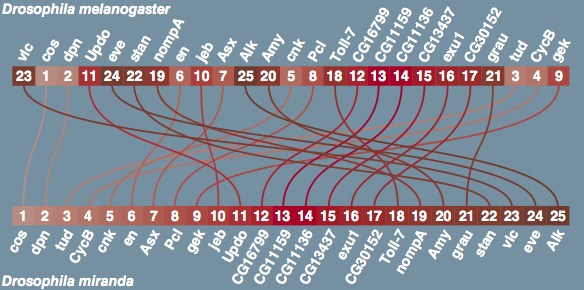
Iris de Vries Mark Landkroon Victor den Haan

Universiteit van Amsterdam Hogeschool van Amsterdam Universiteit van Amsterdam

[Iris.devries@student.uva.nl](mailto:Iris.devries@student.uva.nl) [Mark.landkroon@hva.nl](mailto:Mark.landkroon@hva.nl) [vdenhaan@gmail.com](mailto:vdenhaan@gmail.com)

**1. Inleiding**

De Drosophila melanogaster en de Drosophila miranda zijn fruitvliegsoorten waarvan het genoom goed bestudeerd is (Bachtrog &Andolfatto, 2006; Hoskins et al., 2015). De genen van de twee fruitvliegen zijn identiek, maar staan in een andere volgorde (afbeelding 1). Door middel van mutaties kan het ene genoom in het andere veranderen. Deze mutaties zijn bij de Drosophila beperkt tot het in zijn geheel omkeren van de subdelen van het genoom, ook wel inversies genoemd (afbeelding 2). Als er generaties lang mutaties plaatsvinden, bestaat de mogelijkheid dat het genoom exact zo veranderd dat de volgorde van de genen identiek is aan die van een andere fruitvliegsoort.



*Afbeelding 1*. De genomen van de Drosophila melanogaster en de Drosophila miranda en hun relatie.



*Afbeelding 2*. Mutaties in het genoom van de Drosophila zijn beperkt tot het in zijn geheel omkeren van subdelen van het genoom. Dat wil zeggen dat het niet mogelijk is om enkele genen om te wisselen zonder dat daarbij de hele subreeks wordt omgedraaid. In bovenstaande afbeelding is een mutatie te zien van lengte 5, waarbij de getallen 6, 10, 7, 25 en 20 worden omgedraaid.

In dit artikel wordt besproken welke achtereenvolgende mutaties verantwoordelijk zijn voor het veranderen van het ene genoom in het andere. Er zijn vaak meerdere manieren waarop dit mogelijk is, dit wordt duidelijk bij het “muteren” van onderstaande getallenreeks[[1]](#footnote-1) (afbeelding 3):

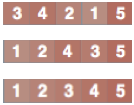


*Afbeelding 3.* Reeks 1



*Afbeelding 4.* Reeks 2

Om deze reeks om te zetten naar reeks 2, is het mogelijk de volgende stappen te doorlopen:



*Afbeelding 5.* Klein aantal stappen tussen reeks 1 en 2.

Reeks 1 kan worden omgezet in reeks 2 in twee stappen. Echter, dit is niet de enige manier waarop dat mogelijk is. Het doorlopen van de volgende stappen geeft hetzelfde resultaat:

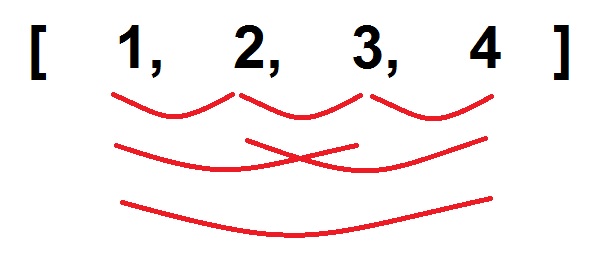


*Afbeelding 6.* Groot aantal stappen tussen reeks 1 en 2.

Dit suggereert dat er een “beste pad” bestaat, of misschien wel meerdere. Dit zou dan het pad met het minste aantal mutaties zijn, maar ook een pad waarin meer kleine mutaties voorkomen. Dit volgt uit het principe van Occam’s scheermes waarbij wordt gesteld dat een gebeurtenis of een fenomeen verklaard moet worden met zo min mogelijk aannames. Elke stap of aanname die geen toegevoegde waarde heeft aan de uitkomst wordt weggelaten (Occam’s razor, 2015). Het is niet mogelijk om met zekerheid het daadwerkelijke verloop van de evolutie te reconstrueren, maar korte en directe paden zijn waarschijnlijker dan zigzagpaden (Hayes, 2007).

Vanwege dit principe is het doel niet om een willekeurig pad te vinden. Het doel is om (1) een algoritme te schrijven dat het pad te vindt dat het genoom in zo min mogelijk stappen omzet en (2) om een algoritme te schrijven dat het pad vindt dat zo min mogelijk genen verplaatst bij het omzetten van het genoom. In dit verslag wordt dat specifiek onderzocht voor het veranderen van het genoom van de Drosophila melanogaster in dat van de Drosophila miranda. Daarnaast worden beide algoritmes getest bij het omzetten van 100 willekeurige genomen in dat van de Drosophila miranda.

Het genoom van al deze fruitvliegen bestaat uit een reeks van 25 genen. Per mutatie zijn er 300 mogelijke omkeermutaties. Dat kan worden uitgelegd aan de hand van een getallenreeks met vier getallen. In afbeelding 7 is te zien dat er in totaal zes mogelijke omkeermutaties zijn. Dit getal staat gelijk aan de som van het aantal genen, min het laatste gen: 3 + 2 + 1 = 6. Bij een genoom van 25 genen zijn de mogelijke mutaties op eenzelfde manier te berekenen: 24 + 23 + ... + 1 = 300. Inversies van grootte 1 worden niet meegenomen, omdat dit in dit geval niets verandert aan het genoom.



*Afbeelding 7.* Alle mogelijke mutaties van een genoom met lengte 4.

Wanneer er geen rekening wordt gehouden met het elimineren van dubbele stappen, loopt bij 15 mutaties het totaal aantal doorloopbare stappen snel op tot 1036. Dit maakt deze opdracht een moeilijk oplosbaar probleem.

**2. Methodes**

Om tot het huidige algoritme te komen, is een overweging gemaakt tussen verschillende zoek-algoritmes. Breadth-First-Search, Depth-First-Search & Iterative Deepening vielen hierbij af nadat deze getest waren, doordat de hoeveelheid data te groot was om binnen aanzienelijke tijd door te rekenen. Uiteindelijk is gekozen om een A\*-algoritme te maken.

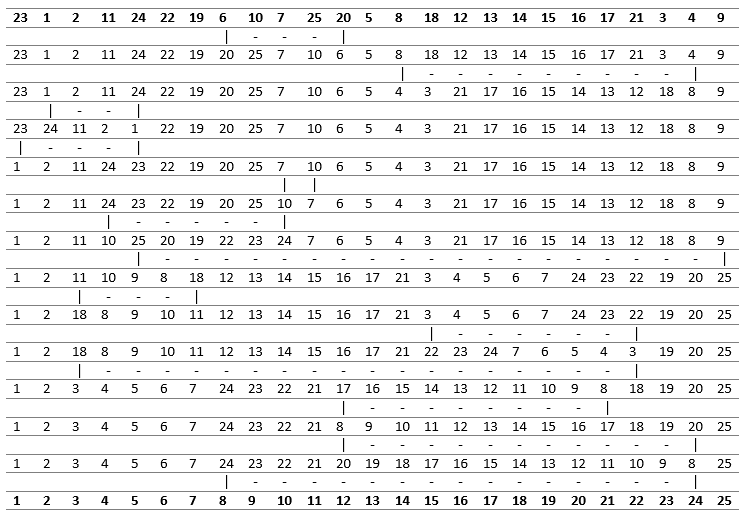
Een belangrijk onderdeel hiervan is de heuristieke functie die aan het algoritme meegegeven wordt. Deze functie geeft voor elk genoom een geschatte afstand tot het doel mee. In het huidige algoritme wordt dit grofweg bepaald door zogenaamde breakpoints te tellen. Een breakpoint ontstaat doordat een gen niet naast zijn uiteindelijke buren staat. In de reeks van afbeelding 3 zijn er tussen de getallen 4 en 2, en 1 en 5 breakpoints. Deze reeks bevat dus twee breakpoints. Om te zorgen dat de reeks in de oplopende volgorde komt te staan, wordt er voor het bepalen van de breakpoints aan het begin en het eind nog een getal geplaatst: 0 4 3 1 2 5 6. Deze getallen veroorzaken wel breakpoints, maar veranderen nooit van plek. Dit brengt het totaal aantal breakpoints voor deze reeks op drie, tussen 0 en 4 zit namelijk nog een breakpoint.

Hierna wordt, afhankelijk van de opdracht, een tweede variabele toegevoegd aan de geschatte afstand. Voor het zoeken van het pad met de minste generaties tussen twee fruitvliegen, wordt het aantal generaties benodigd voor het muteren naar het huidige genoom opgeteld. Voor het zoeken naar het pad met het minste aantal verplaatste genen, wordt het totaal aantal verplaatste genen voor het huidige genoom opgeteld.

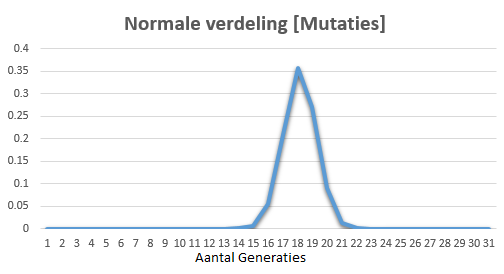
Het huidige heuristiek is niet perfect. Hierdoor is de geschatte afstand niet altijd accuraat genoeg. Dit resulteert soms in een hogere rekentijd. Om ervoor te zorgen dat de test met 100 willekeurige genomen geen onbepaalde tijd zou duren, is er voor gekozen om de rekentijd per genoom te limiteren tot 30 minuten.

**3. Resultaten**

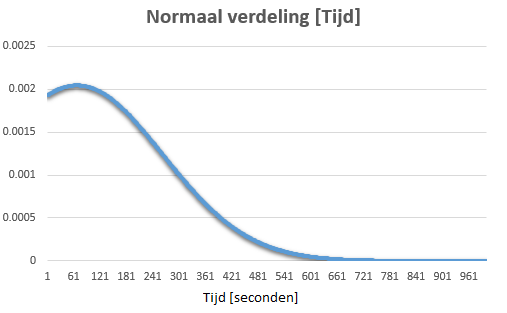
Het genoom van de Drosophila melanogaster kan worden veranderd in dat van de Drosophila miranda in minimaal 13 generaties (afbeelding 8). De som van het aantal verplaatste genen is bij dit pad 128. Wanneer die algoritme op 100 willekeurige genomen wordt gebruikt, onstaat de curve uit grafiek 1. Daaronder, in grafiek 2, is te zien dat het algoritme het meerendeel van de genomen binnen 10 minuten kan berekenen en dat alle genomen tussen de 14 en 22 generaties nodig hebben.



*Afbeelding 8.* Een mogelijke manier om het genoom van de Drosophila melanogaster om te zetten in het genoom van de Drosophila miranda in het minimaal aantal benodigde mutatiestappen, 13.



*Grafiek 1.* Het minimale aantal benodigde generaties bij 100 willekeurige genomen.



*Grafiek 2.* De benodigde tijd voor het berekenen van het minimale aantal benodigde generaties bij 100 willekeurige genomen.

Met een ander heuristiek is het minimaal aantal verplaatste genen bij het evolutietraject tussen de melanogaster en de miranda verlaagd naar 88. Bij dit pad is het aantal generaties 14. In tabel 1 is te zien wat de gemiddelde waarden zijn, de standaarddeviaties, de tijd die het kostte om dit te berekenen en het aantal doorlopen mogelijkheden voor de twee groepen van 100 willekeurige genomen.

|  |  |  |  |
| --- | --- | --- | --- |
| Genome Length | Mutations | Visited | Time [sec] |
| |  | | --- | |  | | 5 | | 10 | | 15 | | 20 | | 25 | | |  |  |  | | --- | --- | --- | | Mean | Min | Max | | 2 | 1 | 4 | | 6 | 4 | 8 | | 9 | 7 | 12 | | 13 | 9 | 16 | | 17 | 13 | 19 | | |  |  |  | | --- | --- | --- | | Mean | Min | Max | | 3 | 1 | 7 | | 25 | 4 | 1058 | | 64 | 8 | 1320 | | 177 | 12 | 1392 | | 1187 | 18 | 9542 | | |  |  |  | | --- | --- | --- | | Mean | Min | Max | | 0 | 0 | 0.04 | | 0.02 | 0 | 1.82 | | 0.26 | 0.01 | 7.79 | | 1.8 | 0.03 | 19.43 | | 67.24 | 0.09 | 1147.01 | |

*Tabel 2.* De gemiddelde prestaties van het algoritme om het pad met het minste aantal generaties te vinden.

|  |  |  |  |
| --- | --- | --- | --- |
| Genome Length | Moved gen’s | Visited | Time [sec] |
| |  | | --- | |  | | 5 | | 10 | | 15 | | 20 | | 25 | | |  |  |  | | --- | --- | --- | | Mean | Min | Max | | 7 | 2 | 12 | | 25 | 9 | 37 | | 49 | 29 | 66 | | 75 | 55 | 96 | | 104 | 77 | 126 | | |  |  |  | | --- | --- | --- | | Mean | Min | Max | | 2 | 1 | 4 | | 7 | 2 | 13 | | 17 | 7 | 203 | | 113 | 12 | 1010 | | 1987 | 35 | 12273 | | |  |  |  | | --- | --- | --- | | Mean | Min | Max | | 0 | 0 | 0 | | 0 | 0 | 0.02 | | 0.04 | 0.01 | 0.42 | | 0.60 | 0.04 | 8.68 | | 132.21 | 0.20 | 1702.77 | |

*Tabel 2.* De gemiddelde prestaties van het algoritme om het pad met het minste aantal verplaatste genen te vinden.

**4. Conclusie en Discussie**

In dit artikel is onderzocht wat het kortste pad is om het genoom van Drosophila melanogaster om te zetten in het genoom van Drosophila miranda. Daarnaast is ook onderzocht wat de manier is om deze omzetting voor elkaar te krijgen waarbij er zo min mogelijk genen verplaatsen. Deze problemen zijn opgelost aan de hand van het maken van een algoritme en twee heuristieken. Daarnaast zijn de twee algoritmes toegepast op 100 willekeurige genomen om deze om te zetten in het genoom van de Miranda.

De resultaten laten een gemiddeld kortste pad zien van 17. Het was mogelijk om het aantal mutaties en het doorlopen pad te vergelijken met uitkomsten van een pad dat was gegenereerd door GRIMM software gemaakt door Glenn Tesler. Deze software wordt veel gebruikt in biologisch onderzoek naar vergelijkbare problemen (GRIMM, 2001). Op deze manier was het mogelijk om met enige zekerheid aan te nemen dat de door het algoritme gevonden paden daadwerkelijk het kortst zijn. Volledig zeker is het helaas niet, omdat zowel GRIMM als de twee heuristieken uit dit verslag niet bij elk genoom het korst mogelijke pad aangeven. Een combinatie van de twee zou dit mogelijk wel kunnen.

Er kan een onderscheid worden gemaakt tussen genomen die gemakkelijk in zo min mogelijk mutaties zijn om te zetten en genomen die dat niet zijn. Het verschil tussen deze genomen heeft waarschijnlijk te maken met het aantal breakpoints dat een genoom bevat. Voor genomen met veel breakpoints zijn meer stappen nodig om het ene genoom in het andere genoom om te zetten.

Het weergeven van reeksen in problemen als getallenreeksen gebeurt vaker. Het geschreven algoritme is dan ook toepasbaar op meer dan alleen het veranderen van deze genenreeksen van de Drosophila. Elk probleem dat valt te reduceren tot het ordenen van een getallenreeks kan in principe worden opgelost met het geschreven algoritme.

In eventueel vervolgonderzoek moet de nadruk worden gelegd op de zekerheid dat een gevonden pad daadwerkelijk het kortste pad of het pad met het minste aantal verplaatste genen is. Daarnaast zouden er vergelijkingen gemaakt kunnen worden met het echte mutatiegedrag van het genoom, om het meest waarschijnlijke genoom beter te kunnen bepalen. Hiervoor is echter ook meer onderzoek vanuit de biologie nodig.

**5. Referenties**

Hayes, B. (2007). Sorting out the genome. American Scientist, 95(5), 386.

Hoskins, R. A., Carlson, J. W., Wan, K. H., Park, S., Mendez, I., Galle, S. E., ... & Krzywinski, M. (2015). The Release 6 reference sequence of the Drosophila melanogaster genome. *Genome research*, *25*(3), 445-458.

Bachtrog, D., & Andolfatto, P. (2006). Selection, recombination and demographic history in Drosophila miranda. *Genetics*, *174*(4), 2045-2059.

Occam’s razor. Retrieved December 18, 2015, from <http://www.collinsdictionary.com/dictionary/american/occam-s-razor>

GRIMM. Retrieved December 18, 2015, from <http://grimm.ucsd.edu/GRIMM/index.html>

1. *Vanuit praktische overwegingen worden genen gerepresenteerd als getallen en daarom het genoom als een getallenreeks.* [↑](#footnote-ref-1)