**Inleiding onderzoek:**

Voor dit onderzoek is het genoom van gist gesequenced om te achterhalen wat de functie van de genen zijn. Deze genen coderen voor eiwitten die op verschillende plekken in de cel kunnen voorkomen. Alle informatie over deze genen staan opgeslagen in de file ’Genes\_relation.data’. In dit bestand staat de volgende informatie over de genen: GeneID’, ’Essential’, ’Class’, ’Complex’, ’Phenotype’, ’Motif’, ’Chromosome’, ‘Function’ en ‘Localization’, zie tabel 1. Ook is er een bestand genaamd interactions\_relations.data hierin staat informatie over de interactie tussen genen. In dit bestand staan de attributen: ‘Gen1’, ‘Gen2’, ‘Type interactie’ en ‘expressie correlatie’, zie tabel 2. Ook zijn er nog 2 ‘.names’ bestanden, hierin staat het format van de bestanden geïllustreerd en welke waarden de attributen kunnen hebben. Deze 4 bestanden vormen de training files, daarnaast zijn er nog 2 toetsfiles gegeven, ‘Genes\_relation.toets’ en ‘Interactions\_relation.toets’. De bestanden hebben hetzelfde format, alleen is bij de toetsfile, ’Genes\_relation.toets’, de locatie onbekend. Het doel is om via machine learning van de trainingfiles, ’Genes\_relation’ en ’interactions\_relation’, een model te maken om te kunnen voorspellen wat de locaties zijn van de genen in de toetsfile ’Genes\_relation.toets’.

*Tabel 1: Het formaat van de Genes\_relation bestanden. Als een eiwit meerdere functies, motifs en/of Phenotype heeft, dan bestaat voor elke combinatie een attribuut.*

|  |  |
| --- | --- |
| Attribuut | Informatie |
| Gen id | Het gen id van het gen. |
| Essential | Hoe essentieel het gen is. Dit kan ‘Essential’, ‘Non-Essential’ en ‘Ambiguous-Essential’ zijn |
| Class | Voor wat voor een soort eiwit het gen codeert (Bv: Transcriptie Factor, Actins, ATPases, etc). |
| Complex | Het complex waar het eiwit in voorkomt. |
| Phenotype | Het phenotype dat een eiwit kan laten ontstaan, als deze tot expressie komt. |
| Motif | Een sequentie in een gen met een biologische functie, die kan variëren. Een gen kan meerdere motifs hebben, dus dit betekent dat deze dan verschillen in sequentie. |
| Chromosome | Het chromosoom waar het gen zich bevindt. |
| Function  Localization | De functie van het eiwit, een gen kan voor meerdere eiwit functies coderen. Ieder eiwit met een andere functie staat op een aparte regel.  De locatie waar het eiwit zich bevindt, per gen id is er maar 1 locatie mogelijk. |

*Tabel 2: Het formaat van de Interactions\_relation bestanden. Een gen kan met meerdere genen een interactie aangaan.*

|  |  |
| --- | --- |
| Attribuut | Informatie |
| Gen id1 | Het gen id van gen1. |
| Gen id2 | Het gen id van gen1. |
| Type  Expression correlation | Dit kan physical, genetic of genetic-physical zijn.  De expressie correlatie tussen de genen. |

Expressie correlatie

In de file “Interactions\_relations.data” staan de bekende expressie correlaties tussen 2 genen. Deze correlatie kan van -1 tot 1 lopen. Een hoge correlatie geeft aan dat het genenpaar elkaar sterk beïnvloed en een correlatie onder de 0 geeft aan dat de expressie van beide genen laag zullen zijn als ze samen in de cel voorkomen.

In de file zijn 3 soorten expressie correlaties te vinden, dit zijn: Physical, Genetic en Genetic-Physical. Een physical interactie betekent dat de 2 eiwitten aan elkaar binden. Een Genetic interactie zal plaatsvinden tussen 2 genen als een van de 2 genen hoog tot expressie komt en bij een Genetic-Physical interactie zullen beide interacties plaatsvinden.

Het effect van deze interacties kunnen variëren van de expressie van een eiwit te onderdrukken of juist te verhogen of de functie te veranderen. Deze correlatie kunnen gebruikt worden als een reductie techniek, een cluster methode (fuzzy clustering), selectiemethode en als een data verkenner. Het is bijvoorbeeld mogelijk om een bepaalde threshold te zetten voor welke genen relevant zijn of welke interessant zijn.

<https://wiki.thebiogrid.org/doku.php/experimental_systems>

**Methode beschrijven die er gebruikt worden bij het maken van de modellen!**