Vrije Universiteit Brussel

Faculty of Science and Bio-engineering Sciences
Department of Computer Science
Computational Modeling Lab

Unlocking the potential of public available gene expression data for large-scale analysis

Jonatan Taminau

Dissertation submitted for the degree of Doctor of Philosophy in Sciences

Supervisor: Prof. Dr. Ann Nowé

Committee members:

Internal members:

Prof. Dr. Ann Nowé
Vrije Universiteit Brussel

Prof. Dr. Bernard Manderick
Vrije Universiteit Brussel

Prof. Dr. Dominique Maes
Vrije Universiteit Brussel

External members:

Prof. Dr. Hugues Bersini
Université Libre de Bruxelles

Prof. Dr. Jacques De Grève
Universitair Ziekenhuis Brussel

Dr. Willem Talloen
Janssen Pharmaceutica NV, Beerse

Dr. Benjamin Haibe-Kains
Institut de Recherches Cliniques de Montréal

Abstract

After more than a decade of microarray gene expression research there is a vast amount of data publicly available through online repositories. It is clear that for the future the new challenges for this technology lie in the integration of this plethora of different data sets in order to obtain more robust, accurate and generalizable results.

A first hurdle for this large-scale integration of studies coming from different labs, using different experimental protocols and even hybridized on different platforms, is the retrieval of the data sets in a uniformed standard. Nowadays it is unfortunately still not possible to retrieve gene expression data in a completely consistent and trackable way and many manual interventions are needed before the actual analysis can be performed. This step is error-prone, leading to obscure errors and reproducibility issues. In this thesis we present the InSilico DB, a tool that provides consistently preprocessed and manually curated genomics data, thereby overcoming many of the current issues related to data acquisition.

In a second hurdle towards the integration of multiple data sets, information from individual gene expression data sets has to be combined and we extensively describe and compare the two main approaches in order to do so: meta-analysis, an approach that retrieves results from individual data sets and then combines the results; and merging, an approach that first combines the actual expression values and then retrieves results on this new data set. Both approaches are described in detail with special attention for their limitations, issues and advantages.

Both for the consistent retrieval of the data and for the integration of multiple data sets we developed two freely available R/Bioconductor packages providing the necessary tools. These two packages seamlessly integrate with each other and we illustrate their power in a final application where we empirically compare both meta-analysis and merging approaches for the identification of differentially expressed genes in lung cancer.

Samenvatting

Na meer dan een decennium of microarray onderzoek is er een grote hoeveelheid data publiek beschikbaar via online repositories. Het is alsmaar duidelijk dat de nieuwe uitdagingen in de nabije toekomst liggen in het combineren van verschillende bestaande data sets om zo meer robuuste, accurate en generaliseerbare resultaten te bekomen.

Een eerste obstakel voor deze grootschalige integratie van studies, komende van verschillende labs en gebruik makend van verschillende experimentele protocollen en technologieën, is het bekomen van de data in een uniform en gestandaardiseerd formaat. Het is vandaag de dag helaas nog niet mogelijk om op een volledig consistente en traceerbare manier data uit deze repositories te verkrijgen en vele manuele interventies zijn nodig vooraleer de effectieve analyse kan uitgevoerd worden. Deze interventies kunnen leiden tot fouten die niet reproduceerbaar zijn. In deze thesis presenteren we de InSilico DB, een online tool die consistent gegenereerde en manueel gecureerde data aanbiedt en zo de huidige problematiek van data acquisitie probeert te verhelpen.

In deze thesis is ook een tweede obstakel geïdentificeerd: het effectief samenvoegen van de informatie van verschillende data sets. We beschrijven uitvoerig de twee gangbare methoden. In *meta-analysis* worden eerst resultaten bekomen van de individuele studies en dan worden die resultaten gecombineerd. In *merging* gaat men eerst de numerieke gene expressie waarden samenvoegen om dan resultaten te bekomen op deze grote gecombineerde data set. Beide methoden worden in detail bespro-

ken met speciale aandacht voor hun limitaties en sterkten.

Zowel voor de consistente acquisitie van de individuele data sets als voor het uiteindelijke integreren, hebben we twee vrij beschikbare en open software pakketten ontwikkeld die de nodige functionaliteit bevatten. Deze twee pakketten zijn reeds opgenomen in het R/Bioconductor framework en werken naadloos met elkaar samen. We illustreren hun mogelijkheden in een finale applicatie waar we meta-analysis en merging met elkaar vergelijken in de context van het vinden van biomarkers in verschillende bestaande long kanker studies.