



Universiteit
Leiden
The Netherlands

Design and development of a comprehensive data management platform for cytomics: cytomicsDB

Larios Vargas, E.

Citation

Larios Vargas, E. (2015, November 25). *Design and development of a comprehensive data management platform for cytomics: cytomicsDB*. Retrieved from <https://hdl.handle.net/1887/36426>

Version: Not Applicable (or Unknown)

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

Downloaded from: <https://hdl.handle.net/1887/36426>

Note: To cite this publication please use the final published version (if applicable).

Cover Page



Universiteit Leiden



The handle <http://hdl.handle.net/1887/36426> holds various files of this Leiden University dissertation

Author: Larios Vargas, Enrique

Title: Design and development of a comprehensive data management platform for cytomics : cytomicsDB

Issue Date: 2015-11-25

Samenvatting

In biomedisch onderzoek wordt veelvuldig gebruik gemaakt van celculturen in het onderzoek. Het onderzoeksveld waarin op grote schaal gebruik gemaakt wordt van cel-systemen wordt aangeduid met Cytomics. High-Throughput Screening (HTS) is een van de meest gebruikte technieken in Cytomics voor het valideren van potentiële kandidaat “targets” voor de ontwikkeling van medicijnen. Hierbij kunnen binnen een experiment grote hoeveelheden verschillende chemische verbindingen worden getest die een cruciale rol spelen in een ziektebeeld en daarmee kandidaat zijn voor het ontwikkelen van medicijnen tegen deze ziekte. Een omgeving voor High-Throughput Screening is samengesteld uit een aantal componenten die onderling samenhang moeten vertonen. Voor een succesvolle toepassing van een dergelijke omgeving zijn er een aantal complicaties die moeten worden overwonnen: (1) het beheren van grote hoeveelheden data, (2) interoperabiliteit, (3) begrijpbaarheid en (4) het delen het hergebruik van data.

In dit proefschrift wordt in 4 hoofdstukken het onderzoek beschreven voor de ontwikkeling van een platform voor onderzoek en beheer van data, genaamd CytomicsDB, waarbij de gehele keten van werkzaamheden, het werkproces, in een systeem wordt ondergebracht. Een geautomatiseerde keten van werkzaamheden wordt voorgesteld waarmee een versnelling van de werkzaamheden teweeg kan worden gebracht; van het begin van het traject bij ontwerp van het experiment tot de data analyse aan het einde van het traject.

In Hoofdstuk 1 wordt een algemene inleiding geven van het werkveld en de concepten die van belang zijn voor het onderzoek. In Hoofdstuk 2 wordt een architectuur beschreven die voor CytomicsDB wordt gebruikt en dit gebaseerd is op een modern relationeel database systeem. Met een dergelijk systeem kunnen bestaande, maar deels verouderde systemen, zogenaamde legacy systemen, worden geïntegreerd; dit zijn de systemen die momenteel in HTS worden gebruikt, maar ook externe opslagsystemen die belangrijk zijn voor de validatie van de data. De voorgestelde architectuur is zeer geschikt voor het delen van data en samenwerking, waarbij verschillende

niveaus van beveiliging van de data zijn ingevoerd die de wetenschapper in staat stellen data te publiceren voor de gehele gemeenschap of juist alleen een selecte groep mede-onderzoekers toegang te geven tot deze data.

Naast data zijn er gegevens over de data, bijvoorbeeld, hoe deze data tot stand zijn gekomen. Deze gegevens worden ook Metadata genoemd, in feite zijn dit eigenlijk data over data. In Hoofdstuk 3 wordt een strategie uitgewerkt voor het beheer van metadata zoals dat in CytomicsDB is ingebed. Er is een semantische schil geconstrueerd waarmee begrijpbaarheid en exploratie van grote volumina data en metadata de betrokken zijn bij een HTS experiment te faciliteren. Tegelijkertijd stelt deze semantische schil de wetenschapper in staat om nieuwe methoden en programma's te integreren waarmee de beeld- en data-analyse kan worden afgehandeld. De resultaten uit deze analyses zullen deel worden van de experiment metadata en komen beschikbaar voor een post data-analyse op een semantisch niveau.

In Hoofdstuk 4 introduceren we het proces van data validatie zoals dat in CytomicsDB wordt uitgevoerd. Het beheer van de HTS informatie is een van de belangrijkste uitdagingen bij het ontdekken van nieuwe medicijnen. Teneinde de consistentie, integriteit en betrouwbaarheid van de data opgeslagen in CytomicsDB te garanderen, is het verplicht in iedere stap van het HTS werkproces een nauwgezette validatie uit te voeren. In CytomicsDB wordt deze validatiestap gefaciliteerd door middel van web-services waarmee iedere nieuwe kritische invoer van data wordt getest aan de hand van een interne of externe opslagsystemen.

In Hoofdstuk 5 wordt de integratie van de CytomicsDB architectuur met een rekencluster, de Leiden Life Sciences Cluster (LLSC), beschreven in samenhang met de omgeving voor data processing die aanwezig is binnen het MonetDB database management system. Voor het efficiënt gebruik van het rekencluster moeten de algoritmen voor beeldanalyse worden aangepast. Deze parallelisatieslag is geëvalueerd en de prestaties van de herschreven algoritmen worden gepresenteerd. De stappen die betrokken zijn bij de analyses worden verder uitgewerkt in een werkproces voor beeld- en data-analyse.

De huidige stand van zaken in HTS experimenten is de uitdaging aan te gaan in de behoefte te kunnen voorzien om complexere platforms voor *Cytomics* te ontwikkelen waarmee ingewikkelde experimenten kunnen worden geanalyseerd en die zich kunnen aanpassen aan de continue veranderingen in het Cytomics veld. De architectuur van CytomicsDB faciliteert de integratie van met andere systemen die binnen HTS experimenten gebruikt worden. De metadata zijn, bovendien, zodanig georganiseerd dat ze voldoen aan een algemeen data-model waardoor de begrijpbaarheid, samenwerking en het delen van data sterk verbeterd is. De validatie van de metadata die in CytomicsDB worden opgeslagen is wordt een sleutel proces, de consistentie van deze metadata wordt verzekerd door een dubbele controle met data uit andere databanken die door de gemeenschap worden beheerd. Tot slot, het platform dat in dit proefschrift wordt gepresenteerd, integreert de voordelen van het toepassen van de beeldanalyse op rekenclusters en daarnaast het gebruik van de functionaliteit voor dataverwerking

die in MonetDB is gebouwd met het softwarepakket MonetDB.R. Hiermee kan een aanzienlijke winst worden behaald met het verwerken van de data waarbij voorkomen wordt dat er heel veel data moeten worden verplaatst voor de verwerking.

