Everything should be linked: linking and visualising data for dynamic multilevel and multidimensional biological data interpretation.

Exploring multi-level effects of structural variations in non-coding genomic regions in cancer

RHWE (Robin) van der Weide*

Cancer Stem cells & Developmental biology
Utrecht Graduate School of Life Sciences

*Supervisor: Joep de Ligt, PhD

Summary of the research

With the increase in popularity and cost-effectiveness of various omics-approaches, more and more data is becoming available to researchers of different fields. The complexity of integrating and analysing information of these approaches increases with every added omics-layer and/or other dimension (e.g. time-series, treatments). The current tools and frameworks for these approaches have two major limitations in their design: scalability and generality (i.e. the possibility to add of more levels and/or dimensions). Moreover, there isn't an option to overview a dataset without filtering, dividing or structuring the data. These limitations restrict the integration of complex dataset, needed to truly understand biology.

Enter the Semantic Web and its Resource Description Framework (RDF): a general and simple framework for making statements about subjects. RDF is already heavily used in fields outside of biology, enabling users to integrate and search data based on semantics. Every RDF-statement (i.e. a Triple) has three parts, in which anyone can say anything about anything: a subject, a predicate and an object. An example of such a Triple is "BRAF1 has the molecular function of binding calcium ion", which has these three parts: a subject (BRAF1), a predicate (molecular function) and an object (binding calcium ion). Another Triple can then say something about the phosphorylated protein levels of this gene in a sample. Connecting these two Triples would enable a researcher to find a possible pattern in the data (i.e. a gene, responsible for calcium ion binding, has a low phosphorylation level in the investigated sample). Since every type of data can be translated to RDF, integration of large datasets of different levels and dimensions becomes possible and a lot more feasible.

One of the other big advantages of using RDF is the ability to combine local and remote RDF-databases (EMBL-EBI has already launched six databases, including UniProt and Reactome), which makes analyses even more powerful. By using the SPARQL Protocol and RDF Query Language (SPARQL), retrieving and manipulating data in RDF is easily readable by both humans and computers. The SPARQL-results can subsequently be visualized as a whole, or filtered by the user.

Here, we propose the use of semantic web technologies and visual analytics to decrease the complexity of integrating and visualizing multi-level and -dimensional biological data. Firstly, we will create the framework needed to design the missing tools for converting the most-used NGS-formats to RDF. Next, methods and tools for visual analytics of the biological RDF-data will be created. With this, we can perform many difficult, previously unmanageable, data-integration studies. Examples of these include analyses on multi-omic networks versus treatments, finding the consequences of complex genomic structural variations, connecting nuclear and mitochondrial data and combining ribosomal profiling with codon-bias, RNA-editing and allele-specific expression.

Layman's summary

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Keywords: *structural variation, multi-level data integration, next-generation sequencing, cancer, visual analytics*

Background, aims and approach

Overall aim

The aim of this project is to integrate and visualise multiple levels and dimensions of (NGS-based) omics-data with methods of the semantic web and, using these methods, further understand the consequences of structural variations in the non-coding regions of the genome on other biological levels, like the transcriptome and proteome. Thus, this proposal has two sub-projects, which rely heavily on each other:

1. **Integration and visualisation**
   Integration of NGS-based data by using semantic web-methodologies and W3C-compliant dynamic visual analytics to improve integrative bioinformatics in general and NGS-based multi-level and -dimensional research in particular.

2. **Multi-level analysis**
   Multi-level and -dimensional integrative bioinformatical analysis to elucidate the consequences of genomic structural variations in non-coding regions in cancer.

Scientific relevance and challenges

The amount of (public) biological data has exploded in the last years -even outpacing Moore's law- which is the result of the advances in omics-technologies, like Next-Generation Sequencing (NGS) and Mass-Spectrometry (MS), in both performance and costs. Aside from the sheer size, a second factor for the highly complex nature of current biomedical research is the addition of other dimensions, like time-series or treatments to the aforementioned omics-levels. While there are plenty of studies on single-level data analysis (Huttenhower and Hofmann, 2010), both basic research and industry are agree that data-integration is key to understand the complex nature of biology more thoroughly (Gomez-Cabrero et al., 2014; Searls, 2005).

of da are Several studies, that have combined multiple levels or dimensions have been performed (Low et al., 2013)

Originality and innovative character

Sahoo et al. (2008) geeft heel veel info over BIO-RDF. Information gain through entailment reasoning is an important advantage of ontology-based data integration.

Methods and techniques

Research plan

Timetable

Collaboration

Knowledge utilisation

Gomez-Cabrero et al. (2014) geeft wensen van de community aan

References

Gomez-Cabrero, D., Abugessaisa, I., Maier, D., Teschendorff, A., Merkenschlager, M., Gisel, A., Ballestar, E., Bongcam-Rudloff, E., Conesa, A., and Tegnér, J. (2014). Data integration in the era of omics: current and future challenges. *BMC Syst. Biol.*, 8(Suppl 2):I1.

Huttenhower, C. and Hofmann, O. (2010). A quick guide to large-scale genomic data mining. *PLoS Comput. Biol.*, 6(5):e1000779.

Low, T. Y., van Heesch, S., van den Toorn, H., Giansanti, P., Cristobal, A., Toonen, P., Schafer, S., Hübner, N., van Breukelen, B., Mohammed, S., Cuppen, E., Heck, A. J. R., and Guryev, V. (2013). Quantitative and qualitative proteome characteristics extracted from in-depth integrated genomics and proteomics analysis. *Cell Rep.*, 5(5):1469–78.

Sahoo, S. S., Bodenreider, O., Rutter, J. L., Skinner, K. J., and Sheth, A. P. (2008). An ontology-driven semantic mashup of gene and biological pathway information: application to the domain of nicotine dependence. *J. Biomed. Inform.*, 41(5):752–65.

Searls, D. B. (2005). Data integration: challenges for drug discovery. *Nat. Rev. Drug Discov.*, 4(1):45–58.