Development of databases and internet resources for biological data

Nadine E. Tatto
nadine.tatto@boku.ac.at; nadine.tatto@ncl.ac; nadine.tatto@boku.ac.at

Muthgasse 18, 1180 Vienna
Univ.Prof. Dipl.-Ing. Dr.nat.tech.n. Diedrich Mattsonovich

Dr. Alessandra Graf, Muthgasse 11, 1190 Vienna
Univ.Prof. Dipl.-Ing. Dtech.n. Friedrich Leisch, Peter-Jordan-Straße 82, 1190 Vienna

Introduction:

In bioinformatics, the need and benefit for internet resources was and is always present, beginning with the first web pages and the rise of bioinformatics as a discipline. Since the amount of data, especially the huge amount of sequence data, is increasing and the collaborative work across the globe on it is rising, the need becomes even stronger. Especially websites for specific organisms, like Wormbase 1, PlasmoDB 2, Toxodb 3, Candida Genome Database 4 and more, are increasing in number. Currently the only way to retrieve and visualize data and information on Komagatella pastoris (a.k.a. Pichia pastoris) is the NCBI - National Center for Biotechnology information 5, which is not updated regularly with the most recent data — and the outdated Libview 6 genome browser with likewise outdated data. Furthermore, additional data is spread over several other online resources, like Lycocme — Universal Protein Resource 7. With the improved sequences and annotations available for Komagatella pastoris CIP-4212 — a.k.a. homographe phyl 8, 9, 10 and other Komagatella strains the opportunity for an internet resource comparable to other organism specific databases and websites is obvious and essential. Additionally, the availability of improved visualization tools – like Ikaros 11, 12 – gives us the opportunity to provide a better visualization of the Komagatella sp. data regarding usability and extensibility. Especially data from sources like RNA sequencing can be visualized in a graphical way with a greater usability.

Question:

How to develop and provide an integrated platform for the P. pastoris research community, with useful tools, data access and data visualization? How should such a platform be designed and which features should it provide? How many and what kind of frameworks are available to package analysis pipelines and which of them are suitable for bioinformatics purposes?

Hypothesis:

My hypothesis is that, utilizing qualitative interviews of potential users and comparison with other species centered online resources, the development of a state-of-the-art online service platform for data representation and analysis for Komagatella sp. can be achieved. This platform will be able to satisfy redeemer users and community requirements.

Material & Methods:

Data:

3 Sequences of P. pastoris strains, S. cerevisiae strains:
- K. phaffii CBS743 12,13
- K. phaffii CBS315 14
- K. pastoris DSM20302 15
- S. cerevisiae 2ASB 16
RNA-Seq data available and newly generated, ChIP-Seq data to be generated.

Community/ User survey:

Determine user requirements on analysis tools, database, data representation and visualization.

Database:

RDIMS - Oracle Enterprise Edition 11g was selected as database system, because of its versatile tools and options.

Website:

Development and implementation of a publicly accessible web interface fulfilling the determined requirements. Possible technologies HTML, JavaScript or Oracle Apex 5 (Oracle Corporation, Redwood City, USA, 2016).

Visualisation:

Suitable, practical and state-of-the-art tool for graphical representation and visualization of sequence data in various formats. For example genome browser Abrowses 17, UCSC Browser 18, Ikaros 11, 12.

Analysis tools:

Development and implementation of online, publicly available analysis pipelines for RNA-Seq and ChIP-Seq analysis. Several self developed plugins for the chosen visualization tool. Suitable development methods and tools are to be chosen. State-of-the-art data analysis tools and workflows are to be used, for example regarding RNA-Seq analysis.

Outlook & Expectations:

As a result a database incorporated with sequence and annotation data of several P. pastoris strains and a website which is populated by this database. Users should be able to access, filter, search and download the data through a publicly accessible website. Dedicated user shall also be able to administer the data inside the database. Integrated into this website a genome browser will be available to visualize the sequence data in addition to further analysis tools and pipelines. Newly developed plugins for the genome browser will be available to facilitate the sequence data directly inside the genome browser, e.g. using Blast directly from the browser. The analysis pipelines will, amongst others data types, enable users to analyse RNA-Seq data in different ways, e.g. only based on count abundance and differential expression analysis after sequence assembly.

References/Literature:


Acknowledgements: My gratitude goes to my supervisors and my employer for the advice, support and opportunity to perform this project.