Development of computational methods for analysing proteomic data for genome annotation

University of Cambridge

Darwin College

A thesis submitted for the degree of

Doctor of Philosophy

Markus Brosch

The Wellcome Trust Sanger Institute,
Wellcome Trust Genome Campus,
Hinxton, Cambridge, CB10 1SA,
United Kingdom.

December 2009

Dedicated to my family

Declaration

This thesis describes work carried out between May 2006 and December 2009 under the supervision of Dr Jyoti Choudhary and Dr Tim Hubbard at the Wellcome Trust Sanger Institute, while member of Darwin College, University of Cambridge. This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text.

This thesis does not exceed the specified length limit of 300 pages as defined by the Biology Degree Committee. This thesis has been typeset in 12pt font using LaTeX2ε according to the specifications defined by the Board of Graduate Studies and the Biology Degree Committee.

Markus Brosch
December 2009.

Summary

Current functional genomics relies on known and characterised genes, but despite significant efforts in the field of genome annotation, accurate identification and elucidation of protein coding gene structures remains challenging. Methods are limited to computational predictions and transcript-level experimental evidence, hence translation cannot be verified. Proteomic mass spectrometry is a method that enables sequencing of gene product fragments, enabling the validation and refinement of existing gene annotation as well as the detection of novel protein coding regions. However, the application of proteomics data to genome annotation is hindered by the lack of suitable tools and methods to achieve automatic data processing and genome mapping at high accuracy and throughput. The main objectives of this work are to address these issues and to demonstrate the applicability in a pilot study that validates and refines annotation of *Mus musculus*.

In the first part of this project I evaluate the scoring schemes of “Mascot”, which is a peptide identification software that is routinely used, for low and high mass accuracy data and show these to be not sufficiently accurate. I develop an alternative scoring method that provides more sensitive peptide identification specifically for high accuracy data, while allowing the user to fix the false discovery rate.

Building upon this, I utilise the machine learning algorithm “Percolator” to further extend my Mascot scoring scheme with a large set of orthogonal scoring features that assess the quality of a peptide-spectrum match. I demonstrate very good sensitivity with this approach and highlight the importance of reliable and robust peptide-spectrum match significance measures.

To close the gap between high throughput peptide identification and large scale genome annotation analysis I introduce a proteogenomics pipeline. A comprehensive database is the central element of this pipeline, enabling the efficient mapping of known and predicted peptides to their genomic loci, each of which is associated with supplemental annotation information such as gene and transcript identifiers. Software scripts allow the creation of automated genome annotation analysis reports.

In the last part of my project the pipeline is applied to a large mouse MS dataset. I show the value and the level of coverage that can be achieved for validating genes and gene structures, while also highlighting the limitations of this technique. Moreover, I show where peptide identifications facilitated the correction of existing annotation, such as re-defining the translated regions or splice boundaries. Moreover, I propose a set of novel genes that are identified by the MS analysis pipeline with high confidence, but largely lack transcriptional or conservational evidence.

Acknowledgements

First and foremost I would like to thank my supervisors Jyoti Choudhary and Tim Hubbard for giving me the opportunity to carry out this project and for all their invaluable advice, support and encouragement. Many thanks also to my thesis committee Richard Durbin, Gos Micklem and John Cottrell for their critical and constructive assessment of my work. I thank the Wellcome Trust for my PhD studentship and Matrix Science for funding conference travel.

I extend my gratitude to Lukas Käll for all the fruitful peptide scoring discussions and beyond, as well as for extending the Percolator software package interfaces, without which parts of my project would not have been possible. Special thanks to John Cottrell and David Creasy for extensive support with Mascot related questions and problems. I am also very grateful for their decision to integrate Mascot Percolator, which was developed as part of this project, into their upcoming Mascot 2.3 release, enabling widespread use of this method. Also, I would also like to thank David Fenyo for X!Tandem related scoring discussions as well as Mario Stanke for help with the Augustus gene prediction software. Many thanks also to Eric Deutsch and Zhi Sun for answering Peptide Atlas related questions as well as providing me with a comprehensive mouse dataset used in this work.

Many thanks to Sajani Swamy, who introduced me to the Mascot software and Parthiban Vijayarangakannan for writing a web application for Mascot Percolator. I would like to express my gratitude to Mercedes Pardo, who "tried" to introduce me to the "web-lab world", as well as Lu Yu and Mark Collins for their great efforts to provide me with the required mass spec data. I am also grateful to the HAVANA team around Jennifer Harrow, who started to investigate my proteogenomic data. Particular thanks to Felix Kokocinski and Jonathan Warren for extensive help with setting up the distributed annotation server and helping with related problems.

I enjoyed working in Jyoti Choudhary and Tim Hubbard's research group and thank every member and ex-member for the enjoyable time at the Sanger Insitute. Jyoti Choudhary, Tim Hubbard, James Wright, Mark Collins and Daniel James kindly read the draft of this thesis, providing me with valuable feedback. Thank you.

On a personal note, I want to thank my partner Daniela Wieser for supporting me in my endeavour and for putting up with my ridiculous working hours.

Markus Brosch,
Wellcome Trust Sanger Institute,
December 2009.

Contents

List of Figures

Nomenclature

AMT	Adjusted Mascot Threshold
E-value	Expectation Value
FDR	False Discovery Rate
FP	False Positive
MATH	Mass Accuracy-Based THreshold
MHT	Mascot Homology Threshold
MIT	Mascot Identity Threshold
MMD	Maximum Mass Deviation
MS	Mass Spectrometry
MS/MS	Tandem Mass Spectrometry
PEP	Posterior Error Probability
PPM	Parts Per Million
PSM	Peptide Spectrum Match
ROC	Receiver Operating Characteristics
SQL	Structured Query Language
TP	True Positive