Report doctoral thesis – Ing. Jiří Novák

Summary and description of the work:

The doctoral thesis "Similarity Search in Mass Spectra Databases" has been submitted by Ing. Jiří Novák to the Faculty of Mathematics and Physics of the Charles University in Prague. In this thesis the PhD candidate addresses an important research topic in computational mass spectrometry-based proteomics – the efficient identification of peptides and proteins from tandem mass spectra. The thesis is divided into nine chapters. Chapter one briefly introduces the biological processes that are underlying the central dogma of molecular biology, which explains the information flow from the genome (DNA) over the transcriptome (RNA) to the proteome (proteins). In this chapter the thesis motivates the importance of the analysis of proteins. Chapter two introduces the concepts that are used in mass spectrometry to analysis mixtures of proteins. The thesis explains fundamentals in laboratory methods for the digestion of proteins to peptides. This process is essential since only peptides can be efficiently analyzed on most of the common mass spectrometry setups. Furthermore, the next important step on most proteomics experiments is the use of high-pressure liquid chromatography (HPLC) to separate complex mixtures of peptides. Mass spectrometry is introduced as the method that is online coupled to the chromatographer to further separate and detect the peptides according to their mass. Mass spectrometry consists of three parts: the ion source, the mass analyzer and the detector. All these parts are introduced and described. Tandem mass spectrometry is a method that enables the 'sequencing' of peptides based on the recording of entire peptide masses and so called fragment ions of peptides that correspond to parts of the peptide. The concept of tandem mass spectrometry, as well as the physical and chemical principles underlying peptide fragmentation are introduced as the last part of this chapter.

The next chapter introduces computational methods that have been establish to process mass spectra. The chapter is logically divided into a section the reports on algorithms for preprocessing of mass spectra, a paragraph on algorithms for peptide identification, followed by paragraphs on protein identification, protein quantification and on frameworks that can be used to analyze shotgun proteomics data. The candidate describes methods for spectral quality filtering, he introduces the most commonly used database search engines, elaborates on *de novo* approaches for identifying peptides and introduces concepts for the statistical assessment. Since the identification of proteins is conceptionally different for the peptide exercise, he introduces the main methods form protein identification. Labeled and label-free quantification, as well as the algorithmic challenges are outlined. The Trans-Proteomic-Pipeline and OpenMS/TOPP are introduced as the major software frameworks in the field.

In its chapter four the thesis reports on time constraints in searching ever growing databases of proteins. It introduces several concepts that aim at reducing the run of database searches. These concepts include among other precursor mass filters, peptide sequence tags and different distance functions. All these concepts are thoroughly introduced and discussed.

The following chapter five reports on metric and non-metric access methods that build the basis for the scientific novelty of the thesis. It introduces the mathematical foundations of metric methods, such as the metric spaces and metric distances. Different distance metrics, such as the Minkowski, the Cosine and the Hausdorff distance are introduced. Different algorithms are outlined that make use of metric distances to perform similarity searches. These algorithms are represented with the respective pseudo codes and the methods are discussed in the context of their performance in similarity search. A special focus is put to the M-tree method, as a database index structure, which appears to provide good performance in secondary memory. As major non metric methods, the chapter reports mainly on the TriGen algorithm and shows how non metric methods along with the triGen algorithm and a comparable tree structure can be implemented.

Chapter 6 describes the major findings and developments of the thesis which is an efficient implementation of a non-metric similarity search strategy for peptide identification based on MS/MS spectra. First it discusses the angle distance and logarithmic distance are introduced as similarity measures. An important part in the suggested strategy is the indexing of the database. From the indexed database a set of theoretical spectra is queried against the experimental set of tandem spectra and adequate peptide sequences are assigned to the spectra. The querying corresponds to the scoring and the k nearest neighbor method is used for assigning

scores. An important finding in this chapter is the fact that most of the existing methods do not consider potential modifications in the experimental spectra while indexing the database. The presented method based on the Hausdorff function is able to cope with modified spectra. Furthermore, the chapter suggest some additional modules in the processing pipeline, such as clustering of spectra as pre- and sequential scans of protein sequence candidates as postprocessing tools.
The methods that are suggested as scientific novelty in chapter six are then compared against the state-of-the-art methods in chapter seven. The candidate outlines a variety of freely available data sets that he used for his performance evaluation. A detailed listing of performance evaluation metrics follows the in this chapter. The author shows that his pipeline 'SimTandem' outperforms most of the state-of-the-art tools in most of the datasets (except for one).
Finally chapter eight outlines the implementation of the suggested algorithms. The methods is made available through a web interface, as well as through a C++ implementation as a TOPP tool within the OpenMS framework. This chapter also provides documentation on the usage of both, the web interface and the command line or GUI application in TOPP/OpenMS.
Chapter nine summarizes the findings of the thesis.

Scientific novelty of the results:

After decades of proteomics research the task to identify peptides and proteins from MS/MS spectra still remains an unsolved problem. Over the years numerous algorithms have been published that enable qualified user to perform this task that is computationally very demanding. It that sense methods that enable efficient identification of peptides need to fulfill three major criteria: (1) the methods need to be sensitive (i.e. they should identify as many spectra as possible); (2) they should perform this task rather fast (with more and more data being accumulated in proteomics, computation will become a bottleneck rather soon); (3) they should be usable by non-experts.
The hereby presented work presents an approach that improves the current state-of-the-art on all three criteria. It outperforms well established tools, it is significantly faster and through intuitive interfaces and has been made accessible for non-experts. Another important point is that the suggested algorithm can cope with modifications in peptide candidates and still provides fast run times. Modified peptides are causing many problems in the identification of MS/MS spectra. Thus, appropriate methods are highly needed. The identification of modifications on proteins is one of the major goals of proteomics research.

Applications and implications to neighboring areas:

- In other areas of bioinformatics database searches are equally important as in proteomics. In genomics millions of small DNA fragments (reads) are mapped against a reference genome. While the genome in that sense is nothing else than a database, most of the current methods implement sequential querying of the reads. It would be an intriguing experiment to test the methods suggested in this thesis for their performance in genome mapping experiments.
- The suggested method uses a scoring (querying) schema that is different from the presented alternative methods and it also appears to be more sensitive. Unfortunately the thesis did not elaborate on the qualitative overlap of spectra that were identified by SimTandem and alternative tools such as OMSSA and X!Tandem. It would be very interesting, however, to integrate SimTandem into Consensus approaches, such as the OpenMS ConsensusID that has been introduced in chapter three of this thesis. The fast runtime would give it an excellent advantage. In combination with other approaches a significant increase in identified peptides can be expected.

Form of the thesis:

The thesis is written in sound English is very well structured. The different chapters introduce both the biological background and the theoretical foundation that are essential to follow the main scientific part. While chapter one touches on the central dogma of molecular biology, I would have expected a bit deeper motivation for proteomics and also in comparison to other technology platforms, such as transcriptomics. Ideally this first chapter should

address questions such as: Why should one look at the proteins and not only to DNA and RNA? Would it make sense to look at all three levels?
In the following I will list some minor issues:

- Page 8: "The reason is that the evolution tends to preserve the structure (i.e., the function) of a protein rather than the sequence." ...deserves a reference
- Chapter 2:"Mass Spectrometry Fundamentals " ...the fundamentals that are presented here are the fundamentals of "Mass spectrometry-based proteomics". Mass spectrometry has many application areas beyond proteomics.
- It would be desirable to have a clear definition of 'shotgun proteomics' at some point
- The description of 'high-performance-liquid chromatography' in chapter two could be a bit more detailed. The paragraph does not describe the underlying principles why some peptides are retained on the column and others not, which is an essential feature for the separation of peptides
- The description of 'electrospray ionization' in chapter two could be a bit more detailed. It would be important to mention that depending on the voltage one can also add negative charges to analytes and multiple charges are also frequently observed on peptides.

Author's ability for creative scientific work:

I strongly recommend the author's ability for creative scientific work. The first and utmost reason is that the author tried to merge two fields: software engineering and mass spectrometry. He identified a very suitable application area of theoretical concepts that were initially developed in a very different context. This kind of interdisciplinary approach will be highly appreciated by both communities. Classical education in the life sciences does not go deep enough into formal methods and in most cases education in formal methods does not provide deep insight in the life sciences. Thus, interdisciplinary PhD studies are very important to drive both fields.

Recommendation:

I consider the submitted thesis as a sound scientific work and based on this work I fully recommend to award the doctoral degree to the author of this thesis.

Tübingen, July 26, 2013

S. Nahnsen

Dr. Sven Nahnsen