SMP: Cell

Project: Combinatoric Prediction of Splice Variants from MS/MS Data

H.W. Mewes – MIPS, GSF München – w.mewes@gsf.de

Introduction

In this project we aim to develop new methods for the search for novel splice variants from human and murine proteins based on MS/MS data. In order to do so, a pipeline of high throughput methods needs to be established to allow for fast and accurate analysis of the acquired data. Once this pipeline has been accomplished, we will systematically search for new splice variants in cooperation with experimental groups from the NGFN.

This pipeline naturally features several crucial steps which need to be tested and optimized independently from each other. The first step consists of the translation of the measured tandem MS spectra into peptide fragment sequences. These spectra are usually obtained by tandem mass spectrometry of protein mixtures that are purified and separated by 2D gel electrophoresis and cleaved by trypsin. There are – in principle – two different ways to perform this translation: The standard procedure is to calculate theoretical spectra from all proteins in a given database and match the measured spectra against them. There are several commercially available tools which follow this approach like Sequest (***1***) and Mascot (***2***). From the GPM project comes a free tool called X!-Tandem (***3***).

An alternative to these database lookup techniques mentioned above is to perform a *de novo* prediction of peptide fragment sequences. Therefore the tandem MS spectra are translated into amino acid sequences and subsequently searched in a sequence database using sequence similarity search tools. The advantage of this approach is the unbiased prediction of the peptide fragments from the available data alone. Several algorithms have been developed for this retranslation of MS spectra into protein sequences, for example Lutefisk (***4***) and PEAKS (***5***).

The gathered peptide fragments must be aligned against the human or murine genome with standard sequence similarity based methods. This procedure yields candidate coding regions on the target genome but needs to get combined with statistical methods in order to distinguish between regions of higher and lower significance. Exon regions with higher confidence are assembled into candidate genes.

The third step is the scoring of these candidate genes against a custom splice variant database that needs to be tailored specifically for our needs. Only at this level a detection of different splice variants will be possible based on the calculated scores.

Project Status

As this project is still in an initial phase, not all of the above described steps are implemented yet. Subsequently the currently implemented protocol is described which must not be regarded as final. As mentioned above the first step is to translate the measured tandem MS spectra into peptide fragment sequences. As we are looking for novel splice variants from possibly unknown proteins, we decided against protein database lookup tools. Although the GPM X!-Tandem is integrated into the workflow for testing purposes. Currently we use Lutefisk which was kindly provided by Richard Johnson. This tool tries to predict peptide fragments *de novo* from a given tandem MS (MS/MS) spectrum in a network based approach (***4***).

The obtained small sequence fragments are then aligned in a first step against the genomic DNA sequence in a quick but not very sensitive BLAST search.

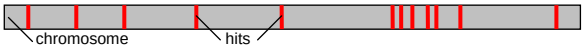

Fig 1: *In the first step we perform a rough search for the protein sequences generated from MS/MS data. Currently we use standard sequence mapping algorithms such as BLAST or BLAT.*

Because of the short length of the query sequences (typically between five and ten amino acids), many hits are generated which are more or less equally spread over the chromosomes. Regions of clustering hits are identified by a self-developed algorithm. The rationale behind this is that the region of the genome where the protein is encoded should have a higher hit density than regions that were only found by chance.

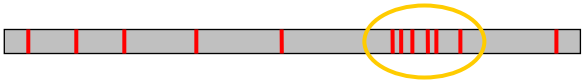

Fig 2: *Clustering the hits on the genome, we identify potential coding regions.*

The identified hit clusters serve as a basis for an advanced investigation. We cut these potential coding regions from the chromosomes and use them for a refined search. This search includes a more accurate alignment process and sequence based reediting of the aligned areas, e.g. joining of neighbouring hits under specific conditions.

Fig 3: *The identified hit clusters serve as a base for a more accurate search. This search includes a more accurate alignment process and sequence based reediting of the aligned areas.*

Currently we test and optimise the performance of our prediction pipeline with theoretical spectra calculated from a set of approximately 600 curated proteins from the German cDNA consortium (***6***). The proteins get digested *in silico* and the resulting spectra of the peptide fragments are calculated by TheoSpec (***7***). This training set is then used to refine the search parameters and the scoring algorithm for the hit clusters.

The next milestone is the development of a database of all possible splice variants of all human and murine genes. As this is still a moving target the database has to be easily maintainable so that updates and corrections to that database can be performed automatically. Based on this database, the predicted potential coding regions are used to score alternative gene models. This might be extended to the

prediction of novel gene models once we get higher sequence coverage from the experimental data.

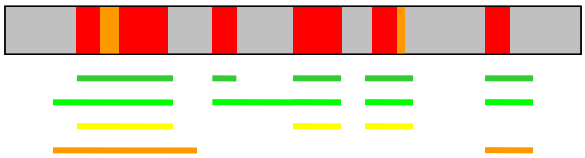

Fig 4: *Finally the predicted potential coding regions are used to score alternative gene models from our splice-candidate database.*

Outlook

Before analyzing a protein by mass spectrometry, it has to be digested using a cutting enzyme, typically trypsin. Due to length limitations, the resulting peptides only cover about 40% on average of the original protein. However, there are other endopeptidases available. On the other hand this means that a dramatic improvement of our method might be achieved if different endopeptidases are used and results are combined. This is one of our main goals for the future.

Our method is similar to those based on ESTs. It is highly biased towards preferentially expressed proteins but has the advantage that only stable splice variants are detected. However, target oriented experiments, e.g. using protein purification techniques, may allow for searching splice variants of specific genes.

In the near future we will test other search tools for the first rough search step, e.g. BLAT (***8***). Depending on its performance we will use this instead of BLAST. The second search might get refined by using more accurate algorithms (e.g. full Smith-Waterman). Moreover we try to increase the prediction quality by incorporating additional information from the genome sequence like donor/acceptor sites or transcription factor binding.

In case of questions please contact M. Erdmann (marco.erdmann@gsf.de) or A. Facius (a.facius@gsf.de).

Lit.: ***1.*** *Eng J K et al. An approach to correlate MS/MS data to amino acid sequences in a protein database. J Am Soc Mass Spectrom. 1994;5(11):976-89.* ***2.*** *Perkins D N et al. Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis. 1999;20(18):2551-67.* ***3.*** *David Fenyö et al. A Method for Assessing the Statistical Significance of Mass Spectrometry-Based Protein Identifications Using General Scoring Schemes. Anal Chem. 2003 February 15;75(4):768-74.* ***4.*** *Taylor J A et al. Implementation and Uses of Automated de Novo Peptide Sequencing by Tandem Mass Spectrometry. Anal Chem. 2001 June 1;73(11):2594-604.* ***5.*** *Ma B et al. PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. Rapid Commun Mass Spectrom. 2003 October 30;17(20):2337-42.* ***6.*** *Available: mips.gsf.de/projects/cdna* ***7.*** *Boehm A M et al. Command line tool for calculating theoretical MS spectra for given sequences. Bioinformatics. 2004 November 1;20(16):2889-91.* ***8.*** *Kent W J. BLAT - The BLAST-Like Alignment Tool. 2002 April;12(4):656-64.*