A software pipeline dedicated to automatic MS/MS data analysis

Erwan Reguer [(1)], Estelle Nugues [(1)], Romain Cahuzac[(2)], Myriam Ferro [(2)], Thierry Vermat [(3)], Emmanuelle Mouton [(2)], and Jérôme Garin [(2)]

[(1)] Equipe Helix - INRIA Rhône-Alpes - 655 avenue de l'Europe - Montbonnot - 38334 Saint Ismier Cedex
erwan.reguer@inrialpes.fr, estelle.nugues@inrialpes.fr
[(2)] Laboratoire de Chimie des Protéines - ERIT-M 201 CEA/INSERM - CEA/Grenoble - 38054 Grenoble
cahuzac@dsvsud.cea.fr, ferro@dsvsud.cea.fr, mouton@dsvsud.cea.fr, garin@dsvsud.cea.fr
[(3)] Société GENOME express - 11 chemin des Prés 38944 Meylan
t.vermat@genomex.com

Keywords. Computational proteomics, MS/MS Spectra analysis, Protein/Gene identification, high-throughput data analysis

Introduction

With the recent improvements of MS/MS QTOF spectrometers biologists can now generate very large amount of spectral data (up to 1500 peptides per day) that can no longer be analyzed manually. There is therefore a growing need for computer systems (pipelines) allowing fully automated protein identification from raw MS/MS data. So far, two main approaches have been proposed to this purpose [3]:

1) Direct identification that consists in the comparison of the raw MS/MS spectrum with all entries of a virtual MS/MS spectra database (such as Mascot [5] approach).
2) Indirect identification which involves two successive steps i) MS/MS spectrum interpretation (i.e. determination of amino acid sequences like in the *de-novo* sequencing approach) followed by ii) protein identification from the corresponding peptides.

In this paper, we present an approach for automatic protein identification dedicated to high-throughput proteomics. This approach follows the line of the indirect protein identification method but, unlike *de-novo* sequencing, does not require the determination of long sequence stretches . It is based on a concept, named Protein Sequence Tag (PST), which has been introduced in [6]. In order to fully exploit this concept we designed two complementary software modules: Taggor for PSTs generation from spectra (MS/MS data interpretation) and PepMap for PSTs localization on protein or genomic data (protein/gene identification).

Protein Sequence Tag (PST)

Protein Sequence Tags (PSTs) can be easily generated from MS/MS spectra analysis [6,4]. A PST is defined by a short peptide sequence (3 to 5 amino acids) flanked by two masses corresponding to the two adjacent polypeptides (Fig. 1).

These two masses represent the two unknown sub-sequences corresponding to the spectrum areas that are not easily interpretable. Using PSTs as results of mass spectrometry analysis offers several advantages:

1) Only few peaks (n+1 where n is the size of the PST sequence) are needed for the PST determination that is therefore more easy to achieve than the complete peptide sequence (especially when the spectrum is noisy).
2) Generating several overlapping PSTs from one spectrum, may contribute to produce *de-novo*-sequencing-like results while avoiding *de-novo* algorithms frequent misinterpretations.
3) From the algorithmic point of view, rapid pattern-matching of short PST sequences can be achieved very efficiently since the small size of PST sequences allow to use hashing techniques.

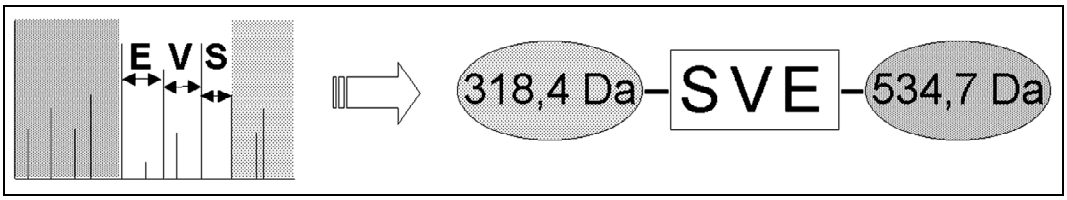

Fig. 1 Peptide Sequence Tag interpretation (right) of a MS/MS spectrum (left)

From spectra to PSTs: *Taggor*

The Taggor module aims at determining a set PST candidates from a processed MS/MS spectrum (i.e. a peak-list) without complete peptide reconstruction or protein database query. The Taggor algorithm involves three steps:

1) All sequences of N amino acids (N equals 3 to 5) are generated.
2) Each of these sequences is aligned to the MS/MS spectrum according to the Y ions reading orientation. If a sequence matches a set of peaks, then flanking masses are computed and a new PST is generated. A score is defined for each PST: it consists in the product of the relative intensities of spectrum peaks that correspond to the PST anchoring sequence.
3) Finally, a set of best PST candidates (usually 20 PSTs) is retained for further analysis.

From PSTs to proteins/genes: *PepMap*

PSTs, generated from MS/MS spectra analysis as described above, are then subjected to a second software component: PepMap. PepMap is responsible of mapping the PSTs to protein databases or, directly, to translated complete chromosomes. PepMap algorithm consists in two steps:

1) Mapping a PST on a polypeptide sequence consists in searching for the PST sequence part, then checking that at least one of the flanking masses corresponds to the adjacent sequences. Working on genomic data involves PSTs mapping on the six translation frames of genomic sequences. By taking into account partial matches (i.e. one of the flanking masses is not recognized while the other is), PepMap may additionally provide important information about intron/exon boundaries (Fig. 2).

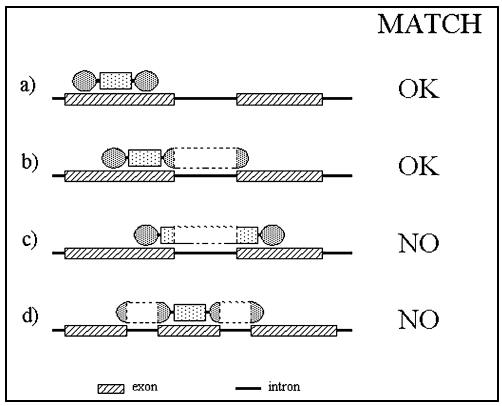

Fig. 2 possible PST matching types

2) An optional clustering phase aims at grouping the PSTs matches belonging to the same protein in order to help identifying the corresponding gene. This clustering phase is, of course, only useful when

applying PepMap to complete, unannotated, chromosome. We devised several algorithms to this purpose but good results are obtained by simple single linkage clustering procedure : a match is clustered with other surrounding matches, if they are closer than a given maximum distance (typically 5000bp for *Arabidopsis thaliana* genome and 15000bp for human genome).

Taggor-PepMap pipeline first results

Given 9 experimental sets on *Arabidopsis thaliana* chloroplast proteins, representing 1298 LC MS/MS spectra (actually peak lists), we have compared the proteins assignments automatically produced by the Taggor-PepMap pipeline to the ones provided by a human expert [1]. Taggor settings were: 3 amino acids length PSTs, a maximum of 20 PSTs generated from a spectrum and mass tolerance was set to 500ppm. Generated PSTs were scanned by PepMap against the *Arabidopsis thaliana* proteins from TIGR (http://www.tigr.org) & TAIR (http://www.arabidopsis.org). We decided to keep each protein entry on which 3 or more PSTs matched completely (no partial matches). From the 1298 spectra, the expert assigned 113 proteins (manually and with the use of Mascot software). From the same dataset Taggor generated 11300 PSTs from which PepMap assigned 101 proteins. 69 of these assignments are same as the expert's ones.
When compared to human assignments Taggor-PepMap missed 44 proteins. Further studies on these proteins allow us to identify three main reasons for this failures: i) expert succeed to assign MS/MS spectrum to a protein even when the studied peak list consist in 1 significant peak (parent ion mass); ii) some spectra cannot be interpreted by reading the Y ions (in those rare cases the "b ions" are needed); iii) human expert can assign a protein from only one MS/MS spectrum whereas our requirement of at least 3 PSTs complete matches per protein generally led to using at least two different MS/MS spectra for this protein..
From the 32 "overpredicted" protein assignments, one has been identified as a human expert miss (a Na+ transporter). The other "false-positives" are under closer examination.

Conclusion

The first results of the combined Taggor-PepMap pipeline are very encouraging. The PST approach greatly reduces MS/MS misinterpretations still providing enough information to identify proteins with an accuracy comparable to human expertise. We are now in the process of integrating the Taggor-PepMap modules into the high-throughput pipeline of the French national proteomic platform in Grenoble. The complete final pipeline will incorporate other modules like spectrum qualification (upstream) and protein characterization (downstream).

References

[1] M. Ferro, D. Salvi, S. Brugière, S. Miras, S. Kowalski, M. Louwagie, J. Garin, J. Joyard, and N. Rolland, Proteomics of the chloroplast envelope membranes from Arabidopsis thaliana. *MCP* published May 28, 2003, 10.1074/mcp.M300030-MCP200.

[2] J.A. Taylor, R.S. Johnson, Implementation and uses of automated de novo peptide sequencing by tandem mass spectrometry. *Analytical Chemistry* 73, 2001.

[3] D. Fenyo, Identifying the proteome: software tools. *Curr. Opin. Biotechnol.* 2000 Aug; 11(4) :391-5.

[4] J.R. Yates, III, Mass Spectrometry from genomics to proteomics. *TIG*, vol. 16, n°1, jan. 2000.

[5] D.N. Perkins, D.J. Pappin, D.M. D.M. Creasy, J.S. Cottrell, Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 1999; 20 (18) :3551-356702.

[6] M. Mann, and M. Wilm, Error-Tolerant Identification of Peptides in Sequence Databases by Peptide Sequence Tags. *Analytical Chemistry*, 1994. 66: pp. 4390 - 4399.