PepLine: A Software Pipeline for High-Throughput Direct Mapping of Tandem Mass Spectrometry Data on Genomic Sequences

ARTICLE *in* JOURNAL OF PROTEOME RESEARCH · JUNE 2008

Impact Factor: 4.25 · DOI: 10.1021/pr070415k · Source: PubMed

11 AUTHORS, INCLUDING:

Myriam Ferro
Atomic Energy and Alternative Energies Co…
75 PUBLICATIONS **3,499** CITATIONS

SEE PROFILE

Marianne Tardif
Atomic Energy and Alternative Energies Co…
20 PUBLICATIONS **1,121** CITATIONS

SEE PROFILE

Yves Vandenbrouck
Atomic Energy and Alternative Energies Co…
28 PUBLICATIONS **575** CITATIONS

SEE PROFILE

Jérôme Garin
Atomic Energy and Alternative Energies Co…
114 PUBLICATIONS **6,359** CITATIONS

SEE PROFILE

research articles Journal of proteome research

PepLine: A Software Pipeline for High-Throughput Direct Mapping of Tandem Mass Spectrometry Data on Genomic Sequences

Myriam Ferro,[†,#] Marianne Tardif,[*,†,#] Erwan Reguer,[‡] Romain Cahuzac,[†] Christophe Bruley,[†] Thierry Vermat,[||] Estelle Nugues,[||] Marielle Vigouroux,[||] Yves Vandenbrouck,[||,§] Jérôme Garin,[†] and Alain Viari[*,‡]

CEA, DSV, iRTSV, Laboratoire d'Etude de la Dynamique des Protéomes, Grenoble, F-38054, France, INSERM, U880, Grenoble, F-38054, France, Université Joseph Fourier, Grenoble, F-38054, France, INRIA Grenoble-Rhône-Alpes, Projet Helix, Montbonnot, F-38334, France, GENOME express, Meylan, F-38944, France, and CEA, DSV, iRTSV, Laboratoire Biologie, Informatique et Mathématiques, Grenoble, F-38054, France

Received July 6, 2007

PepLine is a fully automated software which maps MS/MS fragmentation spectra of trypsic peptides to genomic DNA sequences. The approach is based on Peptide Sequence Tags (PSTs) obtained from partial interpretation of QTOF MS/MS spectra (first module). PSTs are then mapped on the six-frame translations of genomic sequences (second module) giving hits. Hits are then clustered to detect potential coding regions (third module). Our work aimed at optimizing the algorithms of each component to allow the whole pipeline to proceed in a fully automated manner using raw nucleic acid sequences (i.e., genomes that have not been "reduced" to a database of ORFs or putative exons sequences). The whole pipeline was tested on controlled MS/MS spectra sets from standard proteins and from *Arabidopsis thaliana* envelope chloroplast samples. Our results demonstrate that PepLine competed with protein database searching softwares and was fast enough to potentially tackle large data sets and/or high size genomes. We also illustrate the potential of this approach for the detection of the intron/exon structure of genes.

Keywords: proteomics • tandem mass spectrometry • Q-TOF • bioinformatics • genome annotation • six-frame translation • peptide sequence tag • gene structure

1. Introduction

Exploiting mass spectrometry data, either as peptide mass fingerprints (PMF) or by using MS/MS data, by matching to protein databases is currently a widely used technique for protein identification.[1] However, because of the makeup of protein databases, some identifications may be missed. Indeed, many proteins that are listed in these databases are deduced from genomic DNA sequences by gene prediction algorithms.[2] Despite considerable advances, *ab initio* gene structure prediction remains computationally difficult, especially for eukaryotic species, and predictions are often erroneous and still far from comprehensive.[3] As a striking example, experimental verification of the *Caenorhabditis elegans* genome revealed that more than 50% of predicted genes need corrections in their exon−intron structures, despite the fact that the exon boundary predictions were usually satisfactory.[4] It is therefore important to complement *ab initio* prediction approaches by experimental data such as ESTs or cDNAs. In this context, MS/MS based protein identification provides direct access to protein amino acid sequences that can further be mapped to genomic sequences. Correlation between experimental data and genomic databases is a powerful way to annotate genomes and discover new genes.

Liquid chromatography coupled with tandem mass spectrometry (LC-MS/MS) is now a well-established technique for protein identification from complex protein mixtures. One of the main advantages of this approach is that it can be easily scaled up for high-throughput experiments.[5–7] Data produced by LC-MS/MS are typically exploited by matching against protein databases in order to identify proteins. There are basically two main approaches for this: the database-driven approach and the *de novo* sequencing approach. In the database-driven approach, using tools such as Mascot (http://www.matrixscience.com),[8] Sequest,[9] X!Tandem[10] (http://www.thegpm.org/TANDEM), and Phenyx (GeneBio), the MS/MS spectra are directly compared to theoretical MS/MS spectra generated from the virtual tryptic digestion of a protein sequence database. The candidate proteins are then scored according to the number and score of matching spectra. This approach does not require any prior time-consuming interpretation of the spectra to be identified. Moreover, since the

[*] To whom correspondence should be addressed. For M.T.: e-mail, Marianne.Tardif@cea.fr (for enquiries regarding mass spectrometry and results); phone, (33) (0)4 38 78 11 14; fax, (33) (0)4 38 78 50 51. For A.V.: e-mail, Alain.Viari@inria.fr (for enquiries regarding algorithm and implementation); phone, (33) (0)4 76 61 54 74; fax, (33) (0)4 76 61 54 08.

[†] CEA, DSV, iRTSV, Laboratoire d'Etude de la Dynamique des Protéomes, INSERM, and Université Joseph Fourier.

[#] These authors contributed equally to the work.

[‡] INRIA Grenoble-Rhône-Alpes.

[||] GENOME express.

[§] CEA, DSV, iRTSV, Laboratoire Biologie, Informatique et Mathématiques.

10.1021/pr070415k CCC: $40.75

Published on Web 03/19/2008

theoretical spectra can be precomputed once and for all, it appears particularly well-suited for quick processing of large data sets. One limitation of this approach is that homologues are not tolerated, since minor changes in the amino acid composition may drastically change spectra. In the *de novo* sequencing approach[11,12] MS/MS spectra are first interpreted using peptide fragmentation rules.[13] The complete or partial peptide sequences can then be subjected to database searching by using sequence-oriented software such as Blast,[14] MS-Blast,[15] or MS-Pattern (http://prospector.ucsf.edu/). The main advantage of this approach is that it can easily handle mutations between the actual peptide and the proteins of the database and, therefore, cope with protein homologues. However, the sequence interpretation process is still difficult to perform automatically. This is due to the fact that only a limited portion of the spectrum can usually be interpreted accurately. An intermediate solution, known as "sequence tagging", has been proposed by Mann and Wilm.[16] In this approach, Peptide Sequence Tags (PSTs) are generated from MS/MS spectra. A PST is defined by a small sequence tag (usually 3 or 4 amino acids) and the two flanking (N- and C-terminal) masses. The database scanning selectivity is lower than that in either the database-driven or *de novo* approaches, but from the algorithmic point of view, it becomes much easier to interpret spectra automatically.[17,18] Two programs have been proposed to automatically generate PSTs from MS/MS spectra: GutenTag[18] and PepNovoTag.[17] Both programs will be discussed hereafter and compared to the one proposed in our pipeline, Taggor.

In the context of genome annotation, protein identifications obtained by the database-driven approach have previously been used to ascertain gene predictions by mapping Sequest-identified peptides on genomic sequences.[19] Other MS-based approaches aimed not only to confirm gene annotation, but also to refine gene structure predictions or find new genes. These approaches typically performed direct searches on genomic sequences, bypassing the large pool of hypothetical proteins that could be deduced using prediction tools and gene building algorithms.

A few studies demonstrated the use of peptide mass fingerprints (PMF), obtained from MALDI-MS experiments, to identify coding regions.[20,21] In these studies, the genome was translated according to the six reading frames and PMFs were mapped on this translated genome with respect to its *in silico* digest. Although this method proved useful in confirming predicted exons and showed its capability in finding new coding regions, it was generally aimed at identifying the genomic origins of a low number of proteins (contained in 2D-gel spots) rather than performing large-scale annotation. This strategy was assessed using species with relatively simple genomes (*Saccharomyces cerevisiae*, *Escherichia coli*, *Pseudomonas aeruginosa*, *Mycobacterium tuberculosis*). Further studies made use of MS/MS data in order to mine genomic sequences directly, by scanning the six-frame translation products of the genome with a database-driven search tool. A first approach consisted in using Mascot with a database made by translating 600 kb segments of the human genome.[22] Similarly, Kalume et al. used Mascot with a database made from translation of 100 kb segments to annotate the genome of *Anopheles gambiae*.[23] These approaches were not straightforward because database-driven tools like Mascot cannot deal with large nucleic acid sequences, such as whole translated chromosomes. Thus, such approaches required segmentation of the genome and were used in complement to protein and ESTs database searches. Instead of performing prior segmentation of the genome, other approaches built a more realistic representation of the genome by first generating all putative ORFs (possibly of a minimum length) from each complete chromosome and then scanning the corresponding protein database. This procedure is straightforward with prokaryotic genomes. It has recently been used to annotate two bacteria of the *Phytophtora* genus[24] and to identify novel genes that were missed by gene prediction for these organisms. This approach was also undertaken for the complete human genome[25] where the situation is more complex due to the exon/intron structure and the size of the genome. More than 217 million putative ORFs were generated from the complete human genome. These ORFs were used as the database in a search for novel blood proteins using the data from HUPO-Plasma and the X!Tandem search engine. This study allowed the detection of new exons within known coding regions.

A major limitation of the methods described above, both the PMF-based and the uninterpreted MS/MS spectra-based methods, is that the entire tryptic peptide must match a continuous genomic segment. This requirement precludes the possibility of matching peptide masses (in PMF) or MS/MS spectra for peptides that are split over an intron/exon boundary. By contrast, strategies applied to genomic sequences using peptide *de novo* sequencing allowed gaps in matching tryptic peptides and, therefore, may be used to ascertain exon/intron boundaries. In this context Hippler's group developped an error-tolerant search algorithm specifically dedicated to the detection of intron-split peptides (GenomicPeptideFinder, GPF).[26] The authors used a *de novo* sequencing software (deNovoX, Thermo Finnigan) to retrieve peptide sequences from MS/MS spectra which were mapped on the six-frame translation of a genome. This specific approach for assignment of intron-split peptides was assessed on the *Chlamydomonas reinhardtii* EST database and was combined with a Sequest search in order to achieve exhaustive annotation.[27] As mentioned above, an alternative to the complete *de novo* approach is to restrict the spectrum interpretation to peptide sequence tags (PSTs). This was first suggested by Kuster et al.[28] In this approach, the sequence tag of the PST was translated into a corresponding degenerated oligonucleotide sequence. Exact matches of this oligonucleotide on the forward and reverse frames of the *Arabidopsis thaliana* and human genomes were then retrieved, and the flanking masses were calculated and compared to the flanking masses of the PST. This approach proved to be helpful in finding new intron/exon boundaries and new genes. However, PST generation and analysis of the results were mainly performed manually, therefore, ruling out its use with high-throughput experiments. A first step toward high-throughput analysis of very large data sets with a PST approach was performed by MacGowan et al.[29] using an exhaustive set of predicted transcripts instead of human chromosomal sequences in order to reduce the number of spurious intergenic or intron hits. More recently, the PST approach was applied to a large data set composed of 18 million MS/MS spectra from human proteomic samples using PSTs of three amino acids or more[30] and a modified version of the Inspect search algorithm.[31] Again, instead of searching the translated human genome directly, the authors made use of a compact graph representation of all putative exons and splice variants.

In this paper, we describe and evaluate a PST approach similar to those described by Kuster et al., McGowan et al.,[28,29] and Tanner et al.[30] We aimed at developping a fully automated

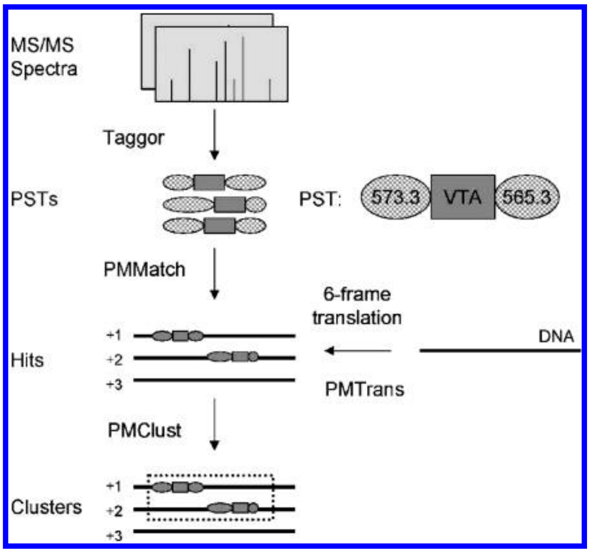

Figure 1. Overview of the PepLine pipeline. PSTs are generated from MS/MS peak lists using the Taggor module. A PST is defined by a small sequence tag (3 amino acids) and the two flanking (N- and C-terminal) masses. PSTs are then mapped on the six-frame translations (PMTrans module) of a DNA sequence using the PMMatch module. Hits belonging to the same DNA strand (eukaryotes) or optionally to the same DNA frame (prokaryotes) are then clustered in order to determine the location of potential genes (PMCluster module).

and simple procedure, with the potential for high-throughput analyses without the requirement of performing any prior interpretation of the genomic sequence. In addition, while not strictly devoted to matching intron-split peptides, our approach allows the detection of exon−intron boundaries and thus is well-adapted to eukaryotic gene structure. Briefly, our software called PepLine allows PSTs to be generated automatically and subsequently mapped and clustered on the six-frame translation of DNA sequences. PepLine is made of three modular components, each of which were algorithmically optimized to work efficiently on very large data sets and can be easily embedded into a software pipeline. As shown in Figure 1, the first component, Taggor, performs the MS/MS spectra interpretation to produce PSTs and was designed to specifically handle QTOF MS/MS data. The second one, PMMatch, maps these PSTs on large DNA genomic sequences and yields sequence hits. Finally, PMClust performs the clustering of closely located hits. We present the three components and give some illustrations of their application and performance.

2. Materials and Methods

2.1. Taggor: from Spectra to PSTs. In contrast to already existing PST generation algorithms[17,18] that will be discussed below, Taggor adopts a "brute-force" approach to generate PSTs from tandem mass spectra. Indeed, since we are interested in quite small sequence tags (length $k = 3$ or 4 amino acids) the complete enumeration of 20^3 or 20^4 possible sequences does not represent a computational difficulty and is achieved within less than a second. For each sequence tag, Taggor then recurses on the list of peaks in the spectrum to determine all the combinations of $(k + 1)$ peaks that are compatible with this particular sequence. Being compatible means that the mass difference between two consecutive peaks equals the corresponding amino acid mass, within a given error tolerance threshold. In its current form, Taggor specifically handles QTOF MS/MS data. Indeed, to generate PSTs, Taggor assumes that the fragmentation corresponds to the y ions series, since y ions are the major peaks in QTOF MS/MS spectra from tryptic peptides.

A given combination of $k + 1$ peaks (corresponding to the sequence tag of a PST) is scored according to the following formula:

$$s = \exp\left(-\left(\sum_{i=0}^{k} r_i - \beta\right)/\alpha\right) \qquad (1)$$

where r_i is the intensity rank of the ith peak in the combination. "Intensity rank" means that the peaks are first sorted according to their intensity and the most intense peak is ranked 1, the second 2, and so forth. The rationale for this score comes from the fact that it has been observed[32] and we have experimentally verified (data not shown) that the fraction of y ions in a spectrum is an exponentially decreasing function of the intensity rank. In other terms, the probability of an ion of rank r to be Y can be written as as $P(Y/r) \propto \exp(-r)$. The β constant is simply computed to yield a score $s = 1$ when using the $(k + 1)$ first intensity ranked peaks of the spectrum, that is, $\beta = (k + 1)(k + 2)/2$. The α parameter depends upon the experimental setup and is better expressed as the minimum rank r_0 that yields a score of 0.5, that is, $\alpha = kr_0/\ln(2)$. For QTOF MS/MS spectra, r_0 was set to 10.

Taggor does not perform genuine *de novo* interpretation of the whole spectrum. This is good if the noise-level of the spectrum is high, that is, if it is difficult to assign a single peptide to the spectrum. On the other hand, if the spectrum is clean and easily interpretable, Taggor will produce several independent but overlapping PSTs and will not take advantage of this fact. To benefit from overlapping PSTs, we introduced an optional second step, where overlapping PSTs are linked together in order to build *de novo* chains. A chain is simply a list of overlapping PSTs, and chains are ranked according to their length. At this point, the score of a PST is corrected using the following formula:

$$s' = s\exp((1 - R)/R_0) \qquad (2)$$

where R is the rank of the longest chain containing the PST and R_0 is a parameter with the following interpretation: if R_0 is large ($R_0 \to \infty$) then the correction factor tends to 1 whatever the chain and there is no *de novo* correction of the initial PST score ($s' \to s$); if R_0 is small ($R_0 \to 1$) then the correction factor is maximum and will penalize PSTs not included in long chains. The default value for R_0 is 1, assuming good quality spectra.

2.2. PMMatch: from PSTs to Genome Hits. The second step of the pipeline consists in locating the PSTs generated by Taggor on translated complete chromosomes. Each PST is mapped onto the six-frame translations of a chromosome (by using the PMTrans utility, Figure 1) and gives rise to hits. A hit is defined as the location of a PST on the translated chromosome. One PST may give rise to several hits at different locations on the chromosome. Conversely, several hits may be associated with the same location (same peptide sequence) on the translated chromosome. For this mapping step, when the value of k (the sequence tag length) remains small (typically $k = 3$ or 4), an efficient and simple way of matching the sequence tag of the PST is to use a bucket indexing technique. A bucket array of size $|\Sigma|^k$ (where $|\Sigma| = 20$ stands for the size of the alphabet) stores the observed PSTs, indexed on their k-amino acid sequence. Then each translated frame of the whole chromosome (of size N) can be scanned in one pass therefore yielding an average complexity of $O((Nn/|\Sigma|^k) + |\Sigma|^k)$, where n is the total number of PSTs to map. The term $n/|\Sigma|^k$ corresponds to the average filling factor of a bucket. In practice, n ($\sim$1000 spectra) is generally smaller than $|\Sigma|^k$ (8000 for $k = 3$) and this factor is therefore less than 1.

When the sequence part of a PST has been mapped, the corresponding frame of the translated chromosome is scanned to the left and to the right in order to match the N- and C-terminal masses of the PST against the actual amino acid sequence. The scan process stops when the cumulated mass matches, or exceeds, the required mass or when a stop codon is encountered. At this point, PMMatch checks if the sequence ends match (i) the enzyme cleavage pattern (internal hit), (ii) a methionine (possible N-terminal hit), or (iii) a stop codon (C-terminal hit). PMMatch also checks if there is no enzyme cleavage pattern within the peptide sequence (miscleavage) and discards the hit if too many miscleavages are observed (the maximum number of miscleavages can be set by the user). PMMatch can optionally handle amino acid modifications during the scan process. One of the main differences between protein and genome scanning lies in the fact that, in the former, peptides can be interrupted by introns. Therefore, PMMatch optionally also allows partial hits, that are hits where the sequence tag and only either the N- or C-terminal mass of the PST are matched.

2.3. PMClust: from Hits to Clusters. The last step of the PepLine procedure consists in clustering the hits in order to propose regions potentially associated with genes or, at least, with exons. In the current version of PepLine, we make use of a simple single-linkage approach.[33] This method has the advantage of being quick (time complexity is linear with the number of hits being clustered) and simple, since it depends only on one intuitive parameter, δ, which is the maximal distance between two consecutive hits in a cluster. It suffers from the well-known "chain effect" that makes it sensitive to the δ parameter. Some experiments are therefore necessary to tune this parameter for each species under study. A typical value of δ is correlated with the mean length of introns. For species where the variance of intron length is small (e.g., *A. thaliana*) this procedure gives satisfactory results and is not very sensitive to the δ parameter (see Results and Discussion). Hits are clustered separately on the direct and reverse DNA strands. Optionally (namely for prokaryotes) hits can also be clustered frame by frame. To minimize the chain effect, PMClust ignores partial hits when constructing the clusters. Optionally, partial hits can be added in clusters after their construction. In the current version, PMClust reports all clusters, including single hit clusters. However, to limit the size of the output, users can specify a minimum threshold in terms of number of hits or of different peptides a cluster should contain in order to be printed out.

2.4. Sample Preparation, MS Analyses, and Peptide Identification. To evaluate each components of PepLine, two different sets of samples were prepared.

The first set of samples was composed of 10 standard proteins, all purchased from Sigma: Lysozyme (P00698), Trypsinogen Beta (P00760), Lactoglobulin (P02754), Carbonic Anhydrase (P00921), Pepsin (P00791), Ovalbumin (Q804A4), Albumin (P02769) Phosphorylase B (P00489), Beta-Galactosidase (P00722), and Myosin (Q28641).

Each protein standard (10 μg) was first loaded on a 12% acrylamide gel for SDS-PAGE analysis. After SDS-PAGE, a discrete band was excised from the Coomassie blue-stained gel. For each protein, the in-gel digestion was carried out as previously described.[34] Peptides were then extracted from gel pieces with 5% (v/v) formic acid solution and acetonitrile. After drying, tryptic peptides were resuspended in 0.5% aqueous trifluoroacetic acid. The samples were injected into a LC-Packings (Dionex) nanoLC system where they were preconcentrated on a 300 μm × 5 mm PepMap C18 precolumn. The peptides were then eluted onto a C18 column (75 μm × 150 mm). The chromatographic separation used a gradient from solution A (5% acetonitrile, 95% water, and 0.1% formic acid) to solution B (5% water, 95% acetonitrile, and 0.1% formic acid) over 60 min at a flow rate of 200 nL/min. The LC system was directly coupled with a QTOF Ultima mass spectrometer (Waters). MS and MS/MS data were acquired and processed automatically using MassLynx 4.0 software.

The second sample was a genuine biological sample: a mixture of proteins extracted from *A. thaliana* chloroplast envelope. The complete extraction and preparation procedure has been described elsewhere.[35] Briefly, chloroplast envelope proteins were extracted from *A. thaliana* leaves using two procedures to wash the chloroplast envelope (NaOH and NaCl treatments). Samples were loaded onto a SDS-PAGE gel. Migration was stopped when the migration front was localized between the stacking and the separating gels to concentrate the protein sample in a single gel band. The gel band was further treated as described above.

MS/MS data of both sets of samples were processed using the MaxEnt3 (MassLynx) procedure with the following parameters: QA threshold, 10; ensemble number, 1; iterations, 80. For the *A. thaliana* data, the processed files corresponding to the two samples considered in the present study were concatenated for subsequent database searching. All MS/MS spectra from both sets were tentatively assigned to peptides using Mascot 2.0 (http://www.matrixscience.com) and the UniProt database (http://www.expasy.uniprot.org, release 6.0). For both the standard and the *A. thaliana* sets of proteins, sulfone, cysteic acid, oxidized methionine and *N*-acetyl (protein) were used as variable modifications. Peptides showing a score higher than 40 were validated without any manual examination. Peptides with a score lower than 40 were systematically checked and/or interpreted manually to confirm or dismiss the Mascot suggestion.

For each sample set, all MS/MS spectra that gave indubitable peptide identification were extracted and clustered into a single peak list file. Spectra that could not be manually assigned to a peptide were discarded. For the standard set of proteins, this resulted in 203 assigned spectra, 78 of them bearing a modification, corresponding to 200 different peptides. For the *A. thaliana* set, this resulted in 291 assigned spectra, 40 of them bearing a modification, corresponding to 217 different peptides (Table 1).

3. Results and Discussion

3.1. Taggor Evaluation. Taggor was run (with an error tolerance of 50 ppm) on both standard and *A. thaliana* sample sets, and the resulting PSTs were evaluated by comparison to the validated peptide sequences. The sequence tag of a PST was correct if it was present in the validated peptide sequence disregarding any distinction between "I" and "L" and between "K" and "Q". As a whole, a PST was considered to be correct if both its sequence tag and its flanking masses were correct. Results are given in Table 1 for two settings of Taggor

Table 1. Comparison of Taggor and PepNovo Results Using Known MS/MS Spectra from Standard and *A. thaliana* Protein Digests

	sample set Standard			sample set *A. thaliana*		
	no. spectra = 203			no. spectra = 291		
	no. peptides = 200			no. peptides = 217		
program	Taggor[a]		PepNovo[b]	Taggor[a]		PepNovo[b]
Parameters	$r = 10$ $s = 0$	$r = 5$ $s = 0.5$	$r = 10$ $s = 0$	$r = 10$ $s = 0$	$r = 5$ $s = 0.5$	$r = 10$ $s = 0$
no. PSTs	1998	615	2020	2797	894	2881
no. correct PSTs	915 (45.8%)	474 (77.1%)	932 (46.1%)	1139 (40.7%)	658 (73.6%)	1008 (35.0%)
no. wrong PSTs	1083	141	1088	1658	236	1873
no. spectra with at least one correct PST	171 (84.2%)	143 (70.4%)	178 (87.7%)	233 (80.1%)	196 (67.4%)	225 (77.3%)
no. identified peptides	169 (84.5%)	143 (71.5%)	175 (87.5%)	183 (84.3%)	155 (71.4%)	176 (81.1%)

[a] Evaluation of PSTs at 200 ppm. [b] PepNovo -tag_length = 3, evaluation of PSTs at 2000 ppm.

parameters '*r*' and '*s*'. The '*r*' parameter corresponds to the maximum number of PSTs produced per spectrum, and the '*s*' parameter is the minimum score allowed for a PST.

It is notable that the results obtained for the standard proteins set were better than those for the *A. thaliana* set. This can be explained by the higher quality of the whole set of spectra in the first set, due to individual analysis of proteins with controlled sample amounts and careful selection of validated peptides. On the other hand, the spectra in the second set, which is more representative of a genuine biological experiment, displayed a wider range of qualities. The more stringent Taggor parameters ($r = 5$, $s = 0.5$) gave better results in terms of selectivity in both cases (77% versus 46% correct PSTs for the standard set and 74% versus 41% for the *A. thaliana* set). This must be balanced against the loss of sensitivity observed: one-third to one-half of correct PSTs was lost. However, when using the most stringent parameters, we have found that this loss was not critical since most PSTs were redundant with regard to the corresponding peptide, as indicated by the number of spectra with at least one correct PST and the number of correctly identified peptides (Table 1). Neither numbers were significantly affected by the Taggor settings. As expected (and as we shall see later), false PSTs are less likely to yield correct hits on the chromosome and to eventually get clustered. Therefore, when Taggor is used in the context of a full pipeline, these settings are not critical. However, if Taggor is to be used for the sole purpose of providing PSTs, then it is necessary to look more carefully at the distribution of correct PSTs as a function of their rank and score. These distributions are shown for both sets in Figure 2. The number of correct PSTs was maximal at early ranks and dropped gradually with rank (Figure 2a,b). Few PSTs were generally gained after rank 6, whereas the number of false PSTs continued to increase. In terms of score, the distribution of correct PSTs was maximal between 0.5 and 0.6, whereas the distribution of false PSTs was maximal at 0 and decreased sharply with score (Figure 2c,d). Therefore, the settings $r = 5$, $s = 0.5$ seemed to give a good compromise between sensitivity and selectivity.

Closer examination of the origin of false PSTs showed that most of them were incorrect because their sequence tag was erroneous (between 80 and 90% of total wrong PSTs). In some cases (about 5% of total wrong PSTs), sequence tags were anchored on b-ions series, resulting in reverse-oriented sequences. This occurs for peptides bearing a highly basic amino acid such as Arg, Lys, or His at their N-terminal end. These peptides produce prominent b-ions series, which is misleading for Taggor. These mirror tags were not very frequent, suggesting that most of the wrong sequence tags were anchored on peaks not belonging to either y- or b-series. Indeed, 40% of total false PSTs generated with the loose Taggor setting ($r = 10$, $s = 0$) had a score below 0.1, suggesting that they were picked up on low-intensity peaks. These PSTs might have been generated either from noise or from ions belonging to minor series in CID fragmentation such as "a" ions or internal fragment ions. The rate of false, low-score PSTs can be controlled by tuning the "*r*" and "*s*" parameters.

Finally, only a small fraction of the PSTs (about 5% of total wrong PSTs) displayed a correct tag but wrong masses. Most often, the N-terminal mass was wrong. This is due to the fact that the calculation of the N-terminal mass depends on the parent ion mass which, in our case, is extracted from the MS/MS data file. For instance, some N-terminal masses may be incorrect due to fragmentation of the second isotope of a peptide ion. This was the case for 5% of the spectra from both the standard and the *Arabidopsis* data sets.

The objective of Taggor is very close to those of GutenTag[18] and PepNovoTag.[17] Besides different algorithmic approaches, the main difference is that both GutenTag and PepNovoTag make use of a quantitative fragmentation model to score the PSTs, whereas Taggor is based on a much simpler scoring scheme. However, Taggor yielded very similar results when compared to PepNovoTag, as shown in Table 1. Moreover, with the ($r = 10$, $s = 0$) parameters, Taggor and PepNovoTag generated 498 and 654 common (i.e., identical) correct PSTs for the standard and the *A. thaliana* data sets, respectively. This means that more than 50% of correct PSTs are generated by both programs. As for wrong PSTs, Taggor and PepNovoTag generated 67 and 144 common wrong PSTs for the standard and the *A. thaliana* data sets, respectively. This represents less than 10% of common wrong PSTs. Thus, the two programs tend to find the same correct PSTs but make different errors. Consequently, it is not advisable to merge their results, since the increase in the total number of correct PSTs would be accompanied by a higher increase in false-positives. Despite our efforts, we did not manage to get satisfactory results on our data sets when running GutenTag. This is probably due to the fragmentation model implemented in GutenTag that may not be adapted for use on QTOF MS/MS data. Indeed, GutenTag has been optimized for Ion-trap data where b- and y-ions are often equally present and makes use of b-ions as corroborative information. In contrast, the fragmentation model of PepNovoTag appears to be of larger scope and seems to fit equally well with QTOF and Ion-trap data.

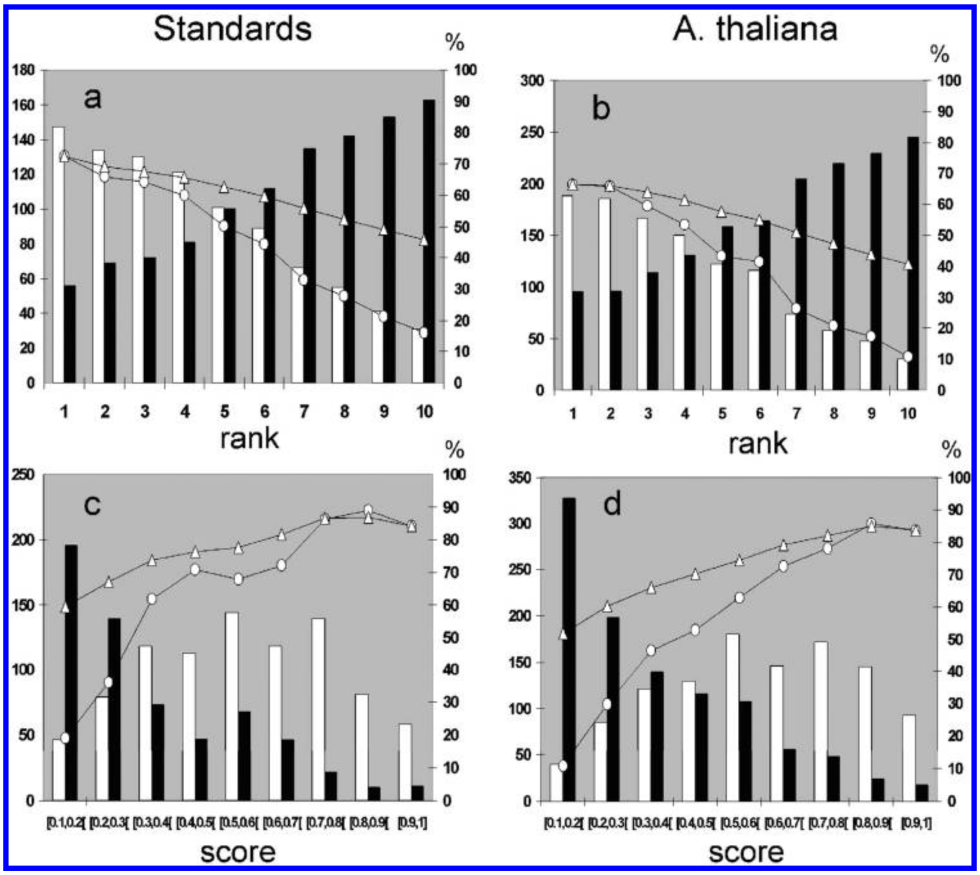

Figure 2. Distribution of correct PSTs and incorrect PSTs generated by Taggor. MS/MS data for MaxEnt3-processing were acquired on a Q-TOF instrument from analysis of protein standards (a and c) or *A. thaliana* sample (b and d). Taggor parameters were defined such that each MS/MS spectrum produced a maximum of 10 PSTs (ranked from 1 to 10, rank 1 corresponding to the highest scoring PST in the spectrum). The minimum score allowed was set to 0. The top panels (a and b) display the number of correct (white bars) and wrong PSTs (black bars) as a function of PST rank. The bottom panels (c and d) display the number of correct (white bars) and wrong (black bars) PSTs as a function of PST score. Only PSTs with a score greater than 0.1 are shown (numbers of correct/wrong PSTs in the [0, 0.1] score category were 19/472 for the standard set and 29/625 for the *A. thaliana* set). The superimposed curves represent the percentage of correct PSTs within each category of rank or score (circles) or the cumulated percentage of correct PSTs (triangles) starting from the best rank (rank 1) or best scores ([0.9, 1]). The cumulative curves depict the actual percentage of correct PSTs when generated with Taggor parameters '*r*' or '*s*' set at the corresponding abscissa.

Table 2. Evaluation of PMMatch results using known MS/MS spectra from *Arabidopsis thaliana* protein digests and PSTs sets from previous Taggor evaluation

Taggor settings	$r = 10$ $s = 0$		$r = 5$ $s = 0.5$	
no. Hits	1632		755	
no. correct Hits	784 (48.0%)		487 (64.5%)	
	from PST correct	from PST wrong	from PST correct	from PST wrong
	784 (100%)	0 (0%)	487 (100%)	0 (0%)
no. wrong Hits	848 (52.0%)		268 (35.5%)	
	from PST correct	from PST wrong	from PST correct	from PST wrong
	576 (68.0%)	272 (32%)	191 (71.3%)	77 (28.7%)

3.2. PMMatch Evaluation. The second step of the pipeline, that is, the mapping of PSTs on whole chromosomes using PMMatch, was tested using the *A. thaliana* MS/MS data set and all *A. thaliana* chromosomes (5 nuclear, 1 mitochondrial, and 1 chloroplastic). Chromosomes were retrieved from TAIR, database version 6 (Nov. 11, 2006). We used the PSTs generated by Taggor in the previous experiments, that is, with the two different Taggor settings given in Table 1. PMMatch was used with default parameters: mass tolerance 50 ppm, 1 maximum miscleavage, and no modified residues allowed. At this stage, only complete hits were considered. A hit was considered "correct" if the corresponding peptide on the translated chromosome was identical to the manually identified peptide, with no distinction between "I" and "L" or between "K" and "Q". Otherwise, it was labeled "wrong". This does not necessarily mean that the hit corresponds to the correct chromosome location (or even that there is a protein-coding gene at this location), but just that there is accordance between both the spectrum and the genome translated sequence. Results in Table 2 show that about half of the hits were correct. As expected, the tighter Taggor parameters ($r = 5$; $s = 0.5$) yielded less hits but a better ratio of correct hits (64%). This table also indicates whether the correct and wrong hits originated from correct or wrong PSTs. Interestingly, although correct hits originated only from correct PSTs, wrong hits originated both from wrong and correct PSTs. Indeed, about 70% of wrong hits were actually produced by correct PSTs (Table 2). This stems from the fact that, even at a mass tolerance of 50 ppm, several different peptides can theoretically match the same PST. In other terms, even if the PST generation step was perfect (i.e., no wrong PST at all) one cannot expect 100% of hits to be correct but, as in this case, between 60 and 70%. Correlatively, another important feature in Table 2 was the low proportion of hits generated by wrong PSTs (all of these hits being wrong). Indeed, with the less stringent Taggor parameters ($r = 10$; $s = 0$), the 1658 wrong PSTs (Table 1) only gave rise to 272 hits, the great majority of wrong PSTs (95%) leading to no hit at all. On average, the number of hits per PST was around 0.2−0.3 for wrong PSTs, whereas it went up to 1.0 for correct PSTs. There is therefore a "cleaning" effect of the chromosome mapping step that tends to "remove" wrong PSTs. For this reason, and in order to improve the coverage of the proteome under study, and subsequently increase the chances of identifying true coding regions, it may be advantageous to use the less stringent Taggor parameters ($r = 10$; $s = 0$). In the previous analysis, we have considered hits, but there is some obvious redundancy in the results since several hits may actually correspond to the same peptide on the translated chromosome. Indeed, of the 217 different peptides manually identified in the *A. thaliana* set, 188 bear no modification and could therefore be potentially located on the chromosome. The 784 correct hits reported for ($r = 10$; $s = 0$) correspond to 120 different peptides and the 487 correct hits obtained for ($r = 5$; $s = 0.5$) correspond to 101 different peptides. This means that 64% and 54% of known peptides could be recovered by PMMatch.

3.3. PMClust Evaluation. Finally, the last step of the pipeline consists in gathering the hits into clusters. This clustering step is based on the idea that the structure and dispersion of genes in genomes should be reflected by a statistically significant aggregation of hits. For this purpose, we made use of the 1632 hits generated with the ($r = 10$; $s = 0$) Taggor parameters described above. The PMClust program (Figure 1) was used with a clustering distance δ of 3000 nucleotides (default parameter), and all clusters were returned, including those composed of only one hit. Each cluster was then labeled "correct" if it contained a majority (i.e., ≥ 50%) of "correct" hits as previously defined (i.e., if the corresponding peptide on the translated chromosome was identical to the manually identified peptide). This allowed us to study the distribution of the number of hits per cluster for correct and wrong clusters (Figure 3, white and black bars). This figure clearly shows that a majority of clusters with only 1 or 2 hits were wrong. The proportion was inverted for clusters of 3 hits or more and, with >4 hits/cluster, almost all clusters were correct. Therefore, for the remainder of this analysis, we have set a minimum threshold of 3 hits for a cluster to be retained. The 1632 hits generated by PMMatch (Table 2) gave rise to 85 different clusters, comprising 764 hits. Of these clusters, 78 (92%) were correct, according to the previous definition, and 7 were wrong. Furthermore, a high proportion (97%) of wrong hits, 819 out of total 848, were not clustered. Conversely, a high proportion (94%) of correct hits, 735 out of total 784, were integrated in clusters, all of which were correct clusters. These distributions demonstrated an additional "cleaning" effect of the clustering step, provided that a hit threshold of 3 was used. To assess the frequency of random mapping, PSTs were mapped on a shuffled version of the *A. thaliana* genome. The six-frame translations of the chromosomes were shuffled, keeping the same amino acid composition. The positions of the stop codons have not been shuffled to roughly keep the same putative gene lengths. Over 98% of the hits generated were wrong (1050 total, 1038 wrong). Clusters were computed, and almost all of them were not retained as correct, according to the criteria described above. The distribution of the number of hits per cluster is shown in Figure 3 (gray bars). Finally, only 16 clusters exhibited at least 3 hits (representing a total of 60 hits). This experiment demonstrated that, despite the high number of random hits produced, these random hits are efficiently eliminated during the clustering step.

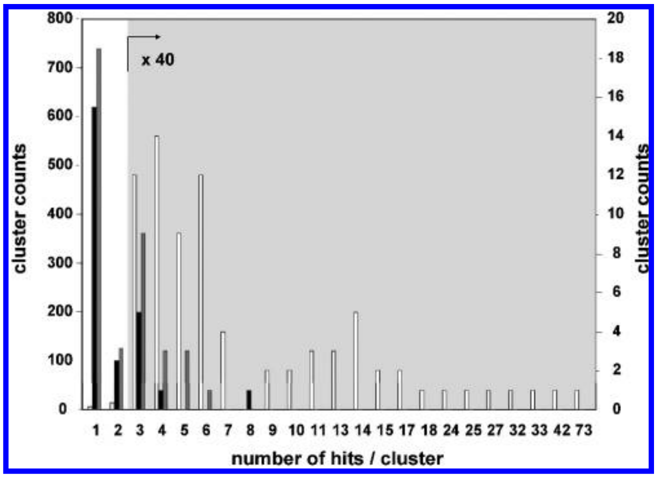

Figure 3. Distribution of the number of hits in correct and wrong clusters. The PSTs generated from the *A. thaliana* set were mapped (PMMatch) on the seven chromosomes of *A. thaliana* (5 nuclear, 1 mitochondrial, and 1 chloroplastic), and hits were further clustered (PMClust) according to their location. The distribution of the number of hits per cluster for both the "correct"clusters (white bars) and the "wrong" clusters (black bars) is shown. In addition, the figure displays the distribution of the number of hits per cluster obtained on the shuffled six-frame translations of the *A. thaliana* genome (gray bars). *Y*-axis on the left side indicates the cluster counts for the white area of the figure (clusters with only one or two hits), while *y*-axis on the right side indicates the cluster counts for the gray area of the figure (clusters with three hits or more). The figure shows that almost all clusters on the "real" genome containing more than 3 hits were actually correct. For clarity, "correct" and "wrong" clusters on the "shuffled" genome were not distinguished because almost all of them were wrong.

Table 3. Evaluation of PepLine Clusters by Comparison with TAIR Annotations and with Mascot Identifications

annotation status	no. clusters	no. correct clusters	no. wrong clusters	% correct clusters
Unannotated[a]	11	6 (2)	5	54.5
Annotated[a]	72	72	0	100.0
Identified[b]	2	0	2	0.0
Not identified[b]	85	78	7	91.8

[a] Clusters spanning, or not, a gene annotation in TAIR. Annotated: the cluster spans a single gene annotation. Unannotated: the clusters spans 0 or (in parenthesis) more than one gene in TAIR. [b] Clusters associated, or not, with proteins identified in Ferro et al.[35]

Finally, to assess the efficiency of PepLine on real data sets, we compared PepLine results to the list of proteins which were originally identified from the 291 *Arabidopsis* spectra. For this purpose, as a first step, we cross-correlated the clusters with the TAIR annotations. A cluster that spans a single gene location

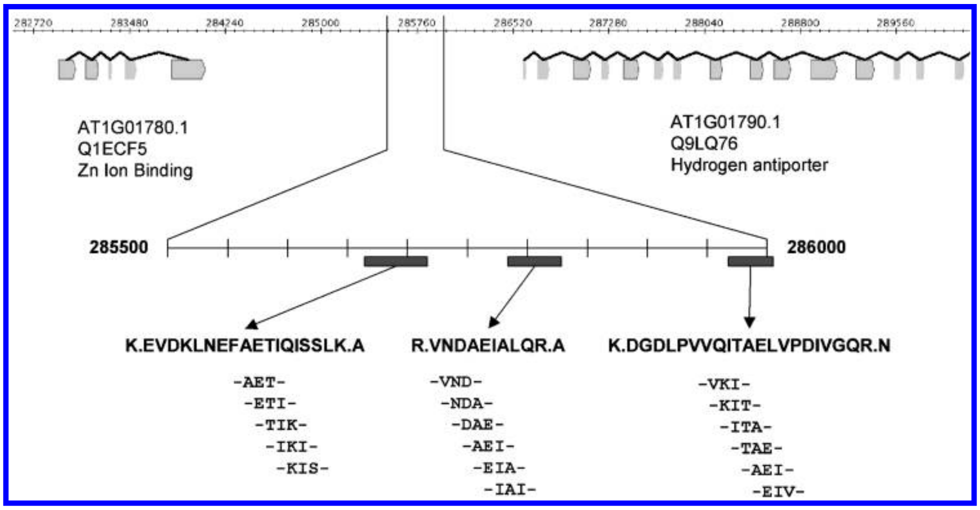

Figure 4. Evidence of an unannotated gene in the TAIR database. The zoomed part of the figure (center) displays an unannotated region of chromosome I, located between genes AT1G01780.1 and AT1G01790.1, where a cluster of three different peptides was found. This cluster indicated the position of a coding region on chromosome I. When investigating other databases, this cluster was associated with a predicted intronless gene, called T1N6.20, on a BAC of *A. thaliana* and referenced as Q9LQ77 in the UniProt database.

in TAIR was labeled "Annotated", otherwise it was labeled "Unannotated" (in most cases, this means that there is no TAIR annotation at all at the location of the cluster). Second, for the "Annotated" clusters, we checked if the TAIR annotation corresponded to a gene that encodes one of the proteins identified in Ferro et al.[35] Then, clusters were labeled "Identified" or "Not Identified" respectively. Results are given in Table 3 and show that almost all of the "correct" clusters (72 out of 78) were indeed associated with proteins annotated in TAIR. Interestingly, all of them had been previously identified in Ferro et al.[35] Considering the fact that 105 different proteins were identified in Ferro et al.,[35] the automated pipeline allowed identification of 69% of the manually identified proteins. More interestingly, we also found 11 clusters, including 6 "correct" clusters, not associated with any TAIR annotation (Table 3). This strongly suggests that these clusters correspond to true, unannotated genes.

3.4. Running PepLine on Unbiased Data Set and Illustration. As mentioned in Materials and Methods, the *A. thaliana* data set was actually a selection of spectra for which a peptide could be manually assigned. In some ways, this may create a bias toward higher quality spectra and hence influence the overall results. To evaluate our pipeline on more realistic data, we reran it using all default parameters on the whole (i.e., unselected) spectra used in Ferro et al.[35] This corresponded to 542 spectra, including the 291 spectra that were manually assigned. Of course, in this case, we could not evaluate the number of "correct" clusters but we could still evaluate how many of them corresponded to a single TAIR annotation. The pipeline yielded 109 clusters, of which 89 (82%) were associated with a single TAIR annotation showing that PepLine is also suitable for analysis of more noisy data sets.

Twenty clusters were found without any TAIR annotation, including the 6 correct clusters mentioned above. Four clusters were classified as wrong clusters. Two clusters actually spanned more than one TAIR annotation and were therefore mistakenly labeled "Unannotated" because they do not span a single gene. Careful examination of the 14 remaining clusters revealed that 12 of them displayed only one peptide. For nine of these 12 clusters, the corresponding peptides were also found in correct "Annotated" clusters (identified with more peptides). These nine "Unannotated" clusters are likely to be false positives. The three additional clusters identified with only one peptide corresponded to contaminants (trypsin and keratin). The remaining two "Unannotated" clusters corresponded to genuine new annotations. One of these clusters with three different peptides is shown in Figure 4. This cluster located to chromosome I between genes AT1G01780.1 and AT1G01790.1, but no associated gene prediction was found in either TAIR or TIGR databases, nor in the EMBL or GenBank *A. thaliana* chromosome I annotations. This cluster suggested the position of a new coding region on chromosome I. While investigating other databases, this cluster was associated with a predicted intronless gene, called T1N6.20, in a BAC of *A. thaliana*, also referenced as Q9LQ77 in the UniProt protein database.

In addition to complete hits, partial hits could also be considered as additional information in order to refine intron/exon boundaries. As explained before, these partial hits consisted in matching all the PST except one of its two flanking masses. This is illustrated in Figure 5 by a cluster corresponding to the At1g06950 gene located on chromosome I which codes for a protein involved in protein import into the chloroplast. Partial hits belonging to this cluster revealed two introns also annotated in the TAIR database at 1925–2002 (nt) and at 2656–2728 (nt). Considering the first intron, the partial hits allowed the exact intron/exon boundaries to be defined. For the second intron, since there was an overlap between the amino acid sequences covered by partial hits on frame 3 and hits on frame 1, we were not able to define its exact boundaries (Figure 5a). However, through manual examination, and by taking into account the GT-AG consensus for intron site, it was possible to predict a precise limit where the codon for the GLY885 residue is shared on both frames. This prediction is in accordance with the TAIR annotation.

Thus, by setting the "partial hit" optional function of PepLine, preliminary results obtained from the chloroplast envelope data set indicated that intron−exon boundaries could be efficiently identified, emphazising the complementarity of our PepLine approach with *ab initio* methods to improve gene structure annotation in higher eukaryote genomes.

3.5. Running Times and Availability. Finally, we would like to stress the fact that, in terms of computational efficiency, the

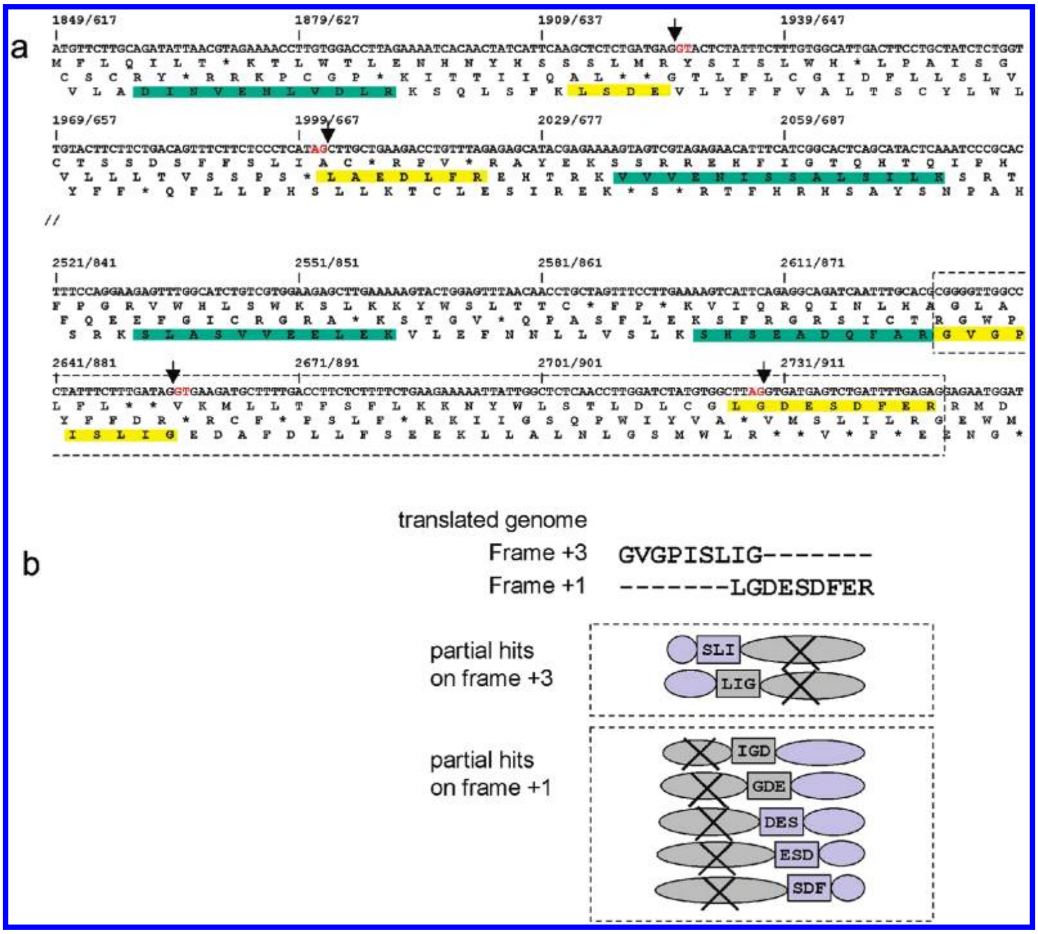

Figure 5. Detection of exon/intron boundaries by PepLine. (a) Zoom over the cluster corresponding to the AT1G06950 gene on chromosome I. The genomic sequence of the AT1G06950 gene (on reverse strand) and the 3-frames translated amino acid sequences are displayed. The numbers correspond to the position of the nucleotidic/proteic residues within the 5511 bases-At1g06950 gene. Peptide sequences matched by complete and partial hits are highlighted by green and yellow boxes, respectively. Partial hits allowed localization of two introns also annotated in the TAIR database (introns at 1926–2002 and at 2656–2728, indicated by vertical arrows). (b) Detailed support for the detection of the 2656–2728 intron. Upper part: amino acid sequences on frames +3 and +1 covered by partial hits delineating the intron/exon boundaries from 2656 to 2728 and showing an overlap according to the frame considered; lower part: representation of the entire set of PSTs generated from a single spectrum and corresponding to partial hits (crossed circles indicate the mass part of the PST that has not matched).

whole pipeline is fully compatible with high-throughput experiments. Using default parameters, the current running time of Taggor is about 30 ms/spectrum (on a Macintosh Intel 2 GHz, 1Go RAM). The analysis of the two MS/MS data sets described in this paper took less than 10 s. As for PMMatch, the current running time is between 0.5 and 1 ms/(PST/Mb). The total time to scan 3000 PSTs on the 120 Mb of complete chromosomes of *A. thaliana* was about 200 s on the same machine. Finally, the clustering step (PMClust) is very quick and takes, under normal conditions, less than a second.

The PepLine software is distributed under the GPL license and is freely available at: www.grenoble.prabi.fr/protehome/software/pepline.

4. Conclusion

We have developed three software modules that, together, allow genome annotation to be carried out from MS/MS data at high-throughput level. These three modules can be used independently, such as Taggor, whose function is to generate PSTs that can be used for searching for PTMs.[15] Taggor is particularly well-suited for QTOF-type data and, although simpler, gave similar results when compared to PepNovoTag. Thanks to the modularity of PepLine, one can use any other program in place of Taggor. When using ion-trap like instruments, existing software such as PepNovoTag or GutenTag shoud be preferred to generate PSTs to feed into PMMatch and PMClust.

Like database-driven approaches (e.g., Mascot or Sequest), the whole pipeline (Taggor + PMMatch + PMClust) proved to be efficient in finding genes previously predicted as known proteins. PepLine is also capable of finding new genes, as exemplified above by the identification of a new coding region in the TAIR genomic sequence. To go further in genome annotation capacity, PepLine may also prove efficient in the discovery of refinement of intron/exon boundaries. A large part of the performance of the pipeline arises from the two consecutive filtering effects by the PMMatch and PMClust modules, even when using minimal parameter settings. Indeed, the PMMatch task run with the “usual” trypsin constraints allowed most of the false PSTs to be discarded because they generated no hits. Finally, the PMClust task eliminated most false hits, merely by setting a minimum of hits to be clustered. Although our experiments have been performed on the *A. thaliana* genome, preliminary results suggest that the same approach could also be undertaken on the complete human genome without the need for restriction to putative transcripts. With recent advances in MS instrumentation and the increasing use of mass spectrometry in biological research, the volume of MS/MS data generated requires appropriate database searching tools in terms of reliability and speed. We believe that tools like PepLine, allowing a direct request on genomic information with MS/MS data, provide ideal complementary approaches to database-driven methods using Mascot or Sequest in order to characterize protein samples to their full extent.

Acknowledgment. We benefitted from the *A. thaliana* samples provided by the Laboratoire de Physiologie Cellulaire Végétale (CEA, DSV/iRTSV/PCV, Grenoble, F-38054, France; CNRS UMR5168; Université Joseph Fourier; INRA). We thank Sabine Brugière for setting up the standard protein MS/MS data, and Maighread Gallagher for reading this paper. We thank John Yates and colleagues for providing us with the GutenTag software license. This project was supported by the French Ministère de la Recherche et de la Technologie, Bioinformatics Call.

Note Added after ASAP Publication. This article was published ASAP on March 19, 2008. A minor text change has been made in the fifth paragraph of the Introduction section. The correct version was published March 26, 2008.

References

(1) Aebersold, R.; Mann, M. Mass spectrometry-based proteomics. *Nature* **2003**, *422* (6928), 198–207.

(2) Mathe, C.; Sagot, M. F.; Schiex, T.; Rouze, P. Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Res.* **2002**, *30* (19), 4103–4117.

(3) Brent, M. R.; Guigo, R. Recent advances in gene structure prediction. *Curr. Opin. Struct. Biol.* **2004**, *14* (3), 264–272.

(4) Reboul, J.; Vaglio, P.; Rual, J. F.; Lamesch, P.; Martinez, M.; Armstrong, C. M.; Li, S.; Jacotot, L.; Bertin, N.; Janky, R.; Moore, T.; Hudson, J. R., Jr.; Hartley, J. L.; Brasch, M. A.; Vandenhaute, J.; Boulton, S.; Endress, G. A.; Jenna, S.; Chevet, E.; Papasotiropoulos, V.; Tolias, P. P.; Ptacek, J.; Snyder, M.; Huang, R.; Chance, M. R.; Lee, H.; Doucette-Stamm, L.; Hill, D. E.; Vidal, M. C. elegans ORFeome version 1.1: experimental verification of the genome annotation and resource for proteome-scale protein expression. *Nat. Genet.* **2003**, *34* (1), 35–41.

(5) Elias, J. E.; Haas, W.; Faherty, B. K.; Gygi, S. P. Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations. *Nat. Methods* **2005**, *2* (9), 667–675.

(6) Sadygov, R. G.; Cociorva, D.; Yates, J. R., III. Large-scale database searching using tandem mass spectra: looking up the answer in the back of the book. *Nat. Methods* **2004**, *1* (3), 195–202.

(7) Washburn, M. P.; Wolters, D.; Yates, J. R., III. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* **2001**, *19* (3), 242–247.

(8) Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **1999**, *20* (18), 3551–3567.

(9) Eng, J.; McCormack, A. L.; Yates, J. R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **1994**, *5* (11), 976–989.

(10) Craig, R.; Beavis, R. C. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **2004**, *20* (9), 1466–1467.

(11) Ma, B.; Zhang, K.; Hendrie, C.; Liang, C.; Li, M.; Doherty-Kirby, A.; Lajoie, G. PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* **2003**, *17* (20), 2337–2342.

(12) Taylor, J. A.; Johnson, R. S. Implementation and uses of automated de novo peptide sequencing by tandem mass spectrometry. *Anal. Chem.* **2001**, *73* (11), 2594–2604.

(13) Hunt, D. F.; Yates, J. R., III; Shabanowitz, J.; Winston, S.; Hauer, C. R. Protein sequencing by tandem mass spectrometry. *Proc. Natl. Acad. Sci. U.S.A.* **1986**, *83* (17), 6233–6237.

(14) Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215* (3), 403–410.

(15) Shevchenko, A.; Sunyaev, S.; Loboda, A.; Shevchenko, A.; Bork, P.; Ens, W.; Standing, K. G. Charting the proteomes of organisms with unsequenced genomes by MALDI-quadrupole time-of-flight mass spectrometry and BLAST homology searching. *Anal. Chem.* **2001**, *73* (9), 1917–1926.

(16) Mann, M.; Wilm, M. Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal. Chem.* **1994**, *66* (24), 4390–9439.

(17) Frank, A.; Tanner, S.; Bafna, V.; Pevzner, P. Peptide sequence tags for fast database search in mass-spectrometry. *J. Proteome Res.* **2005**, *4* (4), 1287–1295.

(18) Tabb, D. L.; Saraf, A.; Yates, J. R., III. GutenTag: high-throughput sequence tagging via an empirically derived fragmentation model. *Anal. Chem.* **2003**, *75* (23), 6415–6421.

(19) Desiere, F.; Deutsch, E. W.; Nesvizhskii, A. I.; Mallick, P.; King, N. L.; Eng, J. K.; Aderem, A.; Boyle, R.; Brunner, E.; Donohoe, S.; Fausto, N.; Hafen, E.; Hood, L.; Katze, M. G.; Kennedy, K. A.; Kregenow, F.; Lee, H.; Lin, B.; Martin, D.; Ranish, J. A.; Rawlings, D. J.; Samelson, L. E.; Shiio, Y.; Watts, J. D.; Wollscheid, B.; Wright, M. E.; Yan, W.; Yang, L.; Yi, E. C.; Zhang, H.; Aebersold, R. Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry. *GenomeBiology* **2005**, *6* (1), R9.

(20) Arthur, J. W.; Wilkins, M. R. Using proteomics to mine genome sequences. *J. Proteome Res.* **2004**, *3* (3), 393–402.

(21) Giddings, M. C.; Shah, A. A.; Gesteland, R.; Moore, B. Genome-based peptide fingerprint scanning. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100* (1), 20–25.

(22) Choudhary, J. S.; Blackstock, W. P.; Creasy, D. M.; Cottrell, J. S. Interrogating the human genome using uninterpreted mass spectrometry data. *Proteomics* **2001**, *1* (5), 651–667.

(23) Kalume, D. E.; Peri, S.; Reddy, R.; Zhong, J.; Okulate, M.; Kumar, N.; Pandey, A. Genome annotation of Anopheles gambiae using mass spectrometry-derived data. *BMC Genomics* **2005**, *6*, 128.

(24) Savidor, A.; Donahoo, R. S.; Hurtado-Gonzales, O.; Verberkmoes, N. C.; Shah, M. B.; Lamour, K. H.; McDonald, W. H. Expressed peptide tags: an additional layer of data for genome annotation. *J. Proteome Res.* **2006**, *5* (11), 3048–3058.

(25) Fermin, D.; Allen, B. B.; Blackwell, T. W.; Menon, R.; Adamski, M.; Xu, Y.; Ulintz, P.; Omenn, G. S.; States, D. J. Novel gene and gene model detection using a whole genome open reading frame analysis in proteomics. *GenomeBiology* **2006**, *7* (4), R35.

(26) Allmer, J.; Markert, C.; Stauber, E. J.; Hippler, M. A new approach that allows identification of intron-split peptides from mass spectrometric data in genomic databases. *FEBS Lett.* **2004**, *562* (1–3), 202–206.

(27) Allmer, J.; Naumann, B.; Markert, C.; Zhang, M.; Hippler, M. Mass spectrometric genomic data mining: Novel insights into bioenergetic pathways in Chlamydomonas reinhardtii. *Proteomics* **2006**, *6* (23), 6207–6220.

(28) Kuster, B.; Mortensen, P.; Andersen, J. S.; Mann, M. Mass spectrometry allows direct identification of proteins in large genomes. *Proteomics* **2001**, *1* (5), 641–650.

(29) McGowan, S. J.; Terrett, J.; Brown, C. G.; Adam, P. J.; Aldridge, L.; Allen, J. C.; Amess, B.; Andrews, K. A.; Barnes, M.; Barnwell, D. E.; Berry, J.; Bird, H.; Boyd, R. S.; Broughton, M. J.; Brown, A.; Bruce, J. A.; Brusten, L. M. J.; Draper, N. J.; Elsmore, B. M.; Freeman, C. D.; Giles, D. M.; Gong, H.; Gormley, D.; Griffiths, M. R.; Hawkes, T. D. R.; Haynes, P. S.; Heesom, K. J.; Herath, A.; Hollis, K.; Hudsen, L.; Inman, J.; Jacobs, M.; Jarman, D.; Kibria, I.; Kilgour, J.; Kinuthia, S. K.; Lane, K. E.; Lees, M. L.; Loader, J.; Longmore, A.; McEwan, M.; Middleton, A.; Moore, S.; Murray, C.; Murray, H. M.; Myatt, C. P.; Ng, S. S.; O'Neil, A.; Parekh, R. B.; Patel, A.; Patel, K. B.; Patel, S.; Patel, T. P.; Philp, R. J.; Platt, A. E.; Poyser, H.; Prendergast, C.; Prime, S.; Redpath, N.; Reeves, M.; Robinson, A. W.; Rohlff, C. R.; Rosenbaum, J. M.; Schenker, M.; Scrivener, E.; Shipston, N.; Siddiq, S.; Southan, C.; Spencer, D. I. R.; Stamps, A.; Steffens, M. A.; Stevenson, D.; Sweetman, G. M. A.; Taylor, S.; Townsend, R.; Ventom, A. M.; Waller, M. N. H.; Weresch, C.; Williams, A. M.; Woolliscroft, R. J.; Yu, X.; Lyall, A. Annotation of the human genome by high-throughput sequence analysis of naturally occurring proteins. *Curr. Proteomics* **2004**, *1* (1), 41–48.

(30) Tanner, S.; Shen, Z.; Ng, J.; Florea, L.; Guigo, R.; Briggs, S. P.; Bafna, V. Improving gene annotation using peptide mass spectrometry. *Genome Res.* **2007**, *17* (2), 231–239.

(31) Tanner, S.; Shu, H.; Frank, A.; Wang, L. C.; Zandi, E.; Mumby, M.; Pevzner, P. A.; Bafna, V. InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. *Anal. Chem.* **2005**, *77* (14), 4626–4639.

(32) Bern, M.; Goldberg, D.; McDonald, W. H.; Yates, J. R., III. Automatic quality assessment of Peptide tandem mass spectra. *Bioinformatics* **2004**, *20* (1), I49–I54.

(33) Rolf, F. J. Single-link Clustering Algorithms. In *Handbook of Statistics*; Krishnaiah, P. R., Kanal, L. N., Eds.; North-Holland Publishing Company: New York, 1982; Vol. 2, pp 267–284.

(34) Ferro, M.; Seigneurin-Berny, D.; Rolland, N.; Chapel, A.; Salvi, D.; Garin, J.; Joyard, J. Organic solvent extraction as a versatile procedure to identify hydrophobic chloroplast membrane proteins. *Electrophoresis* **2000**, *21* (16), 3517–3526.

(35) Ferro, M.; Salvi, D.; Brugiere, S.; Miras, S.; Kowalski, S.; Louwagie, M.; Garin, J.; Joyard, J.; Rolland, N. Proteomics of the Chloroplast Envelope Membranes from Arabidopsis thaliana. *Mol. Cell. Proteomics* **2003**, *2* (5), 325–345.

PR070415K