BIOINFORMATICS APPLICATIONS NOTE
Vol. 29 no. 15 2013, pages 1913–1914
doi:10.1093/bioinformatics/btt298

Gene expression
Advance Access publication May 24, 2013

PIUS: peptide identification by unbiased search

Eduardo P. Costa[1,*], Gerben Menschaert[2], Walter Luyten[3], Kurt De Grave[1] and Jan Ramon[1]

[1]Department of Computer Science, KU Leuven, Celestijnenlaan 200A, B-3001 Heverlee, Belgium, [2]Department of Mathematical Modelling, Statistics and Bioinfomatics, Ghent University, Coupure Links 653, B-9000 Ghent, Belgium and [3]Department of Pharmaceutical and Pharmacological Sciences, KU Leuven, O&N II Herestraat 49, B-3000 Leuven, Belgium

Associate Editor: Burkhard Rost

ABSTRACT

Summary: We present PIUS, a tool that identifies peptides from tandem mass spectrometry data by analyzing the six-frame translation of a complete genome. It differs from earlier studies that have performed such a genomic search in two ways: (i) it considers a larger search space and (ii) it is designed for natural peptide identification rather than proteomics. Differently from other peptidomics tools designed for genome-wide searches, PIUS does not limit the analysis to a set of sequences that match a list of *de novo* reconstructions.

Availability: Source code, executables and a detailed technical report are freely available at http://dtai.cs.kuleuven.be/ml/systems/pius.

Contact: eduardo.costa@cs.kuleuven.be

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on February 1, 2013; revised on May 3, 2013; accepted on May 22, 2013

1 INTRODUCTION

In peptidomics, tandem mass spectroscopy is commonly used to identify peptides: an unknown peptide undergoes fragmentation, its fragment masses are registered in a so-called fragmentation spectrum and the peptide sequence is inferred from this spectrum. This latter step is usually performed by database search methods (Craig and Beavis, 2004): for each protein sequence in the database, potential fragments are predicted with their theoretical fragmentation spectrum; a scoring function then measures how well these spectra match the experimental one, returning the top scoring solutions. However, if the peptide is not derived from one of the database sequences, this strategy is bound to fail. In peptidomics, there are still many unidentified spectra of good quality, presumably partially because of this (Frank, 2009; Menschaert *et al.*, 2010a).

One strategy to deal with this limitation is to use *de novo* sequencing, which infers the amino acid sequence from the mass differences between neighboring peaks in the spectrum. However, from spectra of moderate quality, this strategy cannot extract enough information to unambiguously infer the complete sequence.

Another strategy, which we adopt in this note, is to compare the observed spectrum with the theoretical one of any peptide that could be translated from the genome (Jeong *et al.*, 2010; Kim *et al.*, 2009).

The proposed method, called PIUS (peptide identification by unbiased search), identifies peptides from fragmentation spectra using the six-frame translation of the complete genome. Even though a peptide search against the complete translation of a genome has been investigated in the context of proteogenomics (Gupta *et al.*, 2008; Kalume *et al.*, 2005), these studies were limited to small genomes (Kim *et al.*, 2009) and/or used additional assumptions that hold in the context of proteomics, where protein digestion occurs *in vitro* at predictable cleavage sites, but not in peptidomics, where the proteolytic processing occurs *in vivo* at sites that are difficult to predict. Further, these studies aimed at protein/gene discovery, whereas PIUS aims at peptide identification. The latter requires a much higher recall than the former: identifying a few of the peptides from a genomic region usually suffices to make a gene prediction. All in all, this means that peptidomics presents a markedly different context from proteomics.

Recently, three peptidomics methods have been proposed for genome-wide searches: MS-Dictionary (Kim *et al.*, 2009), MS-GappedDictionary (Jeong *et al.*, 2010) and IggyPep (Menschaert *et al.*, 2010a). The first two methods build a list of *de novo* reconstructions called a spectral dictionary and use it to reduce the number of candidates in the search. IggyPep queries the full genome translation using complete *de novo* reconstructions or partial peptide sequence tags (PST). PIUS differs from these methods because its search is exhaustive, not biased toward a subset of candidates, and it eliminates the need for good-quality PSTs.

2 METHOD

We mention the main aspects of the method; a detailed algorithmic description is available in the Supplementary Material.

In principle, PIUS analyzes all subsequences of the six-frame translation of the genome, checks whether their masses match that of the measured peptide (within an error tolerance) and returns the k highest-scoring matches, given a scoring function. In practice, it organizes the search such that pruning can be performed.

It starts with the first position of the translated genome and adds the subsequent residues one by one (sequence extension), until the mass either matches or becomes larger than that of the measured peptide; in the former case, the subsequence might enter the top k list depending on its score. Then, it stops the sequence extension, moves to the next start position and repeats the same procedure.

To mitigate the computational cost further, PIUS has a pruning procedure, which is based on the quality of the current subsequence as estimated by the scoring function being used. PIUS investigates the question 'is this subsequence unlikely to be a prefix of the correct solution (i.e. a fragment of the solution containing its N-terminal)?'. If so, the subsequence and its extensions are pruned. This is considered to be the case when the subsequence scores less than α times the lowest score of all prefixes of the same length in the top k list. The parameter α thus determines the eagerness to prune.

*To whom correspondence should be addressed.

The default scoring function in PIUS is derived from the one used in SEQUEST (Eng *et al.*, 1994). Although SEQUEST only considers ion series *b* and *y*, PIUS considers the ion series b, b^*, b^0, a, a^*, a^0, y, y^* and y^0. We assign a weight to each ion series to give more importance to abundant ions. To define the weights, we measured the relative frequency of each ion series in a distinct dataset used only for this purpose and for which the correct identifications were known. The user can reuse our weights or define his own.

Our search strategy does not consider sequences originating through splicing, neither does it account for polymorphism. We discuss possible strategies to handle these challenges in Section 3.5 of the Supplementary Material.

3 EXPERIMENTS AND RESULTS

We evaluated PIUS on a subset of 109 spectra from a combined set of peptide mass spectra from different mouse tissue and cell line samples produced by MALDI-TOF-TOF (IWT-50164, 2006–2010; Van Dijck *et al.*, 2011). The subset was obtained as follows. We used a combination of two search algorithms (X!Tandem and OMSSA) within SearchGui to obtain the peptide identification, allowing for the following PTMs: N-term acetylation, C-term amidation, oxidation on M and pyroglutamination on N-term Q. We analyzed the results with Peptide-Shaker and retained only those with a confidence level of 100%, allowing us to use them as a gold standard. Among the 109 spectra, six have a PTM. We compared our results with those of MS-GappedDictionary.

We used PIUS with an error tolerance of 0.5 kDa for both peptide and fragment mass, which is large enough to account for measurement errors in MALDI-TOF-TOF data. We arbitrarily set $k = 10{,}000$ and $\alpha = 0.8$. When searching for unmodified translations, of the 103 unmodified spectra, PIUS returned the gold standard solution (i) as the top 1 for 85 cases, (ii) as top 2 for 3 cases, (iii) in the top 10 of solutions for 8 cases, (iv) in a lower ranking position for 7 cases. These results show that PIUS is often able to reproduce definite identifications by X!Tandem/OMSSA, even though it searches a much larger space.

We also tested PIUS considering the same PTMs used to define the gold standard solutions, allowing one PTM per candidate. For the spectra with PTMs, PIUS found the gold standard solution as (i) the top 1 for five cases and (ii) as the top 3 for one case. For the spectra without PTMs, PIUS found the gold standard solution (i) as the top 1 for 82 cases, (ii) as top 2 for 5 cases, (iii) in top 10 of solutions for 5 cases and (iv) in a lower ranking position for 10 cases. For one case, PIUS pruned the gold standard solution away. These results show that PIUS has a high recall rate even when PTMs are considered.

We configured MS-GappedDictionary as recommended by its authors: the charge range was set to 1, no enzymatic cleavage was specified and all other parameters were left at their default values. Note that this tool does not consider PTMs in the search. Of the 103 unmodified spectra, MS-GappedDictionary found the gold standard solution as (i) the top 1 for 48 cases and as (ii) the top 2 for 2 cases. There were no hits for three cases and the gold standard solution was not among the returned hits for 50 cases. We also tested MS-GappedDictionary with the error tolerance used by PIUS (0.5 kDa). With this lower tolerance, the gold standard solution was (i) the top 1 for 57 cases, (ii) the top 2 for 2 cases and (iii) in the top 10 of solutions for 1 case. There were no hits for 3 cases, and the gold standard solution was not among the returned hits for 40 cases. This contrasts sharply with the good results reported by Jeong *et al.* (2010). A possible explanation for this inferior performance is the source of the mass spectra; Jeong *et al.* (2010) use other instruments that are generally more accurate than MALDI-TOF-TOF.

The current C implementation of PIUS has a throughput of two spectra per hour on an Intel Core i7-2600 when searching the 21 mouse chromosomes, using $\alpha = 0.8$. In case of large-scale experiments, we recommend to use PIUS in a layered peptidomics workflow (Menschaert *et al.*, 2010b). The workflow first searches databases in a conventional manner, and then calls on PIUS for the remaining unidentified (but high quality) spectra. In case this step would become a bottleneck, the algorithm can easily be distributed.

4 CONCLUSION

Many spectra of candidate bioactive peptides remain unidentified to date. To tackle this problem we present PIUS, an open source tool for peptide identification from tandem mass spectrometry data. PIUS searches the entire genome without previous assumptions to reduce the search space and is, therefore, most suited as a tool of last resort. We have validated PIUS with MALDI-TOF-TOF spectra. Even when searching an entire mammalian genome, PIUS obtains correct identifications in a large majority of cases and has a much higher recall rate than MS-GappedDictionary.

ACKNOWLEDGEMENT

The authors thank G. Baggerman and L. Schoofs for the spectra.

Funding: FWO-Vlaanderen [G.0413.09 and postdoc fellowship to G.M.]; European Research Council [240 186].

Conflict of Interest: none declared.

REFERENCES

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

Eng,J. *et al.* (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass. Spectrom.*, **5**, 976–989.

Frank,A. (2009) A ranking-based scoring function for peptide- spectrum matches. *J. Proteome Res.*, **8**, 2241–2252.

Gupta,N. *et al.* (2008) Comparative proteogenomics: combining mass spectrometry and comparative genomics to analyze multiple genomes. *Genome Res.*, **18**, 1133–1142.

IWT-50164 (2006–2010) Set of 19 unpublished spectra from SBO grant 50164 of the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT).

Jeong,K. *et al.* (2010) Gapped spectral dictionaries and their applications for database searches of tandem mass spectra. *Mol. Cell. Proteomics*, **10** (6), M110.002220.

Kalume,D. *et al.* (2005) Genome annotation of *Anopheles gambiae* using mass spectrometry-derived data. *BMC Genomics*, **6**, 128.

Kim,S. *et al.* (2009) Spectral dictionaries integrating de novo peptide sequencing with database search of tandem mass spectra. *Mol. Cell. Proteomics.*, **8**, 53–69.

Menschaert,G. *et al.* (2010a) A hybrid, de novo based, genome-wide database search approach applied to the sea urchin neuropeptidome. *J. Proteome Res.*, **9**, 990–996.

Menschaert,G. *et al.* (2010b) Peptidomics coming of age: a review of contributions from a bioinformatics angle. *J. Proteome Res.*, **9**, 2051–2061.

Van Dijck,A. *et al.* (2011) Comparison of extraction methods for peptidomics analysis of mouse brain tissue. *J. Neurosci. Methods*, **197**, 231–237.