Information Sciences 384 (2017) 126–134

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

PGMiner: Complete proteogenomics workflow; from data acquisition to result visualization

Canan Has[a,b], Jens Allmer[a,b,*]

[a] Molecular Biology and Genetics, Izmir Institute of Technology, Urla, Izmir, Turkey
[b] Bionia Incorporated, IZTEKGEB A8, Urla, Izmir, Turkey

ARTICLE INFO

Article history:
Received 20 February 2016
Revised 22 July 2016
Accepted 3 August 2016
Available online 4 August 2016

Keywords:
Proteogenomics
Mass spectrometry
Bioinformatics
Workflow management
Computational proteomics

ABSTRACT

In parallel with the development of nucleotide sequencing an equally important interest in further describing the sequence in terms of function arose and the latter represents the current bottleneck in the overall research question. Sequencing the transcriptome allows determination of expressed nucleotide sequences and using mass spectrometry allows sequencing on the protein level. Both approaches can only sequence a subset of the existing transcripts. Moreover, for example post translational modification events can only be determined on the proteomics level. Therefore, it is essential to combine proteomics and genomics. For that purpose, proteogenomics data analysis pipelines have been described. Here, we describe a novel proteogenomics workflow which encompasses everything from the acquisition of data to result visualization in the Konstanz Information Miner (KNIME), a state of the art workflow management and data analytics platform. We amended KNIME with a number of processes like peptide consensus prediction, peptide mapping, and database equalizing, as well as result visualization. This enabled construction of our new workflow, entitled PGMiner, which not only includes all data analysis steps, but is highly customizable which is rather cumbersome for most existing pipelines. Furthermore, no burdensome installation processes have to be performed making PGMiner the most user friendly tool available.

1. Introduction

Life science is faced with an ever increasing avalanche of new data and in biological and medical sciences mass spectrometry (MS) and next generation sequencing (NGS) are responsible for the largest part of it. These two data acquisition methodologies are used in genomics to sequence genomes and expressomes (NGS) and to sequence peptides and proteins (MS).

1.1. Genomics

Genomics is the study of genomes, their sequence, structure, and function [48]. Before the start of the human genome project (1990), sequencing was mostly performed using the Sanger methodology, generating perhaps a few thousand nucleotides per lab and day. One of the outcomes of the human genome project were advances in nucleotide sequencing

* Corresponding author.
E-mail address: jens@allmer.de (J. Allmer).

http://dx.doi.org/10.1016/j.ins.2016.08.005
0020-0255/

technology (mostly automation of sequencing tasks) and shortly following its conclusion, next generation sequencing technologies were introduced (2005) which today are able to sequence a complete human genome in one day. NGS technology is in widespread use and is not only employed to sequence genomes but also for sequencing of RNA and therefore transcriptomes. Transcription can, therefore, be monitored using NGS and it further allows the quantification of expression. Thus, NGS provides a platform to identify and quantify RNA expression, some of which may be coding while other parts are not coding for proteins. It has been argued, that translation of RNAs should be straight forward and thus RNA quantity should correlate well with protein abundance. Unfortunately, this is not the case since there are many regulatory and modification reactions that obscure the process. Post transcriptional gene expression modulation (e.g. by MicroRNAs), post translational modifications, protein degradation, and many other processes obfuscate the correlation between mRNA and protein abundance even further. Thus, it is essential to amend the genomic information with proteomic evidence.

1.2. Proteomics

The goal of proteomics is to identify, sequence, determine structure, functions, interactions, and other meta information of proteins in a sample [1]. Mass spectrometry is the work-horse of proteomics enabling high-throughput measurement of peptides in complex samples. Two modes of computational analysis of the resulting MS/MS data are possible. *De novo* sequencing assigns a peptide sequence to an MS/MS spectrum (peptide sequence match; PSM) from the information contained in the spectrum itself [3]. Database search in addition to MS/MS spectra needs a sequence database or a library of previously established PSMs to assign peptide sequences to spectra [15]. Due to current limitations of *de novo* sequencing, database search is the method of choice to assign peptides to MS/MS spectra. One important factor which affects PSM establishment is the content of the databases used. Limiting a database search to only known proteins obviously prohibits the detection of unknown peptides and proteins. In order to reveal novel peptides or proteins produced via alternative splicing, single-nucleotide-polymorphism, or alternative start site selection sequence databases should at least contain the six frame translation of the genome, EST, or RNA-Seq data while many more sequence databases may be useful [22,32]. Incorporating such genomic information is the first step towards proteogenomics.

1.3. Proteogenomics

The proteogenomics field emerged from this interplay of genomics, proteomics, which enabled the incorporation of additional information into gene prediction techniques [50]. Building upon that the next step was incorporation of MS/MS data to validate predicted genomic splice sites [49] and to propose new ones [5].

First studies in the field of proteogenomics included the confirmation and correction of prokaryotic gene models [34]. In addition to evaluating existing prokaryotic gene models by performing database search against the six-frame translation of underlying genomes the determination of novel genes have been achieved [10]. In this fashion, the detection of new prokaryotic and eukaryotic gene models have been achieved.

In addition to the analysis and the establishment of gene models, the detection of disease biomarkers or the response of biological systems to environmental changes can be achieved with proteogenomics studies [43] and has seen promising approaches for cancer detection [20]. Proteogenomics can also support the confirmation of transcripts which may be products of alternative open reading frames [33].

1.4. Data analysis for proteogenomics

Proteogenomics is a complex field and therefore requires the use of a plethora of different data analysis tools. The general workflow includes such steps as data acquisition, database generation, database search with mass spectrometric data, filtering and assessing the significance of such results, mapping of identifications to databases and/or existing gene models, finally the visualization results (Fig. 1). For the purpose of such complex data analysis, pipelines and workflows have been proposed covering some parts of the overall analysis.

Data acquisition involves either the actual measurement of genomic and/or proteomic data or their retrieval from public repositories. Therefore, proteogenomics data analysis workflows need to enable both handling of local files and their retrieval from online sources such as MS/MS data from PRIDE [45] or NGS data from SRA [30]. Sequence databases for database search of MS/MS data involves for example DNA or RNA data.

The raw reference sequences may need to be preprocessed including translation to three or six reading frames, filtering, creation of non-redundant databases, and generation of decoy ones. The selection and processing of sequence databases has a big impact on processing time [11] especially concerning the establishment of PSMs.

A large number of tools such as the open source algorithms OMSSA [18], XTandem [13], and MSGF+ [27] have been developed for database search, but there is only one algorithm (Morpheus) directly targeting proteogenomics [35]. It has been shown that using multiple database search tools in tandem can be beneficial [6]. The results from multiple database search tools may have to be integrated and it is important to establish the significance of the PSMs that were found. In order to assign confidence levels to PSMs various approaches have been employed such as false discovery rate (FDR) [25], percolation [46], and two fold search [23]. The assigned confidence is generally used to filter the PSMs.

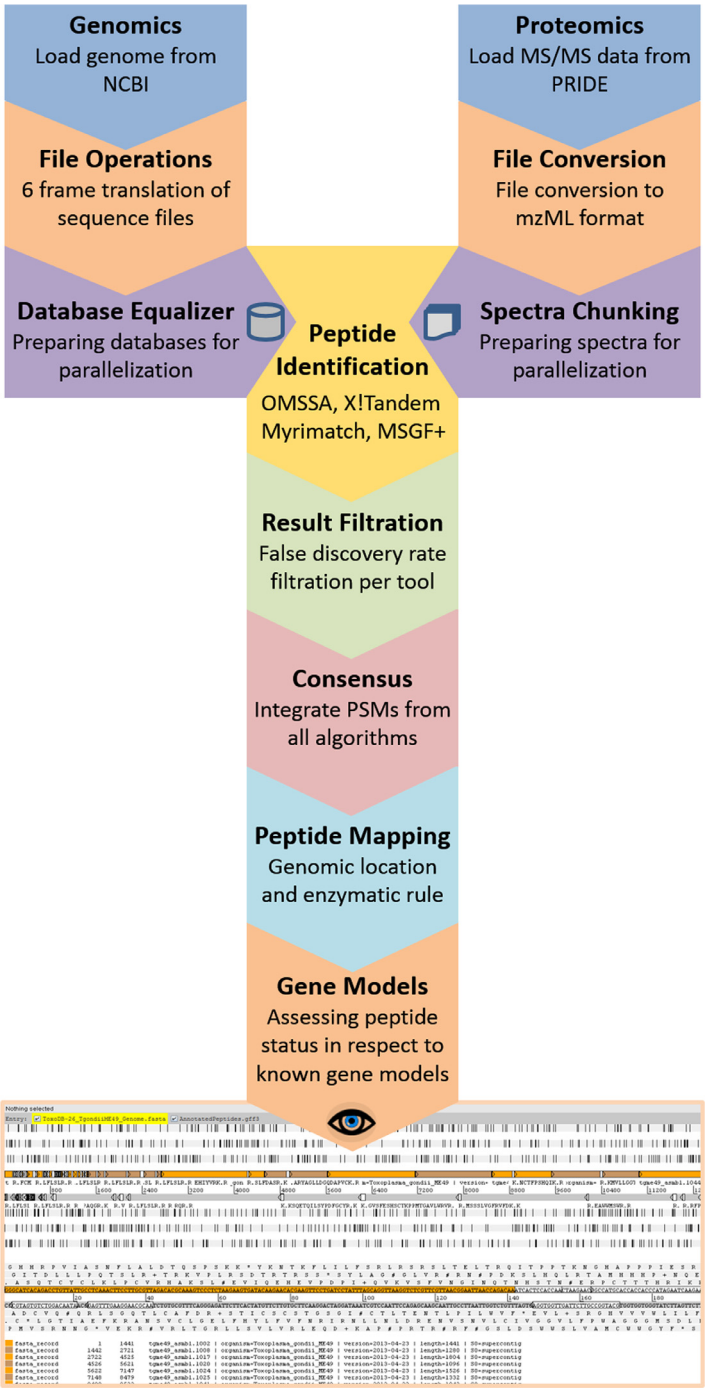

Fig. 1. Overview over the PGMiner workflow. Workflow starts with data acquisition (top) and ends with visualization in a genome browser (bottom).

In addition to the identification of peptides, proteogenomics studies need to map them to genomic locations and/or gene models since database search tools generally do not perform that task. Peptide mapping can be performed by specialized tools such as PGM [38], PGx [7], and the state-of-the-art tool Lelantos (http://jlab.iyte.edu.tr/software/lelantos). In addition to mere mapping, peptides also need to be categorized [21] as exonic, intronic, overlapping with known annotated genomic landmarks, or intergenic [16].

Using these mapping approaches, protein identification (protein inference) can also be performed [12]. Since peptides can ambiguously map to multiple proteins, it is important to establish whether peptides are proteotypic which can be achieved by Lelantos using multiple sequence databases and/or genomic annotations. Protein inference can then be accomplished using available algorithms in KNIME [39].

Finally, it is of interest to visualize the identified peptides and the existing genomic annotations in a platform which allows the visual inspection of multiple lines of evidence such as genomic, transcriptomic and proteomic ones. Known genome browsers for this purpose are Ensembl [17] and UCSC [26], but many more tools such as iPiG [28], the Integrated Genomic Browser (http://bioviz.org/igb/credits.html), and Artemis [36] exist.

Here, we present PGMiner a flexible and comprehensive proteogenomics workflow built using the workflow management and data analytics platform KNIME (Fig. 1). PGMiner includes (i) spectral and sequence data acquisition, (ii) translation of nucleotide sequences and preparation for parallel computations, (iii) peptide identification using multiple database search algorithms, (iv) PSM filtering via FDR, (v) mapping of PSMs to sequence databases and gene annotation, (vi) protein inference, and finally (vii) result visualization in a genome browser. PGMiner is further customizable using existing modules of OpenMS [41] which are integrated with KNIME. In this manner, quantitation facilities could be added to PGMiner which may be of interest to proteogenomics studies [40].

2. PGMiner workflow

PGMiner was developed as a workflow in the Konstanz Information Miner (KNIME; [8]) and was tested on KNIME version 3.1.1 using Java 1.8. KNIME is a data analytics platform and has a visual workflow management environment which uses nodes to model processes and edges to indicate data flow. The OpenMS library makes all its modules available via KNIME and workflows can be constructed using OpenMS tools [2]. Where possible OpenMS nodes were used, but missing operations were implemented as custom nodes for KNIME which are available via our KNIME update site: http://bioinformatics.iyte.edu.tr/PGMiner. In general, no additional installation other than KNIME (and using its update sites) is necessary to use PGMiner (see training video for PGMiner on our YouTube channel BioinformaticsIZTECH or on our web site: http://jlab.iyte.edu.tr/software/PGMiner) whereas existing pipelines often have involved installations procedures. The PGMiner workflow is fully automated and no steps have to be performed manually which is required in some of the existing proteogenomic pipelines. In an attempt to further simplify usage, we also provide a virtual machine image containing a fully functional installation of PGMiner.

All proteogenomics pipelines need suitable data input and often genome data and its annotation acquisition is the first step in proteogenomics. Generally, pipelines expect the user to either use a sequence database provided by the pipeline [24], or needs the user to acquire such information in the appropriate format [35]. PGMiner, is the only tool which offers the direct download of such information from NCBI, but alternatively allows the user to import local files to the workflow.

2.1. Data acquisition

The PGMiner workflow was tested on sequence and proteomics data from the human pathogen *Toxoplasma gondii*. The spectra collection with the accession PXD001042, having 197,204 spectra [19], was retrieved from PRIDE (http://www.ebi.ac.uk/pride/archive/projects/PXD001042) using PGMiner. *T. gondii* was chosen as an example because it has a genome size that can be processed by PGMiner in acceptable time on one computer and since we have an interest in the organism in terms of gene regulation [37].

The current release (http://toxodb.org/common/downloads/Current_Release/TgondiiME49/fasta/data/, release date: 07-Oct-2015) of ToxoplasmaME49 genomic and protein sequence databases were retrieved from ToxoDB (http://toxodb.org) and the GFF annotation file was also downloaded from the same repository (http://toxodb.org/common/downloads/Current_Release/TgondiiME49/gff/data/, version 26, release date: 07-Oct-2015).

ToxoDB is not a commonly used sequence repository, therefore, we do not support data download from it directly and data has to be imported manually to the workflow. However, sequences which are available on NCBI RefSeq can be retrieved via PGMiner. It should be noted that the *T. gondii* GFF file included FASTA formatted sequence information which needed to be removed before the analysis as actual sequences are not expected within GFF files but rather their genomic landmarks, which remained unchanged.

Mass spectrometry data is generally imported manually [29,31,35] and different file formats are acceptable such as mgf (Mascot Generic Format), mzML, and mzXML. There is no competing pipeline which allows users to fetch spectral data from general mass spectrometry data repositories such as PeptideAtlas available at http://www.peptideatlas.org/ or PRIDE (http://www.ebi.ac.uk/pride/archive/). PGMiner enables users to provide local files in mgf, mzXML, or mzML format. Moreover, MS data can be fetched via querying the PRIDE repository. Similar to most existing proteogenomics pipelines PGMiner supports the interconversion among MS file formats.

The ability to directly import data from online repositories simplifies sharing of the workflow since it removes the necessity to share large amounts of data along with the workflow; additionally it simplifies the distribution of the workflow on systems like Amazon Web Services.

2.2. Data preprocessing

In protegenomic studies genomic databases, their three, or six frame translations should provide the basis for any analysis. However, often transcriptomic data such as EST, cDNA, RNA-Seq or protein data such as proteome or gene models are used as sequence databases. In addition to the aforementioned sequence databases, Helmy et al. [22] proposes the use

of specialized forms such as exon graphs and exon-exon junction databases. Proteogenomics tools such as GenoSuite [29], Peppy [35], pGalaxy [24], ProteoAnnotator [19], and BPP [44] offer the translation of genomic sequences. PGMiner accepts both, genomic and protein sequence files in parallel, automatically determines their type, and if necessary translates them.

A known problem for database search of large databases is that many database search tools fail to process them and thereby terminate the pipelines. Therefore, large eukaryotic genomes are rarely used in any proteogenomics study. Even if no crucial problem occurs, pipeline runtime increases with an increase in database size and, therefore, some tools allow the preprocessing of the sequence databases. Peppy creates first segments from input genome then generates possible peptides from segments in a multithread manner and pGalaxy filters sequence database according to HiRIEF [9] technique.

PGMiner uses a different approach by providing database equalizing which enables database search tools to perform database searches more quickly and which further enables seamless result integration and finally supports future parallelization of the database search part of the workflow (one of the most computationally intensive part). The pseudocode of this database equalization approach is given below.

```
Input: A number of sequence files F1, F2, ..., Fn; a positive integer l setting the desired length of sequence
elements;
a positive integer o setting the length of overlapping sequences; a positive integer m giving the desired
number of equalized
files to be produced.
Output: m number of equalized sequence files (E1, E2, ... Em).
Procedure:
1. get m, l, o
2. get Files F; n = size(F)
3. initialize files E[0 .. m]
4. e = 1
5. for i = 0 to n
6.      foreach S in Fi // S represents a sequence
7.          p=0
8.          while p < len(S) --l
9.               E[e++] <- store subsequence(S,p,l)
10.             p += o
11.             if e == m
12.                 e = 0
13.         end while
14.    end foreach
15. end for
16. return E
```

Decoy databases which are necessary for statistical result filtration, are generated on the fly by reversing or shuffling the target sequences i.e. the equalized databases.

2.3. Peptide identification

It has been shown that consensus identification of multiple tools increase number of correctly identified spectra [6,14]. While GenoSuite, PGTools, pGalaxy and ProteoAnnotator include multiple database search algorithm execution for peptide identification step, BPP outsources peptide identification results and Peppy employs only one algorithm for peptide identification.

PGMiner is intentionally designed to perform proteogenomics analysis using multiple database search engines. OMSSA, XTandem, MSGF+, and Myrimatch [42] connectors provided by OpenMS community nodes are used to drive database search algorithms. All available settings of each algorithm can be adjusted by the user but a preselection has been made which generally works sufficiently well. Identification using multiple database search algorithms entails data integration.

2.4. Scoring peptide spectrum matches

To eliminate false-positive hits, some proteogenomics tools use FDR (Peppy, BPP, GenoSuite) or percolation (PGTools). In PGMiner, databases are first equalized and their decoy versions are created by reversing or shuffling their sequences. Each database search algorithm is executed on both target and decoy databases. In the next step, consensus of target database results and decoy database results are calculated for each algorithm, q-values are computed for each PSM, then FDRScores are computed as described in [25] and the ones below an adjustable threshold are removed. Q-value formulation used in this study is as follows:

$$q - value = \left(\frac{FP}{FP + TP}\right)$$

Here FP symbolizes decoy hits and TP symbolizes target database hits. True positive hits are calculated as removal of FP from all target hits that are above threshold, $T_{\text{above threshold}}$.

$$TP = \left(T_{\text{above threshold}} - FP\right)$$

The consensus building node is available on PGMiner which is based on majority vote of multiple algorithms. The pseudocode of consensus building node is described in the following:

```
Input: F1, F2, ..., Fn representing identification files of more than two algorithms e.g.: MSGF+,
X!Tandem, OMSSA. S representing MGF formatted MS/MS spectra files associated with result files F1, F2,
..., Fn
Output: Consensus predictions
Procedure:
1. Initialize T, TreeMap(spectrum, consensus prediction)
2. Initialize H, Hashmap(peptide,tool_support)
3. foreach spectrum s in S
4.      foreach identification file Fi
5.          peptide = get best peptide hit
6.          if peptide is in H
7.              increment tool_support
8.          else
9.              add peptide with tool_support 1
10.      end foreach
11.      if H is not empty
12.          L = list(H.entrySet)
13.          Sort L descendingly by tool_support
14.          p <- highest tool_support having peptide as consensus prediction
15.          T.add(p)
16. end foreach
17. return T
```

2.5. Mapping peptides to sequence databases and gene annotation

Mapping peptides to gene annotations is the central intersection of genomics and proteomics in the field of proteogenomics. Most proteogenomics pipelines employ some specialized tool for mapping identified peptides to the underlying genomic database. PGMiner employs Lelantos (http://jlab.iyte.edu.tr/software/lelantos) for peptide mapping, a tool which employs the WuManber [47] algorithm to achieve unequalled speeds. Our implementation has complexity O(n) [4]. The pseudocode of peptide query matching and location storage step of Lelantos is given below.

```
Input: t, char array representing sequence database; c, sequence type (nucleotide or protein); p, char
array representing patterns (query peptide sequences); w, a positive integer representing word size
Output : result, GFF3 file containing locations of patterns within t
Procedure:
1. n = len(t)
2. Initialize and construct SUFFIX hash, SHIFT map, and PREFIX hash
3. s = len(shortest pattern)
4. tp = s - 1
5. while (tp < n)
6.      word = get w size char array ending at tp
7.      pats = SUFFIX(word) // returns all patterns with word as suffix
8.      if(pats)
9.          foreach pat in pats
10.             pref = get first word from pattern
11.             if pref in PREFIX
12.                 compare p with t //given constraint tp
13.                 if pattern equals to text ending at position tp
14.                     if c equals to ‘‘nucleotide’’
15.                         add genomic location to result
16.                     else if c equals to ‘‘protein’’
17.                         add protein location to result
18.         end foreach
19.         tp += SHIFT(word)
20. end while
21. return result
```

Additionally, enzymatic cleavage rules of the enzyme used during sample preparation are checked following the mapping procedure (if desired) in order to eliminate potentially false mappings.

Mapped peptides are intersected with known information such as gene models. Such annotations can be provided in GFF format. Annotations enable PGMiner to categorize peptide identifications into classes confirming known annotations or conflicting with them. This information is one of the most important outcomes of any proteogenomics analysis and PGMiner further provides information whether conflicting peptide identifications are intergenic, intronic, or overlapping with exons.

Table 1
According to gene annotation mapping results 542 gene models were matched to mapped peptide locations. Gene models were grouped according to confirmed gene models, 3′ overlapping conflict gene models, 5′ overlapping conflict gene models and partially confirmed and additional overlapping peptides gene models.

Status of gene models	Number of gene models	Peptides mapped
Gene models with peptide support for exons	499	1366
Gene models with conflicting 3′ overlapping peptides	19	20
Gene models with conflicting 5′ overlapping peptides	3	3
Confirmed gene models with additional conflicting overlapping peptides	21	118
Other conflicting peptide mappings (Intergenic)	–	1485

2.6. Visualization

Proteogenomics tools like pGalaxy and GenoSuite allow the visualization of identified peptides and proteins in their genomic context. PGMiner is no exception and uses the Integrative Genomics Viewer (http://www.broadinstitute.org/software/igv/) to visualize its findings within the genomic context. In addition to that the Artemis Genome Browser can be used to visualize annotated peptides in their genomic context.

3. Results

Similar to the study by Ghali et al. [19] *T. gondii* data was used. The analysis took approximately 3 h on an AMD 6300 3.5 GHz 6 core 32 GB RAM PC running 64bit Microsoft Windows version 7.

OMSSA, MSGF+, and XTandem database search algorithms were used with the following settings: 1.5 Da precursor mass tolerance and 0.6 Da fragment mass tolerance allowing 1 miss-cleavage. For both algorithms, carbamidomethylation of cysteine residues and oxidation of methionine residues were set as fixed modifications. Since both algorithms accept spectra files in mzML format, downloaded spectra file in mgf format were converted to mzML format. Myrimatch was not employed in this proof of principle but can easily be integrated into the workflow similar to OMSSA, MSGF+, and XTandem.

Each algorithm was executed using ten equalized target database files and their shuffled decoy versions. Afterwards, for each spectrum the best hit was selected according to lowest E-value. Using the OMSSA algorithm 49,291 PSMs were found in the target database and 49,293 PSMs were found in the decoy database. The XTandem algorithm returned 48,611 PSMs using the decoy database and 48,611 PSMs were established using the target database. The MSGF+ algorithm returned 48,905 PSMs on target database and 48,900 PSMS on decoy database.

Before integration of results of the three algorithms, FDR was applied to filter false-positive hits. After 1% FDR, 4017 PSMs passed the filter for OMSSA, 4937 PSMs for XTandem, and 15,922 PSMs for MSGF+. Integration of results led to 5560 consensus peptide-spectrum matches. Integrating Myrimatch and optimization of precursor mass and fragment mass tolerance per algorithm would increase the number of identified consensus peptides. Here XTandem and OMSSA were the limiting factors, returning low amount of results in comparison to MSGF+, which is likely due to our restrictive parameter settings for the search.

Filtered peptides were mapped to genomic locations using Lelantos and the locations which were not in accordance with the expectation of expected enzymatic cleavage were removed. Finally, mapping locations were compared to gene annotations of *T. gondii*. 542 unique genes were confirmed by peptides found in this study. 1485 peptides mapped to intergenic regions of the genome. 1453 matches of unique peptides to exons were found (Table 1). While 499 gene models only included exonic peptides (confirmations), 19 gene models had 3′ overlapping conflicting peptides (Fig. 2, blue track, peptide c) where a 3′ overlap is a peptide whose N-terminus overlaps with the 3′ end of an exon. 3 gene models had 5′ overlapping peptides (Fig. 2, green track, peptides a and b), i.e.: the C-terminus of a peptide overlaps with the 5′ end of an exon. 21 gene models were confirmed with exonic peptides (Fig. 2, green track, peptides 3 and 4) and also included overlapping conflicting peptides. Except for the confirmed gene models, the ones with conflicting overlapping peptides may need to be reviewed in respect to their exon intron structure or possible alternatively spliced products should be considered. Peptides that overlap 5′ with an exon may either indicate that the exon could be extended or may mean that other translation start sites could be considered. Peptides which overlap 3′ with an exon either signify that the exon could be extended or that the annotated stop needs to be changed. All peptide mappings that lead to conflicts with existing gene models may also indicate that there are alternative structures possible for that locus and thus do not directly challenge the existing gene model, but may amend it with alternative versions.

An example for confirmed gene models with additional overlaps is shown in Fig. 2 for the *T. gondii* gene model TGME49_207620 on TGME49 chromosome 1b. Peptides 3 (LEGNAPDVK; spectral support: 2), 4 (DGAIVTDPLLR; spectral support: 1), 5 (FLADQSEFALAR; spectral support: 3), and 6 (NSQDVLAIAK; spectral support: 2) in the green track (Fig. 2) confirm the exons of the gene model. Peptides 1 (LLLCTGSEAR; spectral support: 1) and 2 (VTAVCTMGR; spectral support 1) have a 5′ overlap with the second and third exon, respectively and indicate that the exons should be extended into 5′ direction or show that there may be alternative spliced products possible.

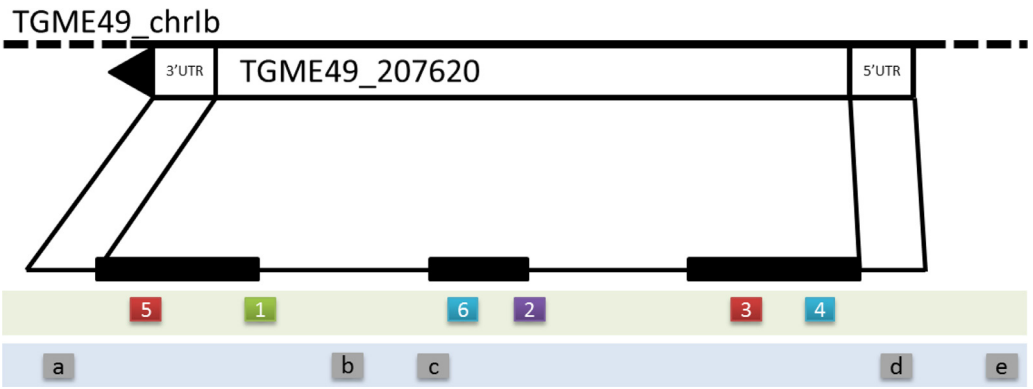

Fig. 2. An example for a confirmed gene model in *T. gondii* with additional 5′overlaps in its genomic context for gene model TGME49_207620 on the TGME49 TGME49_1b chromosome. The upper (green) track contains the 5′ overlapping peptides 1 and 2 as well as peptides confirming the exons (3, 4, 5, and 6) found in this study. Additional mappings that are possible, alas, not found for this example, are provided in the lower (blue) track. Peptides may be mapped with a 3′ overlap to an exon (c), into the 5′ (d) or 3′ (a) UTR, to an intron (b), or intergenic (e.g. outside of any gene model). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4. Conclusions

PGMiner, a comprehensive proteogenomics workflow, has been established in KNIME, a state-of-the-art workflow management and data analytics platform. A novel approach to database handling, mapping and data integration has been introduced by PGMiner making it unique when compared to other tools. PGMiner was tested using *T. gondii* data and 1626 (>50%) peptides identified in this study had conflicts with existing gene models or did not map to one.

Application of the workflow to data from *Toxoplasma gondii* further revealed that of the about 500 gene models identified about 10% had conflicting peptide mappings in respect to their proposed gene structure.

One important distinction of PGMiner is that its installation is straight forward and that the workflow is not broken-up into parts and does not require any user attention once initiated making it the most user friendly among all proteogenomics tools. Additionally, for users familiar with virtual machines, a ready to use image for virtual box is provided on the PGMiner website which relieves the need to install and update KNIME.

Conflicts of interest

Funders and employers didn't take part in any process that produced this work. Authors declare no conflict of interest.

Acknowledgements

The work was supported by the Scientific and Technological Research Council of Turkey [grant number 114Z177] and a Scientific Research Grant from the Izmir Institute of Technology [grant number 2013IYTE04] to JA. Authors would like to thank the KNIME team, Dr. Thorsten Meinl, Dr. Tobias Koetter, Dr. Kilian Thiel, Alexander Fillbrunn, Hermann Azong and Budi Yanto for their support during node development.

References

[1] R.R. Aebersold, M. Mann, Mass spectrometry-based proteomics, Nature 422 (2003) 198–207.
[2] S. Aiche, T. Sachsenberg, E. Kenar, M. Walzer, B. Wiswedel, T. Kristl, M. Boyles, A. Duschl, C.G. Huber, M.R. Berthold, K. Reinert, O. Kohlbacher, Workflows for automated downstream data analysis and visualization in large-scale computational mass spectrometry, Proteomics 15 (2015) 1443–1447.
[3] J. Allmer, Algorithms for the de novo sequencing of peptides from tandem mass spectra, Expert Rev. Proteomics. 8 (2011) 645–657.
[4] J. Allmer, Exact pattern matching: adapting the Boyer-Moore algorithm for DNA searches, Peer J PrePrints (2016) doi: 10.7287/peerj.preprints.1758v1.
[5] J. Allmer, C. Markert, E.J. Stauber, M. Hippler, A new approach that allows identification of intron-split peptides from mass spectrometric data in genomic databases, FEBS Lett 562 (2004) 202–206.
[6] J. Allmer, B. Naumann, C. Markert, M. Zhang, M. Hippler, Mass spectrometric genomic data mining: novel insights into bioenergetic pathways in Chlamydomonas reinhardtii, Proteomics 6 (2006) 6207–6220.
[7] M. Askenazi, K.V. Ruggles, D Fenyö, PGx: putting peptides to BED, J. Proteome Res (2015) acs.jproteome.5b00870.
[8] M.R. Berthold, N.N. Cebron, F. Dill, T.R. Gabriel, T. Kötter, T. Meinl, P. Ohl, K. Thiel, B. Wiswedel, C. Sieb, K. Thiel, B. Wiswedel, KNIME - the konstanz information miner, SIGKDD Explor 11 (2009) 26–31.
[9] R.M.M. Branca, L.M. Orre, H.J. Johansson, V. Granholm, M. Huss, Å. Pérez-Bercoff, J. Forshed, L. Käll, J Lehtiö, HiRIEF LC-MS enables deep proteome coverage and unbiased proteogenomics, Nat. Methods. 11 (2014) 59–62.
[10] N. Castellana, V. Bafna, Proteogenomics to discover the full coding content of genomes: a computational perspective, J. Proteomics. 73 (2010) 2124–2135.
[11] N.E. Castellana, V. Pham, D. Arnott, J.R. Lill, V. Bafna, Template proteogenomics: sequencing whole proteins using an imperfect database, Mol. Cell. Proteomics MCP. 9 (2010) 1260–1270.
[12] M. Claassen, Inference and validation of protein identifications, Mol. Cell. Proteomics. 11 (2012) 1097–1104.
[13] R. Craig, R.C. Beavis, TANDEM: matching proteins with tandem mass spectra, Bioinformatics 20 (2004) 1466–1467.

[14] R.K. Dagda, T. Sultana, J. Lyons-Weiler, Evaluation of the consensus of four peptide identification algorithms for tandem mass spectrometry based proteomics, J. Proteomics Bioinform. 3 (2010) 39–47.
[15] J.K. Eng, B.C. Searle, K.R. Clauser, D.L. Tabb, A face in the crowd: recognizing peptides through database search, Mol. Cell. Proteomics (2011) 10 R111.009522.
[16] D. Fermin, B.B. Allen, T.W. Blackwell, R. Menon, M. Adamski, Y. Xu, P. Ulintz, G.S. Omenn, D.J. States, Novel gene and gene model detection using a whole genome open reading frame analysis in proteomics, Genome Biol (2006) 7 R35.
[17] P. Flicek, I. Ahmed, M.R. Amode, D. Barrell, K. Beal, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, S. Fairley, S. Fitzgerald, L. Gil, C. García-Girón, L. Gordon, T. Hourlier, S. Hunt, T. Juettemann, A.K. Kähäri, S. Keenan, M. Komorowska, E. Kulesha, I. Longden, T. Maurel, W.M. McLaren, M. Muffato, R. Nag, B. Overduin, M. Pignatelli, B. Pritchard, E. Pritchard, H.S. Riat, G.R.S. Ritchie, M. Ruffier, M. Schuster, D. Sheppard, D. Sobral, K. Taylor, A. Thormann, S. Trevanion, S. White, S.P. Wilder, B.L. Aken, E. Birney, F. Cunningham, I. Dunham, J. Harrow, J. Herrero, T.J.P. Hubbard, N. Johnson, R. Kinsella, A. Parker, G. Spudich, A. Yates, A. Zadissa, S.M.J. Searle, Ensembl 2013, Nucleic Acids Res. (2012).
[18] L.Y. Geer, S.P. Markey, J.a Kowalak, L. Wagner, M. Xu, D.M. Maynard, X. Yang, W. Shi, S.H. Bryant, Open mass spectrometry search algorithm, J. Proteome Res. 3 (2004) 958–964.
[19] F. Ghali, R. Krishna, S. Perkins, A. Collins, D. Xia, J. Wastling, A.R. Jones, ProteoAnnotator - open source proteogenomics annotation software supporting psi standards, Proteomics (2014) 1–26.
[20] M. Helmy, N. Sugiyama, M. Tomita, Y. Ishihama, Onco-proteogenomics: a novel approach to identify cancer-specific mutations combining proteomics and transcriptome deep sequencing, Genome Biol 11 (2010) 17.
[21] M. Helmy, N. Sugiyama, M. Tomita, Y. Ishihama, Mass spectrum sequential subtraction speeds up searching large peptide MS/MS spectra datasets against large nucleotide databases for proteogenomics, Genes Cells 17 (2012) 633–644.
[22] M. Helmy, M. Tomita, Peptide identification by searching large-scale tandem mass spectra against large databases: bioinformatics methods in proteogenomics, Genes Genomes Genomics 6 (2012) 76–85.
[23] P. Jagtap, J. Goslinga, J.a Kooren, T. McGowan, M.S. Wroblewski, S.L. Seymour, T.J. Griffin, A two-step database search method improves sensitivity in peptide sequence matches for metaproteomics and proteogenomics studies, Proteomics 13 (2013) 1352–1357.
[24] P.D. Jagtap, J.E. Johnson, G. Onsongo, F.W. Sadler, K. Murray, Y. Wang, G.M. Shenykman, S. Bandhakavi, L.M. Smith, T.J. Griffin, Flexible and accessible workflows for improved proteogenomic analysis using the Galaxy framework, J. Proteome Res. 13 (2014) 5898–5908.
[25] A.R. Jones, J.A. Siepen, S.J. Hubbard, N.W. Paton, Improving sensitivity in proteome studies by analysis of false discovery rates for multiple search engines, Proteomics 9 (2009) 1220–1229.
[26] W.J. Kent, C.W. Sugnet, T.S. Furey, K.M. Roskin, T.H. Pringle, A.M. Zahler, D. Haussler, The human genome browser at UCSC, Genome Res 12 (2002) 996–1006.
[27] S. Kim, N Mischerikow, N Bandeira, J.D Navarro, L Wich, S Mohammed, A.J.R Heck, P.A Pevzner, The generating function of CID, ETD, and CID/ETD pairs of tandem mass spectra: applications to database search, Mol. Cell. Proteomics. 9 (2010) 2840–2852.
[28] M. Kuhring, B.Y. Renard, iPiG: integrating peptide spectrum matches into genome browser visualizations, , PLoS One 7 (2012) e50246.
[29] D. Kumar, A.K. Yadav, P.K. Kadimi, S.H. Nagaraj, S.M. Grimmond, D. Dash, Proteogenomic analysis of Bradyrhizobium japonicum USDA110 using GenoSuite, an automated multi-algorithmic pipeline, Mol. Cell. Proteomics. 12 (2013) 3388–3397.
[30] R. Leinonen, H. Sugawara, M. Shumway, The sequence read archive, Nucleic Acids Res 39 (2011) D19–D21.
[31] S.H. Nagaraj, N. Waddell, A.K. Madugundu, S. Wood, A. Jones, R.A. Mandyam, K. Nones, J.V. Pearson, S.M. Grimmond, PGTools: a software suite for proteogenomic data analysis and visualization, J. Proteome Res. 14 (2015) 2255–2266.
[32] A.I. Nesvizhskii, Proteogenomics: concepts, applications and computational strategies, Nat. Methods. 11 (2014) 1114–1125.
[33] K. Ning, A.I. Nesvizhskii, The utility of mass spectrometry-based proteomic data for validation of novel alternative splice forms reconstructed from RNA-Seq data: a preliminary assessment, BMC Bioinform. 11 (2010) S14.
[34] S. Renuse, R. Chaerkady, A. Pandey, Proteogenomics. Proteomics 11 (2011) 620–630.
[35] B.A. Risk, W.J. Spitzer, M.C. Giddings, Peppy: proteogenomic search software, J. Proteome Res. 12 (2013) 3019–3025.
[36] K R.P. Rutherford, J Parkhill, J Crook, T Horsnell, Artemis: sequence visualization and annotation, Bioinformatics 16 (2000) 944–945.
[37] M.D. Saçar, C. Bağcı, J. Allmer, Computational prediction of MicroRNAs from toxoplasma gondii potentially regulating the hosts' gene expression, Genomics. Proteomics Bioinform. 12 (2014) 228–238.
[38] W.S. Sanders, N. Wang, S.M. Bridges, B.M. Malone, Y.S. Dandass, F.M. McCarthy, B. Nanduri, M.L. Lawrence, S.C. Burgess, The proteogenomic mapping tool, BMC Bioinform 12 (2011) 115.
[39] O. Serang, M.J. MacCoss, W.S. Noble, Efficient marginalization to compute protein posterior probabilities from shotgun mass spectrometry data, J. Proteome Res. 9 (2010) 5346–5357.
[40] P.A. Stewart, K. Parapatics, E.A. Welsh, A.C. Müller, H. Cao, B. Fang, J.M. Koomen, S.A. Eschrich, K.L. Bennett, E.B. Haura, A pilot proteogenomic study with data integration identifies MCT1 and GLUT1 as prognostic markers in lung adenocarcinoma, PLoS One 10 (2015) e0142162.
[41] M. Sturm, A. Bertsch, C. Gröpl, A. Hildebrandt, R. Hussong, E. Lange, N. Pfeifer, O. Schulz-Trieglaff, A. Zerck, K. Reinert, O. Kohlbacher, OpenMS – an open-source software framework for mass spectrometry, BMC Bioinform. 9 (2008) 163.
[42] D.L. Tabb, C.G. Fernando, M.C. Chambers, MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis, J. Proteome Res. 6 (2007) 654–661.
[43] A. Tanca, M. Deligios, M.F. Addis, S. Uzzau, High throughput genomic and proteomic technologies in the fight against infectious diseases, J. Infect. Dev. Ctries. (2013) 7.
[44] J. Uszkoreit, N. Plohnke, S. Rexroth, K. Marcus, M. Eisenacher, The bacterial proteogenomic pipeline, BMC Genomics 15 (Suppl 9) (2014) S19.
[45] J.A. Vizcaíno, R. Côté, F. Reisinger, H. Barsnes, J.M. Foster, J. Rameseder, H. Hermjakob, L. Martens, The Proteomics identifications database: 2010 update, Nucleic Acids Res 38 (2010) D736–D742.
[46] J.C. Wright, M.O. Collins, L. Yu, L. Kall, M. Brosch, J.S. Choudhary, Enhanced peptide identification by electron transfer dissociation using an improved mascot percolator, Mol. Cell. Proteomics (2012) 478–491.
[47] S. Wu, U. Manber, A fast algorithm for multi-pattern searching, Tech. Rep TR94 (1994) 1–11.
[48] J.R. Yates, Mass spectrometry. From genomics to proteomics, Trends Genet 16 (2000) 5–8.
[49] M.M. Yin, J.T.L. Wang, Effective hidden Markov models for detecting splicing junction sites in DNA sequences, Inf. Sci. (Ny). (2001) 139–163.
[50] M.M. Yin, J.T.L. Wang, GeneScout: a data mining system for predicting vertebrate genes in genomic DNA sequences, Inf. Sci. (Ny). 163 (2004) 201–218.