PGMiner reloaded, fully automated proteogenomic annotation tool linking genomes to proteomes

Canan Has[1,2 *], Sergey A. Lashin[3,4], Alexey Kochetov[3] Jens Allmer[1,2 *]

[1]Molecular Biology and Genetics, Izmir Institute of Technology, Urla, Izmir, Turkey

[2]Bionia Incorporated, IZTEKGEB A8, Urla, Izmir, Turkey

[3]Institute of Cytology & Genetics, SB RAS, Novosibirsk, Russia;

[4]Novosibirsk State University, Novosibirsk, Russia

Abstract

Improvements in genome sequencing technology increased the availability of full genomes and transcriptomes of many organisms. However, the major benefit of massive parallel sequencing is to better understand the organization and function of genes which then lead to understanding of phenotypes. In order to interpret genomic data with automated gene annotation studies, several tools are currently available. Even though the accuracy of computational gene annotation is increasing, a combination of multiple lines of experimental evidences should be gathered. Mass spectrometry allows the identification and sequencing of proteins as major gene products; and it is only these proteins that conclusively show whether a part of a genome is a coding region or not to result in phenotypes. Therefore, in the field of proteogenomics, the validation of computational methods is done by exploiting mass spectrometric data. As a result, identification of novel protein coding regions, validation of current gene models, and determination of upstream and downstream regions of genes can be achieved. In this paper, we present new functionality for our proteogenomic tool, PGMiner which performs all proteogenomic steps like acquisition of mass spectrometric data, peptide identification against preprocessed sequence databases, assignment of statistical confidence to identified peptides, mapping confident peptides to gene models, and result visualization. The extensions cover determining proteotypic peptides and thus unambiguous protein identification. Furthermore, peptides conflicting with gene models can now automatically assessed within the context of predicted alternative open reading frames.

1 Introduction

Recent improvements in next generation sequencing (NGS) technology led to an increase in the number of sequenced organisms including ones lacking annotated genes and/or proteins. To account for the missing information, *in silico* gene prediction methods have been employed to predict gene structures, open reading frames, and putative protein coding sequences. Predictions on the protein level are based on sequence homology with known proteins from, for example, model organisms. This methodology is limited to the availability of homologous proteins and by the evolutionary distance among organisms of interest and model organisms [1]. Automatic computer-aided predictions should be supported by

* To whom correspondence should be addressed. Email: cananhas@gmail.com, jens@allmer.de

doi:10.2390/biecoll-jib-2016-293

experimental data. The state-of-the art technology in proteomics for protein identification is mass spectrometry (MS) which provides the opportunity to confirm peptide expression and in turn protein expression. MS data analysis is currently using database search to assign peptide sequences to MS/MS spectra and is limited by sequence availability in databases. With the aid of NGS technology, custom sequence databases can be built by using six- or three-frame translated DNA or RNA sequences. Additionally, available protein sequences, predicted gene models and their derivatives such as alternative spliced forms, exon-exon junction peptides, and single-nucleotide polymorphic sequence variants can be used as databases [2]. Identified peptides can validate gene models but can also allow the discovery of novel coding regions or altered protein sequences that might be related to a certain metabolic state such as disease or environmental stress. In addition to that, correlation of expression among transcriptomics and proteomics expression levels can be investigated for confirmed and novel genes [3].

The field of proteogenomics exists at this intersection of genomics with proteomics [3]. Proteogenomics studies have been validating existing gene models, have discovered novel gene models, and have shown conflicts with existing gene models [4]–[8]. Moreover, proteogenomics strategies have applications in biomarker discovery [9]–[11].

Proteogenomic analyses can be broken down into 6 coarse steps which are: 1) data acquisition, 2) building a custom sequence database, 3) performing database search of MS/MS spectra against this database, 4) statistical significance assessment of peptide-spectrum matches, 5) mapping statistically confident peptides to the genome while taking into account annotated gene models, and, finally, 6) the evaluation and visualization of results.

In this paper, we offer an extension to PGMiner [12] a user-friendly proteogenomic pipeline developed using the KNIME data analytics platform. PGMiner includes the main steps of proteogenomics in a fully automated manner. The workflow enables users to retrieve mass spectrometry based proteomics data and to perform peptide identification by multi-algorithm support. Finally, PGMiner supports machine aided assessment of gene models by mapping identified peptides and proposal of new gene models.

2 Related works

Competing approaches with PGMiner also combine analysis steps into one framework for either eukaryotic, prokaryotic organisms, or both [13]. Some of these tools such as the Bacterial Proteogenomic Pipeline (BPP) [14], Peppy [15], ProteoAnnotator [16] were developed including a GUI while some of them such as PGTools [17] include command-line modules. pGalaxy [18] was developed on the Galaxy data analysis framework and as such is most comparable to PGMiner. All other solutions require the user to provide mass spectrometry data and genome data and its annotation while PGMiner can retrieve them directly from online repositories. All tools support usage of genomic, transcriptomic and protein sequence databases however automatic translation as in PGMiner is not supported by other approaches. Testing competing implementations, large sequence files led to computational runtime problems and caused termination of the pipelines. To tackle this problem, Peppy generates peptide segments from translated genomic sequences, whilst pGalaxy applies HiRIEF [19] based filtering on generated peptides. Both approaches remove some, potentially, important data from screening whereas PGMiner enables search without removing any sequences. While Peppy supports only one database search algorithm, other tools, like PGMiner, enable multiple algorithms to increase identification confidence and number of correctly identified spectra. Main features of these available pipelines were listed in Table 1.

Table 1 Comparison of proteogenomics tools are listed in terms of general proteogenomic workflow steps.

Pipeline	Organism	Data acquisition	Database preproces	Database search algorithms	Statistical assessment	Peptide mapping	Extended features
GenoSuite (2013)	Prokaryote	User input	6-ORF translation	OMSSA X!Tandem InsPecT MassWiz	FDR -Peptide level -Protein level	No algorithm name Against in silico gene annotation	
Peppy (2013)	Eukaryote	User input	Generate peptide segments	Morpheus algorithm	FDR	No algorithm mentioned Against genome and proteome	
Bacterial Protegenomic Pipeline (2014)	Prokaryote	User input	-	Outsourcing results	User dependent	No algorithm mentioned Against genome and proteome	Proteotypic peptides
ProteoAnnotator (2014)	Prokaryote Eukaryote	User input	6-ORF translation	SearchGUI toolkit	FDR	Against in silico gene annotation	
pGalaxy (2014)	Prokaryote Eukaryote	User input	6-ORF translation	ProteinPilot	Two round search ProteinPilot	Blastp Ab initio proteins	
PGTools (2015)	Prokaryote Eukaryote	User input	6-ORF translation	Xtandem OMSSA MSGF+ Comet	FDR PEP	Blastp Ab initio proteins	
PGMiner	Prokaryote Eukaryote	-Fetching via repository -User input	3-ORF 6-ORF translation	OMSSA X!Tandem MSGF+	FDR Peptide level	Wu-Manber BLAST All databases	Proteotypic peptide finding AltORFs finding

3 Implementation

PGMiner is a JAVA based proteogenomic workflow developed in the Konstanz Information Miner (KNIME) [20] version 3.1.1 using Java 1.8. KNIME is a data analytics platform including a visual workflow management environment which uses nodes to model processes and edges to indicate data flow. PGMiner addresses needs in different aspects of proteogenomics such as data acquisition from data repositories, peptide identification, peptide mapping, and proposal of new or corrected gene models and finally visualization of these models (Figure 1).

PGMiner has been developed as a KNIME workflow and all novel nodes we added to KNIME are available from our update site: http:/bioinformatics.iyte.edu.tr/PGMiner. Whilst existing pipelines require elaborative installation procedure and have manually controlled or workflow-independent steps, PGMiner has a simple installation procedure and can then be executed in a fully automated manner. Detailed instructions regarding PGMiner installation are described on our web site: http://jlab.iyte.edu.tr/software/PGMiner.

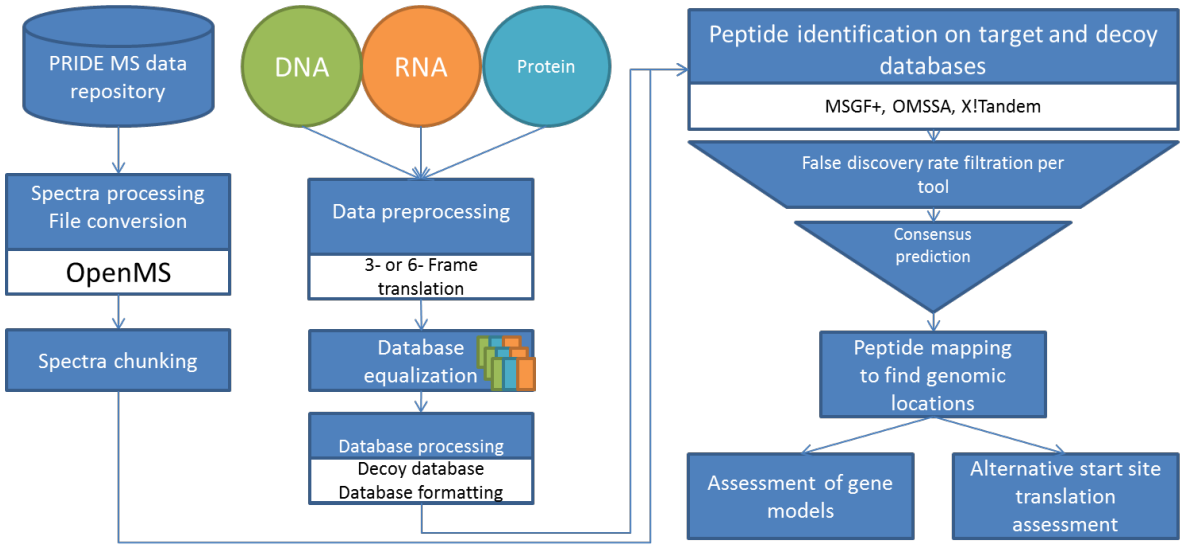

Figure 1 Overview of the PGMiner workflow. MS data can be directly acquired from PRIDE as can sequence data from, if available on ncbi. MS data and sequences can be packaged for parallel processing and peptide identification can be performed by three database search tools (MSGF+, OMSSA, and X!Tandem at the moment. Novel features are the automatic assessment of gene models and selected analysis of alternative translation start sites.

Protein identification is a convoluted process since peptides are shared among proteins and other regions of a genome and, therefore, it is hard to unambiguously identify a protein. PGMiner has been amended with the ability to determine proteotypic peptides. Proteotypic peptides are here defined as peptides occurring only in one location in respect to all mappings of all sequence databases used for PSM establishment to the reference genome. Detection of unambiguously identified proteins and framing other proteins as ambiguous identification is important in large scale studies in order to avoid misinterpretation. PGMiner's proteotypic peptide finder requires gene annotation file in GFF format and sequence files that are compatible with annotation files in terms of sequence accessions. Among selected sequence databases, peptides that are found only in one sequence region are considered as proteotypic. Ambiguities due to different level of associations, i.e. an exonic region might be related to multiple mRNAs and multiple proteins originating from one locus, are resolved in this manner since genomic start and end positions are taken into account.

PGMiner has also been amended to enable prediction of alternative start sites for selected gene models when proteotypic and additional supporting peptides are available. For this, PGMiner mostly follows the linear scanning mechanism where a 40S ribosomal subunit binds to a capped 5'-end of a translation start codon located in an appropriate context [21]–[24]. PGMiner currently only allows the analysis of peptides conflicting with existing gene models, which have been categorized as intronic.

4 Application

In this study, the human pathogen *Toxoplasma gondii* RH strain LC-MS/MS collection (PRIDE accession: PXD003603) was used to demonstrate PGMiner's functionality. Three spectral datasets were available in the collection measured by QTOF Impact HD, Maxis 4G (Bruker Daltonics, Bremen, Germany) and Ion Trap amaZon (Bruker Daltonics), respectively. MS/MS spectra with less than 15 peaks were eliminated. The database search tool nodes of PGMiner: OMSSA, MSGF+, and XTandem were used with the following settings: 0.3Da precursor mass tolerance and 0.35Da fragment mass tolerance for Ion Trap amaZon and Maxis 4G spectra; 50 ppm precursor mass tolerance and 0.1Da fragment mass tolerance for QTOF Impact HD. One miscleavage was allowed and carbamidomethylation of cysteine residues and oxidation of methionine residues were set as fixed modifications.

Genome sequences of RH strain and ME49 strain, annotated proteins, annotated transcripts, open-reading frames and coding sequences were retrieved from ToxoDB release 28 (2016-03-23). Since *T. gondii* is a human pathogen, we filtered human contaminant peptides. To identify those contaminant peptides, we used human *ab initio* gene models and annotated protein sequences. Nucleotide databases were translated to their six reading frames. In total the databases were 689 MB in size and they were processed into 10 equal size databases by using our database equalizer module [12]. The decoy version of each database was generated by shuffling sequences. The best hit per spectrum was selected among hits retrieved from the 10 databases for each spectrum on a per algorithm basis. This step was carried out for decoy hits, as well. Human contaminant peptide matching spectra were excluded. The summary of the results are presented in Table 2.

Table 2: Number of target and decoy peptide-spectrum matches obtained from X!Tandem, OMSSA, and MSGF+ using the toxoplasma genome and human gene models are listed on a per collection basis from PXD003603.

	Target						Decoy					
	X!Tandem		OMSSA		MSGF+		X!Tandem		OMSSA		MSGF+	
Spectra Collection	Filtered Hits	Human Cont. Hits	Filtered Hits	Human Cont. Hits	Filtered Hits	Human Cont. Hits	Filtered Hits	Human Cont. Hits	Filtered Hits	Human Cont. Hits	Filtered Hits	Human Cont. Hits
2012-36-11 ImpactVps26	36251	7431	50155	0	36584	7397	38047	6085	50530	0	37431	6537
2012-36-11 ImpactVps35	37507	8368	52085	0	37870	8266	39731	6540	52480	0	39493	6627
2012-36-15 ImpactVps29	37336	12409	54965	0	34585	11859	43173	7363	55561	0	39086	7331
2012-36-16 ImpactVps35	69545	18443	102584	0	69424	17278	76109	12643	102890	0	72215	14022
2012-36-16 ImpactVps26	69053	16062	96510	0	67928	14665	73196	12673	96900	0	68811	13732
2009-26-11 MaxisVps35	20756	3866	29325	0	19723	3274	21563	3303	29617	0	19822	3127
2009-26-11 amaZonVps26	32801	6113	44071	0	22567	3950	35333	4913	44614	0	22963	3552
Total # of PSMs	303249	72692	429695	0	288681	66689	327152	53520	432592	0	299821	54928

As a result 429,695 target hits and 432,592 decoy hits for OMSSA, 303,249 target and 327,152 decoy hits for X!Tandem and 288,681 target and 299,821 decoy hits for MSGF+ were found. Filtering by 1% FDR led to 11,753 hits for OMSSA, 21,158 hits for X!Tandem and 12,625 hits for MSGF+. Integration of these results identified 12,241 consensus peptide-spectrum matches.

Gene models are either supported through peptides, which in turn are supported via PSMs, or have at least one conflicting peptide mapping (Table 3). Overall, 2,888 unique peptides mapped to 370 unique gene models. Of these peptides 1,052 were identified to be proteotypic. 1,266 peptides were exonic (i.e.: directly supporting annotated gene models) and 13 peptides were overlapping with 5' end of gene models while 31 peptides were overlapping with 3' ends of gene models. 24 gene models had 3' overlapping peptides with 6 of them having also

exonic peptides. 2 gene models had only 5' overlapping peptides and 5 gene models had exonic and 5' overlapping peptides. 339 gene models had only peptides mapped to exons. No intergenic peptides were found. In addition to that there was no alternative start site selection transcript in this dataset, however, the approach was developed for human and may not be applicable for *T. gondii*.

Table 3: According to our results, in total, 370 gene models had peptides mapped to them with 350 only containing peptides supporting the annotation. For other gene models supporting peptides may exist, but in addition peptides which conflict with the available annotation by either overlapping on the 3' side or 5' side with an annotated gene model were found.

Status of gene models	Number of gene models
Gene models with peptide support for exons	339
Gene models with conflicting 3' overlapping peptides	24
Gene models with conflicting 5' overlapping peptides	7

5 Discussion

In this study, we presented an extended version of PGMiner, a new proteogenomic workflow tool, which performs automatic assessment of current gene models for eukaryotic and prokaryotic organisms based on mass spectrometric data. The workflow enables users to acquire data from data repositories and to perform peptide identification by employing multiple database search tools against various sequence databases in a parallel manner. Statistically assessed peptides are further mapped to genome annotations, thereby new gene models can be proposed and current models can be evaluated as confirmed or in need of revision. In order to unambiguously identify gene models, labeling peptides as proteotypic or not is important and the extended version of PGMiner allows users to make such assessment according to user-selected databases. Peptides which are labelled as intronic can be further checked whether they are related to alternative start site selection transcript products.

Acknowledgements

This study was supported by the Scientific and Technological Research Council of Turkey [grant number 114Z177].

References

[1] N. Castellana and V. Bafna, "Proteogenomics to discover the full coding content of genomes: a computational perspective.," *J. Proteomics*, vol. 73, no. 11, pp. 2124–2135, 2010.

[2] M. Helmy and M. Tomita, "Peptide Identification by Searching Large-Scale Tandem Mass Spectra against Large Databases: Bioinformatics Methods in Proteogenomics," *Genes Genomes and Genomics*, vol. 6, no. 1, pp. 76–85, 2012.

[3] A. I. Nesvizhskii, "Proteogenomics: concepts, applications and computational strategies.," *Nat. Methods*, vol. 11, no. 11, pp. 1114–25, Nov. 2014.

[4] D. A. Bitton, D. L. Smith, Y. Connolly, P. J. Scutt, and C. J. Miller, "An Integrated Mass-Spectrometry Pipeline Identifies Novel Protein Coding-Regions in the Human Genome," *PLoS One*, vol. 5, no. 1, p. e8949, Jan. 2010.

[5] N. E. Castellana, S. H. Payne, Z. Shen, M. Stanke, V. Bafna, and S. P. Briggs, "Discovery and revision of Arabidopsis genes by proteogenomics," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 105, no. 52, pp. 21034–21038, 2008.

[6] T. Dandekar, M. Huynen, J. T. Regula, B. Ueberle, C. U. Zimmermann, M. A. Andrade, T. Doerks, et al., "Re-annotating the Mycoplasma pneumoniae genome sequence: adding value, function and reading frames.," *Nucleic Acids Res.*, vol. 28, no. 17, pp. 3278–3288, 2000.

[7] F. Desiere, E. W. Deutsch, A. I. Nesvizhskii, P. Mallick, N. L. King, J. K. Eng, A. Aderem, et al., "Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry.," *Genome Biol.*, vol. 6, no. 1, p. R9, 2005.

[8] M. Helmy, M. Tomita, and Y. Ishihama, "OryzaPG-DB: Rice Proteome Database based on Shotgun Proteogenomics," *BMC Plant Biol.*, vol. 11, no. 1, p. 63, 2011.

[9] A. Tanca, M. Deligios, M. F. Addis, and S. Uzzau, "High throughput genomic and proteomic technologies in the fight against infectious diseases," *J. Infect. Dev. Ctries.*, vol. 7, no. 03, Mar. 2013.

[10] P. A. Stewart, K. Parapatics, E. A. Welsh, A. C. Müller, H. Cao, B. Fang, J. M. Koomen, et al., "A Pilot Proteogenomic Study with Data Integration Identifies MCT1 and GLUT1 as Prognostic Markers in Lung Adenocarcinoma," *PLoS One*, vol. 10, no. 11, p. e0142162, 2015.

[11] T. K. Sigdel and M. M. Sarwal, "The proteogenomic path towards biomarker discovery.," *Pediatr. Transplant.*, vol. 12, no. 7, pp. 737–747, 2008.

[12] C. Has and J. Allmer, "PGMiner: Complete proteogenomics workflow; from data acquisition to result visualization," *Inf. Sci. (Ny).*, Aug. 2016.

[13] P. K. Sarkar, P. K. Prajapati, V. J. Shukla, B. Ravishankar, and A. K. Choudhary, "Toxicity and recovery studies of two ayurvedic preparations of iron.," *Indian J. Exp. Biol.*, vol. 47, no. 12, pp. 987–92, Dec. 2009.

[14] J. Uszkoreit, N. Plohnke, S. Rexroth, K. Marcus, and M. Eisenacher, "The bacterial proteogenomic pipeline.," *BMC Genomics*, vol. 15 Suppl 9, no. Suppl 9, p. S19, 2014.

[15] B. A. Risk, W. J. Spitzer, and M. C. Giddings, "Peppy: proteogenomic search software.," *J. Proteome Res.*, vol. 12, no. 6, pp. 3019–25, Jun. 2013.

[16] F. Ghali, R. Krishna, S. Perkins, A. Collins, D. Xia, J. Wastling, and A. R. Jones, "ProteoAnnotator - Open Source Proteogenomics Annotation Software Supporting PSI Standards.," *Proteomics*, pp. 1–26, 2014.

[17] S. H. Nagaraj, N. Waddell, A. K. Madugundu, S. Wood, A. Jones, R. A. Mandyam, K. Nones, et al., "PGTools: A software suite for proteogenomic data analysis and visualization," *J. Proteome Res.*, vol. 14, no. 5, pp. 2255–2266, 2015.

[18] P. D. Jagtap, J. E. Johnson, G. Onsongo, F. W. Sadler, K. Murray, Y. Wang, G. M. Shenykman, et al., "Flexible and accessible workflows for improved proteogenomic analysis using the Galaxy framework.," *J. Proteome Res.*, vol. 13, no. 12, pp. 5898–908, Dec. 2014.

[19] R. M. M. Branca, L. M. Orre, H. J. Johansson, V. Granholm, M. Huss, Å. Pérez-Bercoff, J. Forshed, et al., "HiRIEF LC-MS enables deep proteome coverage and unbiased proteogenomics.," *Nat. Methods*, vol. 11, no. 1, pp. 59–62, Jan. 2014.

[20] M. R. Berthold, N. Cebron, F. Dill, T. R. Gabriel, T. Kötter, T. Meinl, P. Ohl, et al., "KNIME: The Konstanz Information Miner," in *SIGKDD Explorations*, vol. 11, no. 1, 2008, pp. 319–326.

[21] C. E. Aitken and J. R. Lorsch, "A mechanistic overview of translation initiation in eukaryotes," *Nat. Struct. Mol. Biol.*, vol. 19, no. 6, pp. 568–576, 2012.

[22] O. M. Alekhina and K. S. Vassilenko, “Translation initiation in eukaryotes: versatility of the scanning model.,” *Biochem. Biokhimiĭa*, vol. 77, no. 13, pp. 1465–77, 2012.

[23] R. J. Jackson, C. U. T. Hellen, and T. V Pestova, “The mechanism of eukaryotic translation initiation and principles of its regulation.,” *Nat. Rev. Mol. Cell Biol.*, vol. 11, no. 2, pp. 113–127, 2010.

[24] N. Malys and J. E. G. McCarthy, “Translation initiation: variations in the mechanism can be anticipated.,” *Cell. Mol. life Sci. C.*, vol. 68, no. 6, pp. 991–1003, 2011.

[25] B. Vanderperre, J.-F. Lucier, and X. Roucou, “HAltORF: a database of predicted out-of-frame alternative open reading frames in human.,” *Database (Oxford).*, vol. 2012, p. bas025, Jan. 2012.