Improving a Knowledge Base for Use in Proteomic Data Analysis

Jonathan L. Lustgarten[1], Vanathi Gopalakrishnan[1,2], William R. Hogan[1,3], Shyam Visweswaran[1,2]
[1]Department of Biomedical Informatics and the [2]Intelligent Systems Program
University of Pittsburgh
200 Meyran Ave, M-183, Pittsburgh, PA 15260, USA
[3]University of Pittsburgh Medical Center
Pittsburgh, PA, USA
JLL: jll47@pitt.edu, VG: vanathi@pitt.edu, WRH: wrh9@pitt.edu, SV: shv3@pitt.edu

Abstract

We have developed a knowledge base containing information linking experimentally validated *m/z* ratios to proteins that a user can query to retrieve candidate proteins corresponding to a *m/z* ratio of interest. In this paper, we describe an improvement to the knowledge base that allows a user to query for *m/z* ratios associated with a protein of interest.

1 Introduction

Several significant challenges remain in the analysis of high-dimensional data generated in mass spectrometry–based proteomic experiments. One goal of such experiments that utilize mass spectrometry data is protein identification. Given a spectrum of mass-to-charge ratios (*m*/*z*) the aim is to identify those proteins and peptides that are present with high intensity in the data. The chief challenge in the assignment of putative protein identification to an experimentally identified *m*/*z* ratio stems from the variability of that value. Typically, the *m*/*z* ratio corresponding to a protein varies due to several factors: the biofluid from which the sample is obtained (e.g., plasma, cerebrospinal fluid, tissue lysate, etc.) [Gopalakrishnan, *et al.*, 2004], the mass spectrometry platform, and the occurrence of different forms of a protein resulting from post-translational modifications, alternative splicing and the existence of isoforms [Lustgarten, *et al.*, 2008].

The researcher who needs to identify candidate proteins corresponding to a *m/z* ratio, typically, either searches the literature or online knowledge bases such as the UniProt [The UniProt, 2007] and the Expasy [Gasteiger, *et al.*, 2003]. We have recently developed a knowledge base called the EPO-KB (Empirical Proteomic Ontology Knowledge Base) [Lustgarten, *et al.*, 2008] containing information linking experimentally validated *m/z* ratios to proteins that was curated from the biomedical literature. The EPO-KB can be queried via a Web interface (http://www.dbmi.pitt.edu/EPO-KB) to retrieve proteins that have experimentally determined *m/z* ratios that are close (based on a distance score) to the user-specified *m/z* ratio. Thus, EPO-KB enables identification of candidate proteins corresponding to a *m/z* ratio of interest that can then be evaluated further by experimental methods such as immunochemistry.

A second purpose of generating proteomic data is to identify unique proteomic patterns such as a set of *m/z* ratios that discriminates well disease samples from healthy samples. Such analyses may be assisted by the identification and removal from the data of those *m/z* ratios that represent non-specific inflammatory proteins such as serum amyloid A. We have now added the ability to perform reverse look-ups in EPO-KB, i.e., identification of all *m/z* ratios that have been linked to a particular protein in the literature. We further describe this functionality below.

2 Enabling protein to *m/z* ratios search

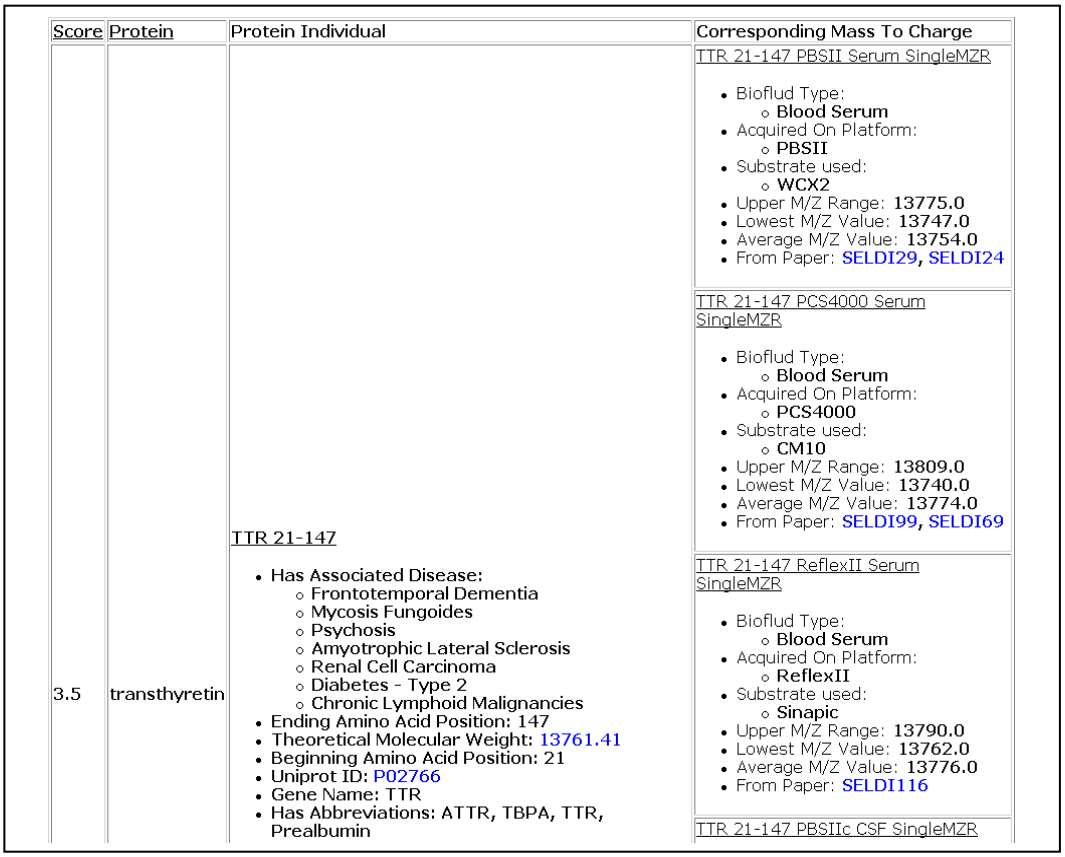
To allow querying of EPO-KB for proteins, we needed a standardized nomenclature to identify proteins uniquely. We chose the UniProt knowledge base [The UniProt, 2007] which provides a unique identifier for each protein and also contains information such as the sequence, possible modifications, and the gene that produces the protein.

For reverse look-ups, EPO-KB can be queried by a user-specified UniProt identifier. This retrieves all *m/z* ratios in the knowledge base associated with the protein and includes *m/z* ratios associated with fragments, mutations, and post-translational modifications of the protein. The retrieved *m/z* ratios can further be filtered on the basis of mass spectrometry techniques and the sample biofluid.

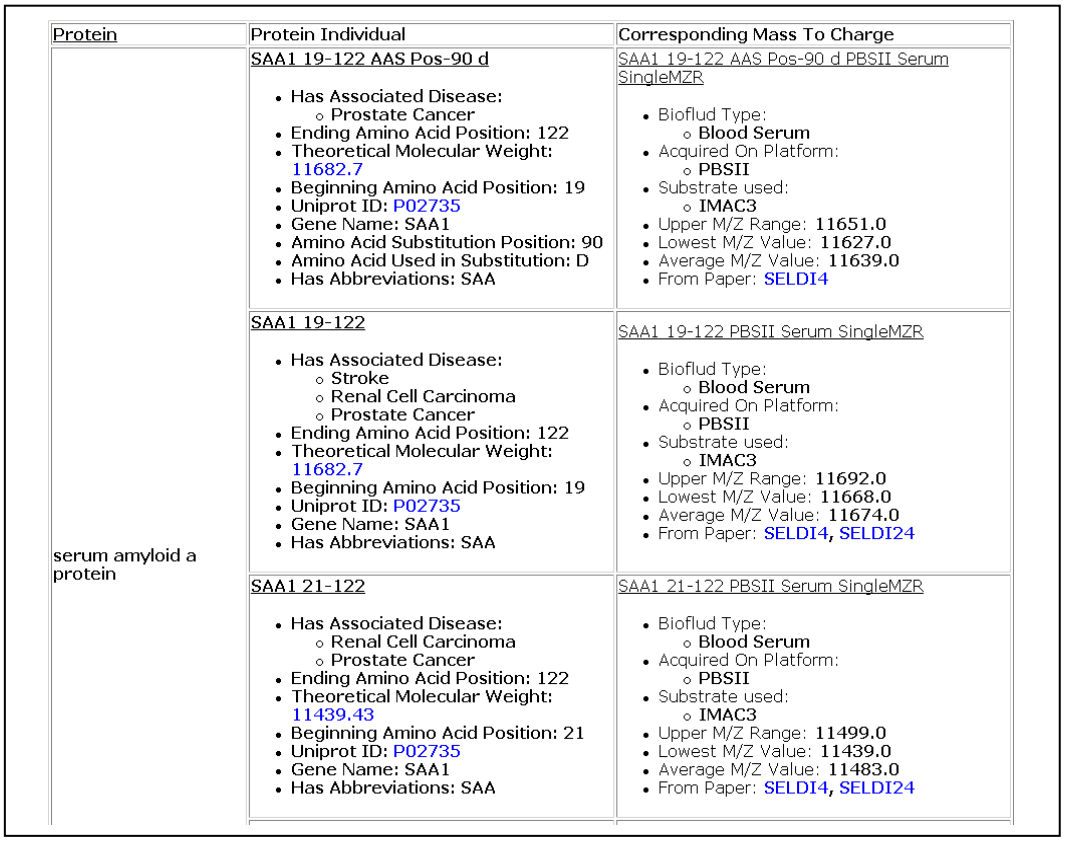
3 Analysis of proteomic data using EPO-KB

We now illustrate with examples the utility of the two methods of querying EPO-KB. A primary use of EPO-KB is for biomarker (*m/z* ratio) identification in experimental spectra. The researcher may use the EPO-KB to retrieve candidate proteins for *m/z* ratios of interest or confirm the identity of already identified *m/z* ratios. For example, the top panel in the Figure shows a partial list of candidate proteins identified by EPO-KB for *m/z* ratio 13750.

Another use of EPO-KB is in the identification of *m/z* ratios associated with a protein of interest. As an example, a researcher may want to analyze known disease specific proteins; however, identifying which *m/z* ratios are linked to those proteins is difficult. In another example, a researcher may want to identify and filter out *m/z* ratios in



Score	Protein	Protein Individual	Corresponding Mass To Charge
3.5	transthyretin	TTR 21-147 • Has Associated Disease: ◦ Frontotemporal Dementia ◦ Mycosis Fungoides ◦ Psychosis ◦ Amyotrophic Lateral Sclerosis ◦ Renal Cell Carcinoma ◦ Diabetes - Type 2 ◦ Chronic Lymphoid Malignancies • Ending Amino Acid Position: 147 • Theoretical Molecular Weight: 13761.41 • Beginning Amino Acid Position: 21 • Uniprot ID: P02766 • Gene Name: TTR • Has Abbreviations: ATTR, TBPA, TTR, Prealbumin	TTR 21-147 PBSII Serum SingleMZR • Bioflud Type: ◦ Blood Serum • Acquired On Platform: ◦ PBSII • Substrate used: ◦ WCX2 • Upper M/Z Range: 13775.0 • Lowest M/Z Value: 13747.0 • Average M/Z Value: 13754.0 • From Paper: SELDI29, SELDI24 TTR 21-147 PCS4000 Serum SingleMZR • Bioflud Type: ◦ Blood Serum • Acquired On Platform: ◦ PCS4000 • Substrate used: ◦ CM10 • Upper M/Z Range: 13809.0 • Lowest M/Z Value: 13740.0 • Average M/Z Value: 13774.0 • From Paper: SELDI99, SELDI69 TTR 21-147 ReflexII Serum SingleMZR • Bioflud Type: ◦ Blood Serum • Acquired On Platform: ◦ ReflexII • Substrate used: ◦ Sinapic • Upper M/Z Range: 13790.0 • Lowest M/Z Value: 13762.0 • Average M/Z Value: 13776.0 • From Paper: SELDI116 TTR 21-147 PBSIIc CSF SingleMZR



Protein	Protein Individual	Corresponding Mass To Charge
serum amyloid a protein	SAA1 19-122 AAS Pos-90 d • Has Associated Disease: ◦ Prostate Cancer • Ending Amino Acid Position: 122 • Theoretical Molecular Weight: 11682.7 • Beginning Amino Acid Position: 19 • Uniprot ID: P02735 • Gene Name: SAA1 • Amino Acid Substitution Position: 90 • Amino Acid Used in Substitution: D • Has Abbreviations: SAA	SAA1 19-122 AAS Pos-90 d PBSII Serum SingleMZR • Bioflud Type: ◦ Blood Serum • Acquired On Platform: ◦ PBSII • Substrate used: ◦ IMAC3 • Upper M/Z Range: 11651.0 • Lowest M/Z Value: 11627.0 • Average M/Z Value: 11639.0 • From Paper: SELDI4
	SAA1 19-122 • Has Associated Disease: ◦ Stroke ◦ Renal Cell Carcinoma ◦ Prostate Cancer • Ending Amino Acid Position: 122 • Theoretical Molecular Weight: 11682.7 • Beginning Amino Acid Position: 19 • Uniprot ID: P02735 • Gene Name: SAA1 • Has Abbreviations: SAA	SAA1 19-122 PBSII Serum SingleMZR • Bioflud Type: ◦ Blood Serum • Acquired On Platform: ◦ PBSII • Substrate used: ◦ IMAC3 • Upper M/Z Range: 11692.0 • Lowest M/Z Value: 11668.0 • Average M/Z Value: 11674.0 • From Paper: SELDI4, SELDI24
	SAA1 21-122 • Has Associated Disease: ◦ Renal Cell Carcinoma ◦ Prostate Cancer • Ending Amino Acid Position: 122 • Theoretical Molecular Weight: 11439.43 • Beginning Amino Acid Position: 21 • Uniprot ID: P02735 • Gene Name: SAA1 • Has Abbreviations: SAA	SAA1 21-122 PBSII Serum SingleMZR • Bioflud Type: ◦ Blood Serum • Acquired On Platform: ◦ PBSII • Substrate used: ◦ IMAC3 • Upper M/Z Range: 11499.0 • Lowest M/Z Value: 11439.0 • Average M/Z Value: 11483.0 • From Paper: SELDI4, SELDI24

Figure. Top: screenshot showing partial list of candidate proteins identified by EPO-KB for *m/z* ratio 13750. Bottom: screenshot showing partial list of *m/z* ratios identified by EPO-KB for serum amyloid A (UniProt identifier P02735).

the data that correspond to proteins that are known to be increased in concentration in the sample but are non-specific to the disease process being examined. In such cases, the researcher may use the EPO-KB to retrieve all *m/z* ratios in the knowledge base that correspond to a protein of interest. For example, the bottom panel in the Figure shows a partial list of *m/z* ratios identified by EPO-KB for the protein serum amyloid A.

4 Future directions

We plan to expand the EPO-KB in several ways including the addition of the ability to search for diseases that will utilize a disease ontology with unique disease identifiers. In addition, we plan to add associated microarray and single nucleotide polymorphism data to the knowledge base.

Acknowledgments

We thank Henrik Ryberg and Chad Kimmel for their assistance in populating the knowledge base, and Gary Garvin for his assistance in programming the website. We thank the Department of Biomedical Informatics for providing server space for the EPO-KB. This work was funded by a training grant from the National Institutes of Health (#5 T15 LM007059 NLM).

References

[Gopalakrishnan, *et al.*, 2004] Vanathi Gopalakrishnan, Eric Williams, Srikanth Ranganathan, Robert Bowser, Merit E. Cudkowic, Max Novelli *et al.* Proteomic data mining challenges in identification of disease-specific biomarkers from variable resolution mass spectra. Proceedings of SIAM Bioinformatics Workshop 2004:1-10.

[Lustgarten, *et al.*, 2008] Jonathan L. Lustgarten, Chad Kimmel, Henrik Ryberg, William Hogan. EPO-KB: A searchable knowledge base of biomarker to protein links. Bioinformatics 2008;24(11):1418-1419.

[The UniProt, 2007] Consortium The UniProt. The Universal Protein Resource (UniProt). Nucl. Acids Res. 2007;35(suppl_1):D193-197.

[Gasteiger, *et al.*, 2003] E. Gasteiger, A. Gattiker, C. Hoogland, I. Ivanyi, R. D. Appel, A. Bairoch. Expasy: The proteomics server for in-depth protein knowledge and analysis. Nucleic Acids Res 2003;31(13):3784-3788.

[Tolson, *et al.*, 2004] Jonathan Tolson, Ralf Bogumil, Elke Brunst, Hermann Beck, Raimund Elsner, Andreas Humeny *et al.* Serum protein profiling by SELDIi mass spectrometry: Detection of multiple variants of serum amyloid alpha in renal cancer patients. Lab Invest 2004;84(7):845-856.