Group 4 – Number 5

RUHR-UNIVERSITÄT BOCHUM

RUB

PURE

ProteomeXchange

mpc MEDIZINISCHES PROTEOM CENTER

ProCon – a Proteomics Conversion Tool for Standardized File Generation and Repository Submission

Gerhard Mayer, Christian Stephan, Helmut E. Meyer, Martin Eisenacher*

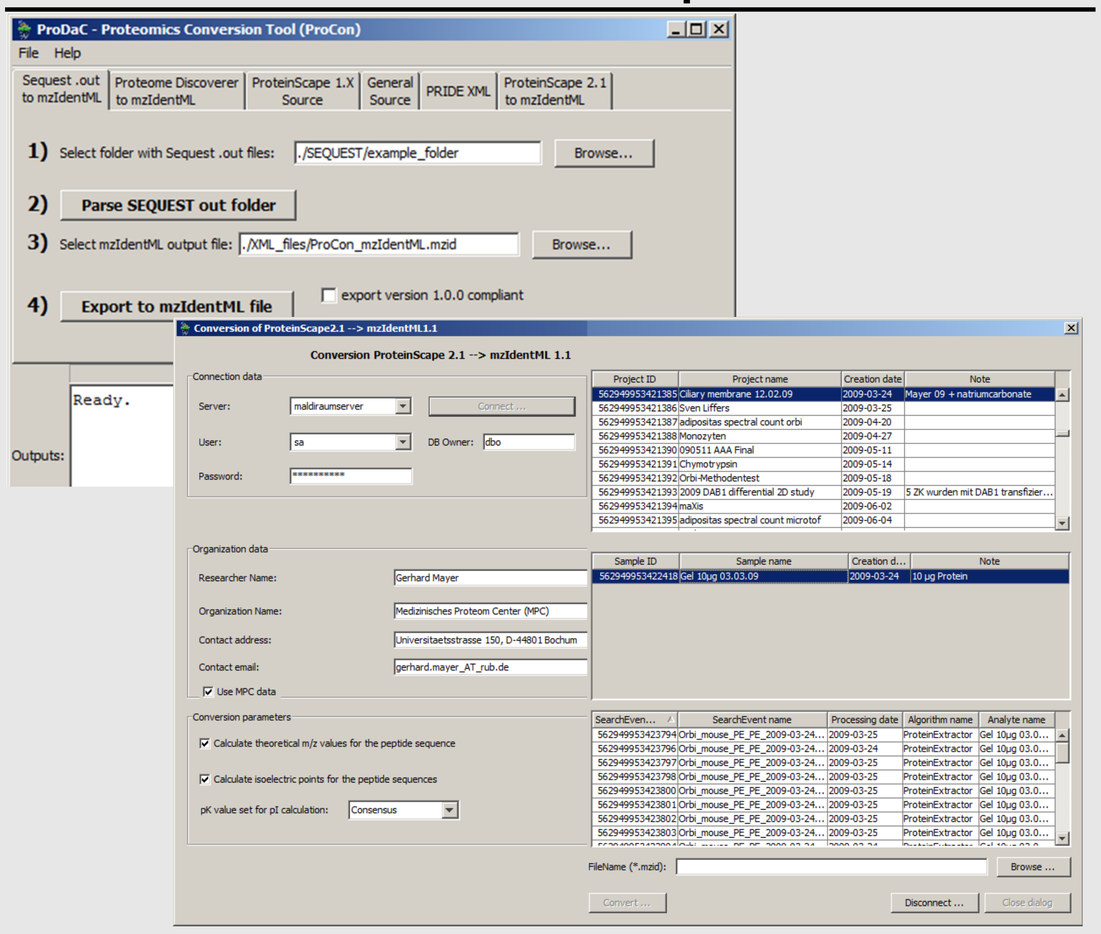
Medizinisches Proteom-Center, Ruhr-Universität Bochum, Germany, * e-Mail: martin.eisenacher@rub.de

Introduction

ProCon – A Proteomics Conversion Tool

In an EU FP6 project (ProDaC, www.fp6-prodac.eu) a major goal was the development of software tools to export, submit and import data sets using standard formats. In a current EU FP7 project (ProteomeXchange, www.proteomexchange.org) a standardized submission pipeline is developed for unique identifier generation and data exchange between large proteomics repositories. The ProCon tool, implemented at the Medizinisches Proteom-Center (MPC) in Bochum (http://www.medizinisches-proteom-center.de/ProCon), converts data into standard file formats (mzIdentML, PRIDE XML) enabling data submission through the ProteomeXchange pipeline.

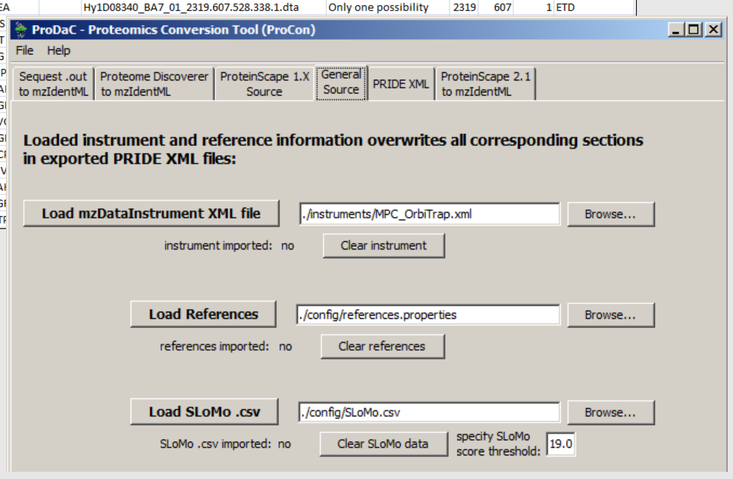
ProCon User Interface Examples



Mapping SloMo Phosphorylation Sites to PRIDE Files during ProteinScape 1.3 Conversion

Implementing a user request Phosphorylation information of the SloMo tool [2] can be integrated into a PRIDE file by loading specially formatted csv columns during ProteinScape 1.3 conversion.

Frag.Run.Cmpd.Charge	Peaks Per Window	Peptide Score	Peptide Seq	Score	Spectrum	Notes	Run	Cmpd	Charge	Fragmentation
ETD.2319.952.1			STGAS	0.00	Hy1D08340_BA7_01_2319.952.502.268.1.dta	Multiple top hits	2319	952	1	ETD
ETD.2362.1067.1			VT*AAA		Hy1D08369_BA4_01_2362.1067.512.382.1.dta	Only one possibility	2362	1067	1	ETD
ETD.2373.536.1			GS*GGR		Hy1D08378_BA1_01_2373.536.513.297.1.dta	Only one possibility	2373	536	1	ETD
ETD.2350.801.1			GAS*AK		Hy1D08362_BA5_01_2350.801.513.329.1.dta	Only one possibility	2350	801	1	ETD
ETD.2375.1327.1			GGS*GR		Hy1D08380_BA3_01_2375.1327.514.276.1.dta	Only one possibility	2375	1327	1	ETD
ETD.2373.920.1			GGGS*R		Hy1D08378_BA1_01_2373.920.514.329.1.dta	Only one possibility	2373	920	1	ETD
ETD.2315.1251.1			GGGS*R		Hy1D08336_BA3_01_2315.1251.514.357.1.dta	Only one possibility	2315	1251	1	ETD
ETD.2375.1158.1			GPGES*		Hy1D08380_BA3_01_2375.1158.525.873.1.dta	Only one possibility	2375	1158	1	ETD
ETD.2319.607.1			AGT*EA		Hy1D08340_BA7_01_2319.607.528.338.1.dta	Only one possibility	2319	607	1	ETD
ETD.2360.1604.1			AGSQS							
ETD.2348.872.1			SGADT							
ETD.2318.1137.1			AHSSG							
ETD.2315.1493.1			GAAGP							
ETD.2360.717.1			GS*VAI							
ETD.2348.1019.1			AS*AGI							
ETD.2348.990.1			AKS*VG							
ETD.2353.914.1			AGT*GI							
ETD.2348.406.1			AAT*CI							
ETD.2360.1041.1			S*VGIV							
ETD.2369.970.1			GPT*AI							
ETD.2319.834.1			PGS*GI							
ETD.2317.1020.1	2	5.27	AS*VTF							



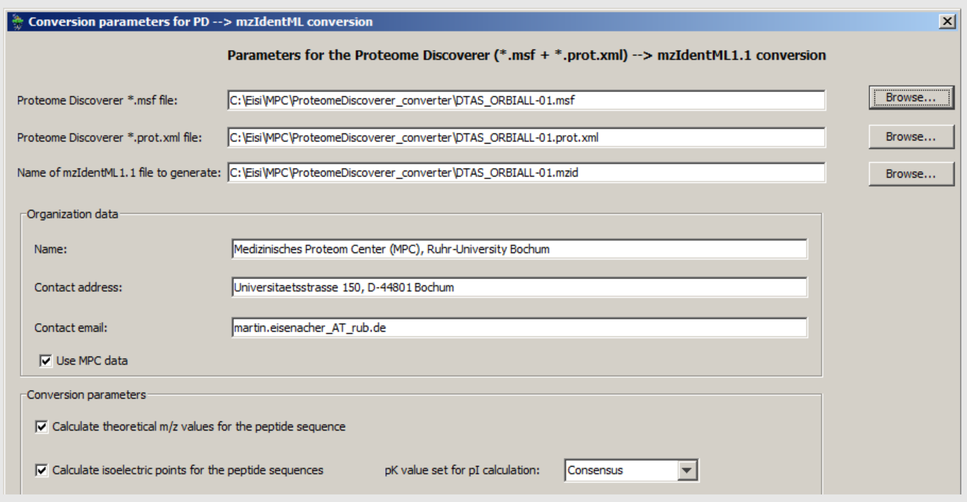
ProCon Functionality

- conversion of Proteome Discoverer® (Thermo Fisher Scientific Inc., San Jose, CA) 1.2 and 1.3 .msf and .prot.xml file information into mzIdentML
- conversion of peptide information of Sequest™ .out files (Thermo Fisher Scientific Inc., Waltham, MA, USA) into mzIdentML
- conversion of ProteinScape® (Bruker Daltonik GmbH, Bremen, Germany) version 1.3 results (SearchEvents and gels) to PRIDE XML
- conversion of ProteinScape 2.1 results (SearchEvents) to mzIdentML

A Java API was developed for accessing the SQL database content of ProteinScape. By using this API the Analyte tree of a project is parsed and the user chooses the samples and results. Then the mzIdentML file is build up in the main memory and ultimately written by using the JAXB (Java Architecture for XML binding) marshaller functionality. During this process also the relevant CV terms are built in by accessing the .obo files via the Ontology Lookup Service [1] (OLS).

ProteomeDiscoverer® to mzIdentML Conversion

After (1) setting meta-project information (such as institute and contact e-mail), (2) specifying the .prot.xml and .msf file paths, and (3) other conversion parameters (such as method for isoelectric point calculation), the conversion into mzIdentML can be initiated and is processed automatically. Internally the Open Lookup Service (OLS) is used to get the needed CV terms and the NCBI taxonomy identifiers used are acquired by accessing the Entrez Programming Utilities. (http://www.ncbi.nlm.nih.gov/books/NBK25501).



ProCon Perspectives

- The current ProteomeDiscoverer® conversion works both with ProteomeDiscoverer® versions 1.2 and 1.3 (both the .prot.xml and .msf files are needed for getting the information about the protein ambiguity groups / protein inference information). It's planned to accelerate the conversion for version 1.3, because here all the relevant information is contained and easily accessible in the .msf (SQLite) database. It will also integrate the ProteomeDiscoverer API [3].
- conversion of ProteinScape 2.1 gel results to mzIdentML
- conversion of ProteinScape 2.1 results to PRIDE XML
- mzQuantML conversion, e.g. for spectral counting workflows

Summary

ProCon complements other converters (e.g. http://code.google.com/p/pride-converter-2) to make data conversion easy even for laboratories without bioinformatics expertise. That supports journal submission and community data access. Thus meta-analyses of proteomics data and linkage to other 'omics data will lead to new insights regarding biomarkers or functional disease mechanisms.

ProCon-related Websites

http://www.medizinisches-proteom-center.de/ProCon

Submission: http://www.proteomexchange.org

Standards: http://www.psidev.info

References

1. Cote RG, Jones P, Martens L, Apweiler R, Hermjakob H. The Ontology Lookup Service: more data and better tools for controlled vocabulary queries. Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W372-6. Epub 2008 May 8 [PMID: 18467421]
2. Bailey CM, Sweet SM, Cunningham DL, Zeller M, Heath JK, Cooper HJ (2009) SLoMo: automated site localization of modifications from ETD/ECD mass spectra. J Proteome Res. 2009 Apr;8(4):1965-71 [PubMed PMID: 19275241]
3. Colaert N, Barsnes H, Vaudel M, Helsens K, Timmerman E, Sickmann A, Gevaert K, Martens L (2011) Thermo-msf-parser: an open source Java library to parse and visualize Thermo Proteome Discoverer msf files. J Proteome Res. 2011 Aug 5;10(8):3840-3 [PMID: 21714566]

Funding

This work is funded by the European Commission within the 7th European Union Research Framework Programme (Contract No.: 260558). ME is funded by P.U.R.E. (Protein Unit for Research in Europe), a project of North Rhine-Westphalia, a federal state of Germany.