

Software

Open Access

Efficient analysis and extraction of MS/MS result data from Mascot™ result files

Florian Grosse-Coosmann, Andreas M Boehm and Albert Sickmann*

Address: Protein Mass Spectrometry and Functional Proteomics Group, Rudolf-Virchow-Center for Experimental Biomedicine, Universitaet Wuerzburg, Versbacher Strasse 9, D-97078 Wuerzburg, Germany

Email: Florian Grosse-Coosmann - florian.coosmann@virchow.uni-wuerzburg.de; Andreas M Boehm - andreas.boehm@virchow.uni-wuerzburg.de; Albert Sickmann* - albert.sickmann@virchow.uni-wuerzburg.de

* Corresponding author

Published: 07 December 2005

Received: 15 July 2005

BMC Bioinformatics 2005, **6**:290 doi:10.1186/1471-2105-6-290

Accepted: 07 December 2005

This article is available from: <http://www.biomedcentral.com/1471-2105/6/290>

© 2005 Grosse-Coosmann et al; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Background: Mascot™ is a commonly used protein identification program for MS as well as for tandem MS data. When analyzing huge shotgun proteomics datasets with Mascot™'s native tools, limits of computing resources are easily reached. Up to now no application has been available as open source that is capable of converting the full content of Mascot™ result files from the original MIME format into a database-compatible tabular format, allowing direct import into database management systems and efficient handling of huge datasets analyzed by Mascot™.

Results: A program called mres2x is presented, which reads Mascot™ result files, analyzes them and extracts either selected or all information in order to store it in a single file or multiple files in formats which are easier to handle downstream of Mascot™. It generates different output formats. The output of mres2x in tab format is especially designed for direct high-performance import into relational database management systems using native tools of these systems. Having the data available in database management systems allows complex queries and extensive analysis. In addition, the original peak lists can be extracted in DTA format suitable for protein identification using the Sequest™ program, and the Mascot™ files can be split, preserving the original data format. During conversion, several consistency checks are performed. mres2x is designed to provide high throughput processing combined with the possibility to be driven by other computer programs. The source code including supplement material and precompiled binaries is available via <http://www.protein-ms.de> and <http://sourceforge.net/projects/protms/>.

Conclusion: The database upload allows regrouping of the MS/MS results using a database management system and complex analyzing queries using SQL without the need to run new Mascot™ searches when changing grouping parameters.

Background

For instance, protein identification via MDLC combined with tandem mass spectrometry techniques or other shotgun approaches usually generate huge data sets and compels application of software programs such as Sequest™

[1], Profound [2] or Mascot™ [3]. This produces peptide sequences that need to be grouped in order to obtain protein identifications with several peptides per hit, which increases reliability of the results. Mascot™ groups the peptide results of a single search run automatically.

```

MIME-Version: 1.0 (Generated by Mascot version 1.0)
Content-Type: multipart/mixed; boundary=gc0p4Jq0M2Yt08jU534c0p

--gc0p4Jq0M2Yt08jU534c0p
Content-Type: application/x-Mascot; name="parameters"

LICENSE=Licensed to: Rudolf-Virchow-Center for Experimental Biomedicine,
University Wuerzburg, (2 processors).
MP=
NM=
COM=/tmp/3645/LCQ4529.491.491.2.dta
IATOL=0
IA2TOL=
[...]
--gc0p4Jq0M2Yt08jU534c0p
Content-Type: application/x-Mascot; name="masses"
[...]
--gc0p4Jq0M2Yt08jU534c0p
Content-Type: application/x-Mascot; name="header"
[...]
--gc0p4Jq0M2Yt08jU534c0p
Content-Type: application/x-Mascot; name="summary"

qmass1=1328.823873
qexp1=665.419212,2+
qmatch1=3708747
qplughole1=38.637717
num_hits=50
hl q1|23059634,2.78e+01,1.00,14998.54
hl_text=|ref|ZP_00084586.1| hypothetical protein [Pseudomonas fluorescens
Pfo-1]
hl q1=0,1328.693527,0.130346,54,65,5.00,DLVDLDGEIGRK,11,00000000000000,26.1
8,3,000000200000000000,0,0,10935629.000000
[...]
--gc0p4Jq0M2Yt08jU534c0p
Content-Type: application/x-Mascot; name="peptides"
q1 p1=0,1328.693527,0.130346,5,DLVDLDGEIGRK,11,0000000000000000,26.18,0000020
00000
00000,0,0;"gi|23059634":0:54:65:3
[...]
--gc0p4Jq0M2Yt08jU534c0p
Content-Type: application/x-Mascot; name="proteins"
"gi|23059634"=14998.54,"|ref|ZP_00084586.1| hypothetical protein
[Pseudomonas
fluorescens Pfo-1]"
--gc0p4Jq0M2Yt08jU534c0p
Content-Type: application/x-Mascot; name="query1"
mass_min=209.800000
mass_max=1229.600000
int_min=1.000000
int_max=3307740.000000
num_vals=194
num_used1=-1
Ions1=228.000000:3.664e+05,327.300000:4.716e+05, [...]
--gc0p4Jq0M2Yt08jU534c0p
Content-Type: application/x-Mascot; name="index"
[...]

```

Protein Identification Hit

Peptide Hit

List of Proteins

Input Spectra („Queries“)

Figure 1 An example of the MIME format of Mascot™ result files is shown in this figure. Wrapped lines are indented. Some lines are removed due to space savings, marked by [...]. The original example file contains 322 lines. Cross-reference links are marked in red.

Table 1: The command line options of mres2x. Parameters for setting the Mascot™'s username, changing line break characters as well as debugging mode exist, too. The usage of mres2x is: mres2x -d destination -o type [-rvfpSuU] filemask_of_input_files, where the last parameter defines the input file(s) including the path and can even be a single file. The input must be in original Mascot™ format, not HTML. The files from the file mask must be in the same directory if the output format is not tab. In case of tab format output, the destination must be a single file, otherwise a folder. mres2x explicitly expands input file masks. A description of the parameters also can be found in the file Overview.html (see additional file 1) included in the source code package.

Option	Parameters	Explanation
type	s_dta m_dat Tab	Describes the output format. Supported types are: s_dta Sequest™'s dta format. Only spectra data will be exported. m_dat split the input into several output files in Mascot™'s output format, one for each query. Tab write out a tabbed format for direct database upload.
-r		Use CR LF instead of LF as linefeed in data blocks. Some OS need special line feed characters in text files.
-v		Increase verbosity mode by one per occurrence of -v. A maximum of two -v is allowed.
-f		Overwrite files/allow usage of non-empty directories. Usually, the destination directory must be empty.
-p		Preserves files on unsuccessful program termination. Useful for debugging purposes.
-S		Show message indicator even if stderr is a terminal.
-u	name	Set the username to name, if no entry is present and if the tab output format is selected.
-U	name	Set the username to name in all cases if the tab output format is selected. This allows changing the username in the result files of Mascot™.

Recombination and merging of search runs is not supported. The data volume limits of Mascot™'s result display tool defined by the underlying computing resource are easily reached and exceeded when applied to a shotgun approach, excluding the opportunity to analyze a huge MDLC experiment at once.

Generally, scientists require their protein identification results in tabular format in order to visualize, filter or sort them by several criteria. Concerning Sequest™, some open source tools for extracting data from its result files already exist, such as Out2Summary from the SASHIMI Project[4] or Sequest Browser™ [1]. For Mascot™, which produces text files in MIME format [[5-10]], such a tool is currently not available as open source. Tools like the ExtParser module integrated in Phenyx [11] convert the preprocessed HTML output of Mascot™'s result display tool rather than the original result file. The parser Mascot2XML of SASHIMI project[4] reads original Mascot™ data and converts into pepXML [12]. This program is available as open source, but does not export all information contained in the Mascot™ file.

For efficient import in spread sheet applications and relational database systems, a straight-forward format is needed, in order to achieve the best performance.

The MIME format of Mascot™ result files looks as shown in figure 1. Obviously, this format cannot be imported into spread sheet applications or database programs because it contains internal references.

Here, we present the command line tool mres2x that is capable of converting results from original, unprocessed MIME formatted Mascot™ output files (extension .DAT)

into a comprehensive tabular format. Extraction of included peak lists into Sequest™'s DTA format is supported, too. Another option allows splitting the original Mascot™ output into several files in Mascot™'s native format according to the number of series of measurements.

An example of running mres2x on Unix/Linux producing tab format output in **mascot.tab** of the file **mascotresult.res** stored in **/tmp** is the following command line:

```
./mres2x -d ./mascot.tab -o tab /tmp/mascotresult.res
```

Implementation

mres2x is implemented in C [13], and therefore is portable to several platforms, most notably Windows™ and Unix™. It offers a command line interface. All functionality is controlled by command line parameters that are shown in table 1. A detailed documentation of all command line switches is given in the file **Overview.html** (see additional file 1) included in the source code package.

The program uses one or more Mascot™ result files as input for processing. Its output can be directed to the program's standard output or to a file in case of the tabular output format. Otherwise, an existing directory must be specified as output destination. The converted tabular format is up to 40 percent smaller in size than the original data without any loss of information. It is designed for direct import into relational database management systems, but also can be used with spread sheet applications or other programs for further processing and validation. The tabular format is documented extensively in the file

tabformat.html (see additional file 2), where the format of the original Mascot™ result files is implicitly documented, too.

mres2x can be used to split huge Mascot™ result files into single files using the **-o s_dat** switch, each containing a single query and its corresponding results. This increases performance of reusing the separated results. Typical examples of use are display, analysis or validation by standard tools, such as the bundled result browser of Mascot™. Nevertheless, the main purpose of mres2x is conversion of huge MIME formatted files into a more readable and compact format for efficient direct import into database management systems, using their native import tools.

Several data analysis steps are performed in order to check the validity of Mascot™ files even while processing the input data. Values are checked for their range at this stage. The most detailed validation is performed when producing the tabular format. A full cross-reference check is performed here. Thereby, it is assured that the output is fully consistent. The cross-referenced structure is shown in figure 1.

In case of errors, a cleanup is performed which removes any result files produced so far and the OS is informed by a non-zero return code of the program. It depends on the error whether further analysis of the input file is performed by mres2x. If possible, the algorithm collects errors and prints them out before termination. If included in the input file, Mascot™ warnings are printed to standard error and are available by the calling program.

On success, the message "**Operation ended successfully.**" is written to the standard output. Wrapping programs can easily test for this message or for a return code of zero.

An example of the output in tabular format is displayed in figure 2. The success codes (in this case E1, O0) at the end of a query section (B to E) or a file section (I to O) allow usage of database transaction rollback in case of errors.

mres2x has thoroughly been tested with several thousands of data sets produced by Mascot™ version 2.1.0 and earlier.

Results

We compared the performance of mres2x with the result viewer named **master_results.pl** that comes with Mascot™ version 2.0.04, using a 368.74 megabytes large MIME formatted Mascot™ result file, containing 1,565,945 lines obtained from 60,000 MS/MS spectra. Conversion of this file with mres2x lasts 1 minute, 20 sec-

onds, whereas display of this file with **master_results.pl** using the binary library **msparser** coming with Mascot™ takes more than 15 minutes on the same computer; the version fully implemented in Perl of **master_results.pl** would be even slower.

Conclusion

We introduced a tool capable of converting Mascot™ result files efficiently into other formats, most notable the one designed for direct database import. mres2x is designed to provide high throughput processing combined with extensive error checking and the possibility to be driven by other computer programs. Therefore, mres2x is suitable for integration into computer automated high throughput environments, using direct import into database management systems.

mres2x reads Mascot™ result files and extracts all information in order to store it to another file or files. It currently supports three output formats: First, the original Mascot™ output file can be split into several files with the same format according to the number of series of measurements. Second, the original input peak lists can be extracted into DTA format. Third, a file in tabular format for direct bulk database upload can be created.

In contrast to other formats, such as pepXML [12], proXML [12] and mzIdent [12], mres2x avoids the overhead implied by the need of interpreting the intermediate XML format over again. This allows for importing data directly in a relational database system or spread sheet applications. XML is a storage space consuming format [14] and parsing and interpretation of XML is a time consuming task, decreasing performance of the whole process [15]. Same with other intermediate formats, such as SQT [14]. The tab output format of mres2x is not intended to meet all requirements of the currently discussed file format standardization [[16-19]] and is not designed as a substitute of either XML format mentioned before. mres2x is designed to be used for direct bulk database uploads of Mascot™ results by means of the corresponding database management system, such as **SQL*Loader** of Oracle™ or **bcp** of SQL Server™. However, it creates an easy to parse tabular format which makes the creation of translating software to produce other formats nearly trivial. This allows export to any other industry standard.

Storing the results in a database management system allows efficient complex queries on the data such as regrouping of peptide results to protein results without need to research the MS/MS data again and yields time and resource savings as well as increased flexibility.

```

h "PROGRAM"; "CODEPAGE"; "PRG_VERSION"; "DATA_VERSION"; "MAX_FILES"
i "TYPE"; "AVERAGE"; "CLEAVAGE"; "DB1"; "DB2"; "FILENAME"; "PROGRAM"; "ICAT"; "INST
RUMENT"; "SEARCHES"; "FRAGMENT_TOL"; "FRAGMENT_TOLU"; "PEPTIDE_TOL"; "PEPTIDE_
TOLU"; "VARIABLE_MODS"; "FIXED_MODS"; "PFA"; "USER"; "TIMESTAMP"; "IDENTITY_THR
ES"; "QUERIES"; "COMMENT"; "CHARGE"; "SEG"
b "QUERY"; "CHARGE"; "MASS"; "PRECURSOR"; "MATCH"; "IDENTITY_THRES"; "HOMOLOGY_TH
RES"; "TITLE"; "PEAKLIST"
p "PROTEIN"; "HITNUMBER"; "TOTAL_SCORE"; "TOTAL_MASS"; "MISSED_CLEAVAGE"; "QUERY
"; "PEPTIDE"; "PEPTIDE_MASS"; "PEPTIDE_START"; "PEPTIDE_SCORE"; "OCCURANCES"; "
MATCHING_FRAGMENTS"; "MATCHING_PEAKS"; "SERIES_FOUND"; "SERIES_FOUND_STR"
f "PROTEIN"; "PROTEIN_NUMBER"; "HITNUMBER"; "TOTAL_MASS"; "MISSED_CLEAVAGE"; "QU
ERY"; "PEPTIDE"; "PEPTIDE_MASS"; "PEPTIDE_START"; "PEPTIDE_SCORE"; "OCCURANCES
"; "MATCHING_FRAGMENTS"; "MATCHING_PEAKS"; "SERIES_FOUND"; "SERIES_FOUND_STR"
e "SUCCESS"
o "SUCCESS"
t "RETURNCODE"; "LISTED"
H "MRES2X"; "ISO-8859-15"; "1.64"; "2.00"; "1
IO "MIS"; "0"; "None"; "NCBIInr_030819"; "NCBIInr_030819.fasta 484217701"; "LCQ4529.
491.491.2"; "MASCOT 2.0.04"; "0"; "Default"; "1,2,5,6,8,9,13,14"; "0.100000"; "Da";
0.200000"; "Da"; " "; " "; "5"; "seqdaemon"; "1092825009"; "31.234674"; "1"; "/tmp/3645/LCQ45
29.491.491.2.dta"; "0+"; " "
B1 1;2;1328.823873;665.419212;3708747;65.692272;38.637717;"; "228.000000:3.
664e+05,327.300000:4.716e+05,442.300000:6.463e+05,602.300000:6.895e+05
[...]
P0 "gi|23059634"; "1;2.78e+01;14998.54;0;1; "DLVLDLGEIGRK"; "1328.693527;54;26.1
8;3;5;11; "000000200000000000"; "y"
P1 [...]
F0 "gi|23059634"; "0;1;14998.54;0;1; "DLVLDLGEIGRK"; "1328.693527;54;26.18;3;5;1
1; "000002000000000000"; "b++"
F1 [...]
E1 1
O0 1
T 0;1
    
```

Description & Parameters

Input Spectra („Queries“)

Protein Identification Hit

Peptide Hit

Figure 2

An example of the output of mres2x in tabular format, one record per line. The lines begin with a prefix, indicating the line type. Lower case letters indicate description lines; the corresponding data has upper case letters as prefix which may be directly followed by running numbers. The format is semicolon separated. Chapters are marked and commented in red. The format is described in the file Tabformat.html (see Additional file 2), included in the source code package.

As the tab output format contains one result record per line, filtering and processing directly after conversion is easily possible, such as for false positives as well as allowing for assembling identifications. The records of protein and peptide results can be distinguished after processing, as the first character of each line indicates the record type.

Availability and requirements

For compilation a standard C compiler is needed. mres2x can be compiled and run on Windows™ and Unix/Linux. The program is freely available via <http://www.protein-ms.de> and <http://sourceforge.net/projects/protms/> for download.

Authors' contributions

FGC implemented the program and made a draft of the manuscript. AB initiated the development. AB and AS contributed with ideas and proofread the manuscript. AB supervised the final testing. All authors have read and approved the final manuscript.

List of abbreviations used

- HTML hypertext markup language
- MDLC multidimensional liquid chromatography
- MIME multipurpose internet mail extensions

MS mass spectrometry

OS operating system of a computer

SQL structured query language

XML extended markup language

Additional material

Additional File 1

Documentation of *mres2x*, this document describes *mres2x* and how to use it.

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2105-6-290-S1.pdf>]

Additional File 2

Format of the tab output format, this document describes the output of *mres2x* when tab format is selected. It implicitly documents the format of Mascot™'s result files.

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2105-6-290-S2.pdf>]

Acknowledgements

This work was supported by the Deutsche Forschungsgemeinschaft (SI 835/3-1; FZT82).

References

- Eng JK, McCormack AL, Yates JR: **An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database.** *Journal of the American Society for Mass Spectrometry* 1994, **5**:976-989.
- Zhang W, Chait BT: **ProFound: An Expert System for Protein Identification Using Mass Spectrometric Peptide Mapping Information.** *Analytical Chemistry* 2000, **72**:2482-2489.
- Perkins DN, Pappin DJC, Creasy DM, Cottrell JS: **Probability-Based Protein Identification by Searching Sequence Databases Using Mass Spectrometry Data.** *Electrophoresis* 1999, **20**:3551-3567.
- Pedrioli PGA: **The SASHIMI Project.** [<http://sashimi.sourceforge.net/>].
- Freed N, Borenstein N: **RFC 2045 - Multipurpose Internet Mail Extensions (MIME) Part One: Format of Internet Message Bodies.** [<http://www.faqs.org/rfcs/rfc2045.html>].
- Freed N, Borenstein N: **RFC 2046 - Multipurpose Internet Mail Extensions (MIME) Part Two: Multipurpose Internet Mail Extensions.** [<http://www.mhonarc.org/~ehood/MIME/2046/rfc2046.html>].
- Freed N, Borenstein N: **RFC 2047 - Multipurpose Internet Mail Extensions (MIME) Part Three: Message Header Extensions for Non-ASCII Text.** [<http://www.mhonarc.org/~ehood/MIME/2047/rfc2047.html>].
- Freed N, Borenstein N: **RFC 2048 - Multipurpose Internet Mail Extensions (MIME) Part Four: Registration Procedures.** [<http://www.mhonarc.org/~ehood/MIME/2048/rfc2048.html>].
- Freed N, Borenstein N: **RFC 2049 - Multipurpose Internet Mail Extensions (MIME) Part Five: Conformance Criteria and Examples.** [<http://www.mhonarc.org/~ehood/MIME/2049/rfc2049.html>].
- Masinter L: **RFC 2388 - Returning Values from Forms: multipart/form-data.** [<http://www.faqs.org/rfcs/rfc2388.html>].
- GeneBio: **GeneBio Phenyx.** [<http://www.phenyx-ms.com/>].
- Keller A, Eng J, Zhang N, Li X, Aebersold R: **A Uniform Proteomics MS/MS Analysis Platform Utilizing Open XML File Formats.** *Molecular Systems Biology* 2005.
- Kernighan BW, Ritchie DM: **The C Programming Language.** 2nd edition. , Prentice-Hall Int.; 1990.
- McDonald WH, Tabb DL, Sadygov RG, MacCoss MJ, Venable J, Graumann J, Johnson JR, Cociorva D, Yates JR: **MS1, MS2, and SQT - three Unified, Compact, and Easily Parsed File Formats for the Storage of Shotgun Proteomic Spectra and Identifications.** *Rapid Communications in Mass Spectrometry* 2004, **18**:2162-2168.
- Boehm AM, Galvin RP, Sickmann A: **Extractor for ESI Quadrupole TOF Tandem MS Data Enabled for High Throughput Batch Processing.** *BMC Bioinformatics* 2004, **5**.
- Orchard S, Hermjakob H, Binz PA, Hoogland C, Taylor CF, Zhu W, Jr. RKJ, Apweiler R: **Further Steps Towards Data Standardisation: The Proteomic Standards Initiative HUPRO 3rd Annual Congress, Beijing 25-27th October, 2004.** *Proteomics* 2005, **5**:337 -3339.
- Orchard S, Hermjakob H, Julian RK, Runte K, Sherman D, Wojcik J, Zhu W, Apweiler R: **Common Interchange Standards for Proteomics Data: Public Availability of Tools and Schema Report on the Proteomic Standards Initiative Workshop, 2nd Annual HUPRO Congress, Montreal, Canada, 8-11th October 2003.** *Proteomics* 2004, **4**:490 -4491.
- Orchard S, Zhu W, Julian RK, Hermjakob H, Apweiler R: **Further Advances in the Development of a Data Interchange Standard for Proteomics Data.** *Proteomics* 2003, **3**:2065-2066.
- Taylor CF, Paton NW, Garwood KL, Kirby PD, Stead DA, Yin Z, Deutsch EW, Selway L, Walker J, Riba-Garcia I, Mohammed S, Deery MJ, Howard JA, Dunkley T, Aebersold R, Kell DB, Lilley KS, Roepstorff P, Yates JR, Brass A, Brown AJP, Cash P, Gaskell SJ, Hubbard SJ, Oliver SG: **A Systematic Approach to Modeling, Capturing, and Disseminating Proteomics Experimental Data.** *Nature Biotechnology* 2003, **21**:247 -2254.

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp

