SHORT COMMUNICATION

Do we want our data raw? Including binary mass spectrometry data in public proteomics data repositories

Lennart Martens[1], Alexey I. Nesvizhskii[2], Henning Hermjakob[3], Marcin Adamski[4], Gilbert S. Omenn[4], Joël Vandekerckhove[1] and Kris Gevaert[1]

[1] Department of Biochemistry, Faculty of Medicine and Health Sciences, Ghent University, Ghent, Belgium
[2] Institute for Systems Biology, Seattle, WA, USA
[3] EMBL Outstation, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK
[4] Department of Human Genetics, University of Michigan, Ann Arbor, MI, USA

With the human Plasma Proteome Project (PPP) pilot phase completed, the largest and most ambitious proteomics experiment to date has reached its first milestone. The correspondingly impressive amount of data that came from this pilot project emphasized the need for a centralized dissemination mechanism and led to the development of a detailed, PPP specific data gathering infrastructure at the University of Michigan, Ann Arbor as well as the protein identifications database project at the European Bioinformatics Institute as a general proteomics data repository. One issue that crept up while discussing which data to store for the PPP concerns whether the raw, binary data coming from the mass spectrometers should be stored, or rather the more compact and already significantly processed peak lists. As this debate is not restricted to the PPP but relates to the proteomics community in general, we will attempt to detail the relative merits and caveats associated with centralized storage and dissemination of raw data and/or peak lists, building on the extensive experience gained during the PPP pilot phase. Finally, some suggestions are made for both immediate and future storage of MS data in public repositories.

Received: October 27, 2004
Revised: January 14, 2005
Accepted: March 1, 2005

Keywords:
Bioinformatics / Databases / Mass spectrometry

The completion of the human genome project, with the corresponding rise of the field of proteomics, led to the creation of the HUPO projects as the next major collaborative scientific enterprise in the life sciences [1]. In order to achieve the high-aiming goals of these projects in a reasonable time frame, collaborations between multiple labs around the world have been set up, with each of these labs analyzing standard samples using distinct protocols and hardware. The Plasma Proteome Project (PPP), as the pioneering project in the larger HUPO consortium, is the first of these to have amassed a large body of proteomics data during its recently completed pilot phase [2]. Centralized data storage and subsequent dissemination of these data to the scientific community has been addressed through the initial data collection and management work of Marcin Adamski at the University of Michigan, Ann Arbor [3] and the protein identification database (PRIDE) [4] project of the European Bioinformatics Institute (EBI). During the construction of these resources, a lot of discussion was attributed to the storage of the MS data. In particular the storage of the raw, binary data that the machines report has been discussed thoroughly.

As the question of storing raw data has recently been taken up by editors of proteomics journals as well [5], and furthermore affects the proteomics community at large [6],

Correspondence: Dr. Lennart Martens, Department of Biochemistry, Faculty of Medicine and Health Sciences, Ghent University, A. Baertsoenkaai 3, B-9000 Ghent, Belgium
E-mail: lennart.martens@UGent.be
Fax: +32-9264-9484

Abbreviations: EBI, European Bioinformatics Institute; **GNU**, GNU's Not Unix; **GZIP**, GNU ZIP; **PPP**, Plasma Proteome Project; **PRIDE**, proteomics identifications database; **RAID**, redundant assay of inexpensive disks we here present a series of advantages and limitations inherent to the publication of raw data compared to processed peak lists, building on the unique experiences obtained through the PPP.

There seems to be a general consensus in the proteomics community today to request submission of the source data on which reported identifications are based [5]. This will allow other researchers to verify and validate the published conclusions independently. Publishing source data also has the benefit of allowing additional (computational) analyses by other researchers, which could lead to the uncovering of new, biologically relevant information that was missed in the original analysis.

These source data can take a number of forms, but by far the most common representations are either the proprietary, binary "raw" formats that the mass spectrometers churn out during their analyses or the text-based, processed peak lists that are typically submitted to search engines for identification of the peptides that produced those spectra. In the case of fragmentation spectra, the peak lists contain the parent peptide m/z and charge (if the charge is known) and a listing of measured m/z values and their intensities for the fragment peaks. Search engines then attempt to match these fragment peaks to *in silico* generated fragmentation spectra of all peptides in a search database. The peak lists are often called MS/MS spectra and due to the extensive automation of acquisition software, they are often the only format encountered by researchers. These files can take a variety of formats, yet all are essentially text-based, small (a few kilobytes *per* file), readily readable by both humans and software programs and easily compressible (two-fold to three-fold compression ratios are routine using GNU ZIP (GZIP) (GNU – GNU's Not Unix)). Additionally, each of these peak list formats can conveniently be transcribed in any other format. A few common examples are SEQUEST files (dta), Micromass peak lists (pkl), and MASCOT Generic Format files (mgf). There is a slight variability in the amount of information these different formats can accommodate, but in general conversion between formats tends to be conservative. Furthermore, the mzData format, a community standard recently developed by the HUPO Proteomics Standards Initiative (PSI) [7] that elicits broad support among both instrument and software vendors, will ultimately eliminate the need for these format conversions.

As noted above, peak lists present an already processed view on the originally recorded data. Typically proprietary, vendor-supplied software is used to extract these peak lists from the raw data. Frequently applied processing techniques during this extraction phase include noise-filtering, centroiding, deconvolution, and deisotoping of the peaks. As there is no standard protocol for these processing steps, problems often arise because what one scientist regards as standard processing might seem "lossy" conversion to another, leading some to label these peak lists as an unfit distribution medium for MS data.

The raw data formats in contrast are much larger in size (typically well above 10 MB *per* file) and are usually stored in a proprietary, binary format. This makes the files impractical to read for both users and third-party software programs, all the more so because the exact format description is typically not disclosed by the vendors. Since the binary format can already be a compressed representation of the data, standard compression algorithms such as GZIP do not always reduce the size of these files. A simple analysis was performed to illustrate both size differences and the effects of data compression (Fig. 1). The much larger size of the raw data does, however, allow these files to contain much more information than peak lists. Raw files contain all the individual peaks as registered by the instrument detector and, for LC-MS machines, can store elution profiles and times for the LC part. Depending on the vendor and make of the machine, other useful instrument-related information can be stored in these files as well.

Recently, several interesting developments have been described that can put this wealth of additional information in raw files to good use [8, 9]. The key to interpreting these raw data directly has been the development of specific software to parse the binary content of these raw files into intelligible data, a tedious and time-consuming task that typically needs to be redone each time a new machine or a new version of an existing machine or its operating software appears. Furthermore, this reverse-engineering of a proprietary format is typically frowned upon by vendors. Next to the above-mentioned caveats associated with proprietary raw data formats, there is also the very real problem of "aging" that comes with any binary formatted data. As time goes by, support for certain formats tends to evaporate and within the space of several years, readers can no longer be found for the format. A detailed review of the issues concerning proprietary data formats and science can be found in [10].

The mzXML format of the Institute of Systems Biology [11], designed as an intermediate format between raw data and peak lists, could bring some solace if it were supported by vendors, but a more pervasive effort on behalf of the entire community to standardize raw data formats is more likely to succeed in eliciting such global support.

When it comes to storing mass spectrometric data in proteomics data repositories, the discussion tends to focus on an "either-or" decision. Most proponents for the storage of raw data currently have (limited) facilities to parse this kind of data, and are therefore able to exploit the richer information therein. The other camp, which advocates the storage of the processed peak lists, tends to lack this software, making the raw data essentially inaccessible to them (unless they happen to possess the particular, proprietary instrument software that allows the transformation to peak lists). It is our opinion that the choice should not be an exclusive one. In fact, we are convinced that both formats have a distinct and additive value at this time and as such fulfill complementary roles.

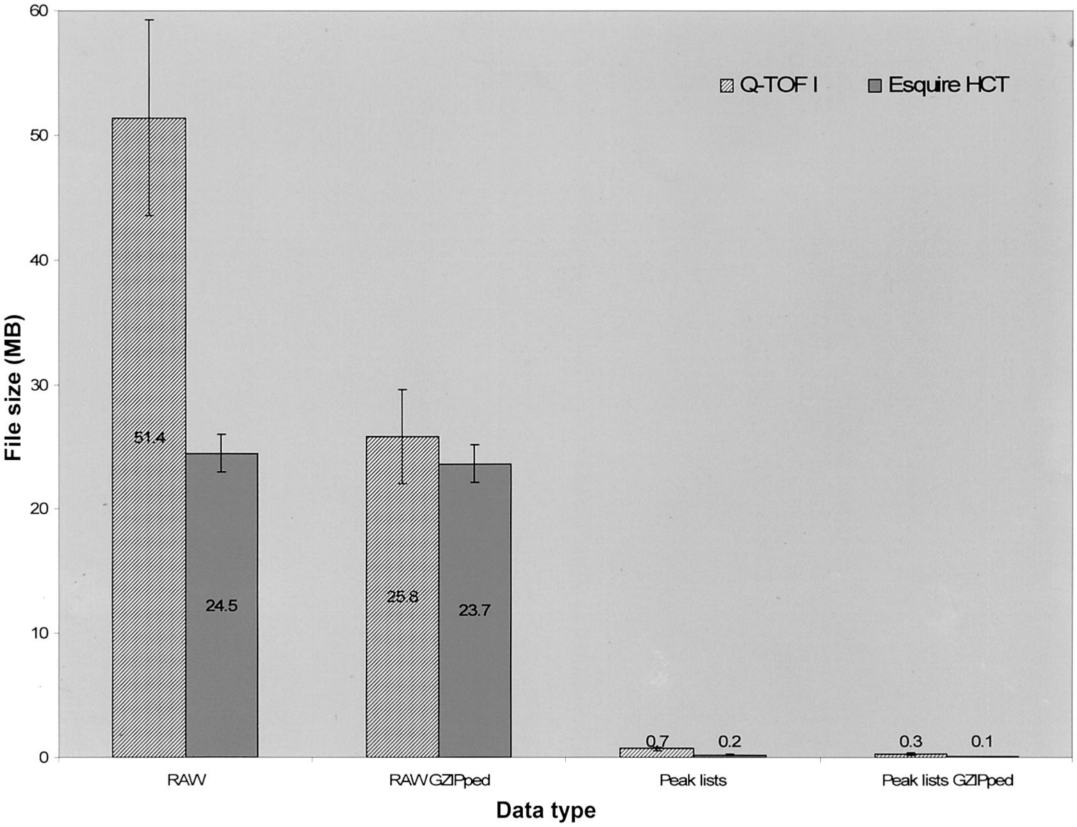

Figure 1. Comparing compressed and uncompressed file sizes for RAW data and the corresponding peak lists. Figures for the data are based on the averages of multiple separate files for each measurement. Error bars denote one SD on the averages. For the raw data, the sizes were averaged over ten individual files. Q-TOF I (Micromass, Cheshire, UK) peak list data consist of 720 individual files, the Esquire HCT (Bruker Daltonik, Bremen, Germany) IT peak list data count 1050 distinct files. Both file sets were grouped into ten subgroups, with each subgroup corresponding to the spectra extracted from a single parent raw file. File format chosen for the peak lists was the intermediately verbose MASCOT Generic Format (http://www.matrixscience.com/help/data_file_help.html). Peak lists have been tarred by GNU tar (http://www.gnu.org) to compensate for size-bloating due to the minimal file size limit of the NTFS file system. Compression for both RAW files and peak lists was done using GZIP with default compression settings. Note the extreme difference in file sizes between raw data files and peak lists. Also notable is the difference in compression efficiency between Q-TOF I RAW files and their Esquire HCT counterparts, especially since the compressed results are highly similar, indicative of a built-in compression in the Esquire HCT files. Compressibility of the peak lists can be deduced from the data labels and is always greater than 50%.

When a reevaluation of the peak lists using a different search algorithm or using a newer sequence database as search base is the scope of the research done with the original data, peak lists typically are the most readily accessible and efficient sources of MS data. For more advanced purposes however, such as obtaining large training sets for machine learning approaches for the prediction of peptide elution times [12] or, in the case of quantitative proteomics experiments based on stable isotope labeling [8], the raw formats present the only data source rich enough for these analyses.

Therefore, in the PPP, peak lists are part of the core data structure, whereas submission of raw files is considered an optional yet highly encouraged addition. The reason for this optional inclusion of raw data is purely technical in origin, as the sheer size of the files involved pushes infrastructure requirements for both storage of the data and their subsequent distribution to their limits.

Typically, funding for these infrastructure issues is evaluated using a standard cost/benefit model, yet for raw data files, the costs will surely outweigh the benefits in the short term. Storing raw files will require large amounts of disk space, which typically should be made redundant (*e.g.*, using RAID systems), thus disk space requirements will be at least twice the size of the data. Back-ups of this amount of data also present a nontrivial challenge. Due to typical low compression ratios, the amount of uncompressed tape media space (which tends to be more expensive than hard drive space) required will be roughly equivalent to the total data size. The distribution of the data after they have been successfully stored, also accounts for a large part of the cost involved since bandwidth does not come free, either. As an illustration of the data storage requirements, we consider the raw data for a single ICAT [13] or COFRADIC [14] run through a complete proteome (30–40 separate LC-MS/MS runs, with a 2 h gradient each) to have a compressed size of roughly 1.5 GB for older or less sophisticated machines, up to a massive 45 GB for newer, state-of-the-art instruments! It can be expected that future machines will generate even larger files as instrument accuracy and resolution increases. Put in perspective, a single proteome thus requires at least three times as much storage space as the NCBI nonredundant protein database (ftp://ftp.ncbi.nih.gov/blast/db/nr.tar.gz) in FASTA format, or three times as much as the full Swiss-Prot database [15] in the native text format! And although a 100 GB low end hard disk can currently be purchased for about US $100, a conservative cost estimate from the EBI averages to a total cost of US $2000 *per* 100 GB stored for data on a public high-availability FTP server, including distribution and back-up costs!

Even though a truly distributed system (every lab hosting its own raw data) maximizes cost-efficiency through distribution of both the storage and bandwidth cost, it is typically undesirable in the long run as the turn-over for availability of academic sites tends to be quite high. The installation of centralized repositories, located at dedicated institutes such as the EBI or the National Center for Biotechnology Information (NCBI), would be far more reliable in the long run, yet these organizations typically suffer from a lack of resources to host this amount of data. Compared to sequence databases, for instance, the growth in data storage requirements (and hence the rise of the cost) will be far greater for raw data, whereas the benefits (typically calculated in number of downloads or resulting publications) will most probably be less. The lack of open formats for the raw data adds to the difficulty of establishing funding for centralized repositories, which brings us to a catch-22: for a true incentive towards routine dissemination of raw data for published papers, we need open standards for the data formats used, but in order to push such open standards on the vendors, a large user community is needed that can actively define these standards as well as demand support for them from the vendors.

As a conclusion, the following recommendations can be made concerning the dissemination of MS data: (1) peak lists should be made available by default. There is no reason not to make these publicly available, and there are no real storage or distribution issues to be considered. (2) raw data have some clear benefits over peak lists, yet currently lack both standardized formats as well as the required infrastructure for centralized storage and distribution. Therefore, information on how to obtain raw data should at the very least be referenced in the published results for the time being. This can easily be done by providing links to individual lab websites from the journal websites (note that this is a version of the "truly distributed system" discussed above). (3) Efforts should be started at centralized repositories to create the necessary infrastructure so that in the mid- to long-term, source data will preferentially be submitted in the raw format. Meanwhile, (4) vendor support should be enlisted for open formats or at least open access to software tools that allow users to read and interpret the different formats of raw data. Since these latter developments are mutually dependent, the most important breakthrough to achieve seems to be the establishment of centralized repositories. Perhaps some lessons can be learned in this respect from the microarray community, as they have faced (and largely overcome) similar problems in the recent past [16].

The authors would like to thank David States for sharing his experiences from the data gathering efforts executed by his core bioinformatics unit at the University of Michigan, Ann Arbor; Ilan Beer for interesting discussions and his views on the subject matter; Peter Stoehr for the storage cost estimates; Jimmy Eng for contributing raw data file sizes for several instruments; and An Staes for her help in preparing the chart in Fig. 1. K.G. is a Postdoctoral Fellow and L.M. a Research Assistant of the Fund for Scientific Research-Flanders (Belgium) (F.W.O. Vlaanderen). Parts of the data used in this paper were generated in the context of the GBOU-research initiative (Project number 20204) of the Flanders Institute of Science and Technology (IWT).

References

[1] Hanash, S., Celis, J. E., *Mol. Cell. Proteomics* 2002, *1*, 413–414.

[2] Omenn, G. S., *Proteomics* 2004, *4*, 1235–1240.

[3] Adamski, M., Blackwell, T. W., Menon, R., Martens, L., Hermjakob, H., Taylor, C. F., Omenn, G., States, D., *Proteomics* 2005, *5*, this issue.

[4] Martens, L., Hermjakob, H., Jones, P., Adamski, M., Taylor, C. F., States, D., Gevaert, K., Vandekerckhove, J., Apweiler, R., *Proteomics* 2005, *5*, this issue.

[5] Carr, S. A., Aebersold, R., Baldwin, M., Burlingame, A., Clauser, K., Nesvizhskii, A., *Mol. Cell. Proteomics* 2004, *3*, 531–533.

[6] Prince, J. T., Carlson, M. W., Wang, R., Lu, P., Marcotte, E. M., *Nat. Biotechnol.* 2004, *4*, 471–472.

[7] Orchard, S., Hermjakob, H., Randall, K. J., Jr., Runte, K., Sherman, D., Wojcik, J., Zhu, W., Apweiler, R., *Proteomics* 2004, *4*, 490–491.

[8] Li, X. J., Zhang, H., Ranish, J. A., Aebersold, R., *Anal. Chem.* 2003, *75*, 6648–6657.

[9] Beer, I., Barnea, E., Ziv, T., Admon, A., *Proteomics* 2004, *4*, 950–960.

[10] Wiley, H. S., Michaels, G. S., *Nat. Biotechnol.* 2004, *22*, 1037–1038.

[11] Pedrioli, P. G., Eng, J. K., Hubley, R., Vogelzang, M., Deutsch, E. W., Raught, B., Pratt, B. *et al.*, *Nat. Biotechnol.* 2004, *22*, 1459–1466.

[12] Petritis, K., Kangas, L. J., Ferguson, P. L., Anderson, G. A., Pasa-Tolic, L., Lipton, M. S., Auberry, K. J. *et al.*, *Anal. Chem.* 2003, *75*, 1039–1048.

[13] Gygi, S. P., Rist, B., Gerber, S. A., Turecek, F., Gelb, M. H., Aebersold, R., *Nat. Biotechnol.* 1999, *17*, 994–999.

[14] Gevaert, K., Goethals, M., Martens, L., Van Damme, J., Staes, A., Thomas, G. R., Vanderkerckhove, J., *Nat. Biotechnol.* 2003, *21*, 566–569.

[15] Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E. *et al.*, *Nucleic Acids Res.* 2004, *32 Database issue*, D115–D119.

[16] Ball, C. A., Sherlock, G., Brazma, A., *Nat. Biotechnol.* 2004, *22*, 1179–1183.