1.1 Metabolomics Studies

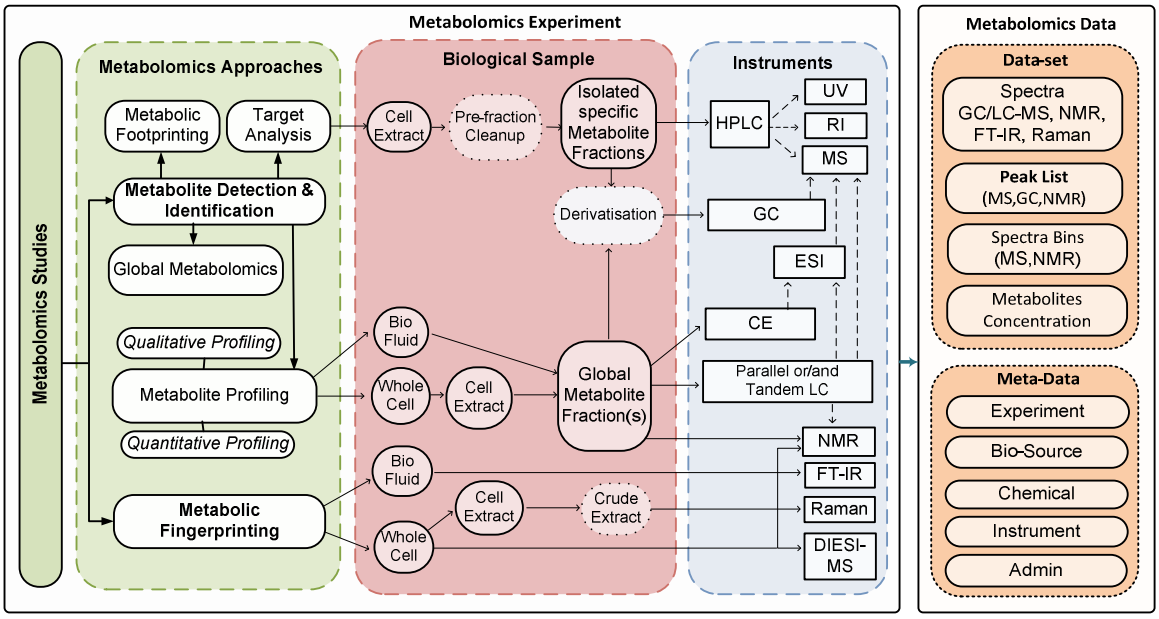

Figure 1

1.2 Metabolomics Data

Metabolomics data consists of both the data set as acquired by the instruments and its associated meta-data. The data set is acquired by *chemical analysis instruments* e.g. NMR, LC/GC-MS, HPLC, FT-IR, etc (McGregor, 1997, Brown and DeAntonis, 1997, Department of Chemistry - University of Arizona, 2006, Dettmer et al., 2007, Dunn and Ellis, 2005, Hites, 1997, Krishna et al., 2007, Sherman Hsu, 1997) in assays. The choice of the instrument depends on the goals of the investigation and their relation with the aims of the study and the design of the assay on one hand, and with the metabolic approaches on the other hand (Goodacre et al., 2004).

Assay data set is usually generated in a form of spectra which vary in their detailed structure depending on the data acquisition instrument and on the transformation used to convert the spectra from one format into another e.g. Fourier transformation for NMR, peak lists, spectra bins or concentration profile (Xia et al., 2009). Metabolomics metadata on the other hand, concern the recorded information in the study regarding the factors which might influence the data set e.g. bio-source, sample preparation, metabolic approach, data acquisition instruments, administration, chemical and other study related factors (Goodacre et al., 2007, Spasic et al., 2006, Sumner et al., 2007, Jenkins et al., 2005).

Factors related to the nature of metabolomics data including size, data types, data structures, and format must also be considered when setting the objectives. The selection of data mining techniques might be influenced by the number of attributes, number of examples (Goebel and Gruenwald, 1999), or the their ratio. Some of the techniques might be able to handle some data types while others might require conversion in later stages. The quality of data might be vital for the success and soundness of data mining results this including issues such as missing values (Michalski et al., 1998, Pelckmans et al.), outliers and unusual distributions of data (Xi, 2008). Several procedures might be required to improve the quality of the data and make it more suitable for modeling; those can be done either through data pre-processing or acclimatization.

1.2.1 References:

BROWN, P. & DEANTONIS, K. 1997. High-performance Liquid Chromotography. *In:* SETTLE, F. A. (ed.) *Handbook of instrumental techniques for analytical chemistry.* Upper Saddle River, NJ/ London: Prentice Hall PTR.

DEPARTMENT OF CHEMISTRY - UNIVERSITY OF ARIZONA. 2006. *Introduction to Mass Spectrometry* [Online]. Arizona: University of Arizona. Available: http://www.chem.arizona.edu/massspec/intro_html/intro.html [Accessed 22/11/2006 2006].

DETTMER, K., ARONOV, P. A. & HAMMOCK, B. D. 2007. Mass spectrometry-based metabolomics. *Mass Spectrometry Reviews,* 26, 51-78.

DUNN, W. B. & ELLIS, D. I. 2005. Metabolomics: Current analytical platforms and methodologies. *Trends in Analytical Chemistry,* 24, 285-294.

GOEBEL, M. & GRUENWALD, L. 1999. A survey of data mining and knowledge discovery software tools. *SIGKDD Explor. Newsl.,* 1, 20-33.

GOODACRE, R., BROADHURST, D., SMILDE, A., KRISTAL, B., BAKER, J., BEGER, R., BESSANT, C., CONNOR, S., CAPUANI, G., CRAIG, A., EBBELS, T., KELL, D., MANETTI, C., NEWTON, J., PATERNOSTRO, G., SOMORJAI, R., SJÖSTRÖM, M., TRYGG, J. & WULFERT, F. 2007. Proposed minimum reporting standards for data analysis in metabolomics. *Metabolomics,* 3, 231-241.

GOODACRE, R., VAIDYANATHAN, S., DUNN, W. B., HARRIGAN, G. G. & KELL, D. B. 2004. Metabolomics By Numbers: Acquiring Understanding Global Metabolite Data. *Trends in Biotechnology,* 22, 245-252.

HITES, R. A. 1997. Gas Chromotography Mass Spectrometry. *In:* SETTLE, F. A. (ed.) *Handbook of instrumental techniques for analytical chemistry.* Upper Saddle River, NJ/London: Prentice Hall PTR.

JENKINS, H., JOHNSON, H., KULAR, B., WANG, T. & HARDY, N. 2005. Toward supportive data collection tools for plant metabolomics. *Plant Physiology,* 138, 67-77.

KRISHNA, C., SOCKALINGUM, G., BHAT, R., VENTEO, L., KUSHTAGI, P., PLUOT, M. & MANFAIT, M. 2007. FTIR and Raman microspectroscopy of normal, benign, and malignant formalin-fixed ovarian tissues. ***Analytical and Bioanalytical Chemistry,*** **387, 1649-1656.**

MCGREGOR, M. 1997. Nuclear Magnetic Resonance Spectroscopy ***In:*** **SETTLE, F. A. (ed.)** ***Handbook of instrumental techniques for analytical chemistry.*** **Upper Saddle River, NJ / London: Prentice Hall PTR.**

MICHALSKI, R. S., BRATKO, I. & KUBAT, M. 1998. ***Machine Learning and Data Mining: Methods and Applications,*** **Chichester, John Wiley & Sons.**

PELCKMANS, K., DE BRABANTER, J., SUYKENS, J. A. K. & DE MOOR, B. 2005. Handling missing values in support vector machine classifiers. ***Neural Networks,*** **18, 684-692.**

SHERMAN HSU, C. P. 1997. Infrared Spectroscopy ***In:*** **SETTLE, F. A. (ed.)** ***Handbook of instrumental techniques for analytical chemistry.*** **Upper Saddle River, NJ / London: Prentice Hall PTR.**

SPASIC, I., DUNN, W., VELARDE, G., TSENG, A., JENKINS, H., HARDY, N., OLIVER, S. & KELL, D. 2006. MeMo: a hybrid SQL/XML approach to metabolomic data management for functional genomics. ***BMC Bioinformatics,*** **7, 281.**

SUMNER, L. W., AMBERG, A., BARRETT, D., BEALE, M. H., BEGER, R., DAYKIN, C. A., FAN, T. W. M., FIEHN, O., GOODACRE, R., GRIFFIN, J. L., HANKEMEIER, T., HARDY, N., HARNLY, J., HIGASHI, R., KOPKA, J., LANE, A. N., LINDON, J. C., MARRIOTT, P., NICHOLLS, A. W., REILY, M. D., THADEN, J. J. & VIANT, M. R. 2007. Proposed minimum reporting standards for chemical analysis. ***Metabolomics,*** **3, 211-221.**

XI, J. 2008. Outlier Detection Algorithms in Data Mining. ***Proceedings of the 2008 Second International Symposium on Intelligent Information Technology Application - Volume 01.*** **IEEE Computer Society.**

XIA, J., PSYCHOGIOS, N., YOUNG, N. & WISHART, D. S. 2009. MetaboAnalyst: a web server for metabolomic data analysis and interpretation. ***Nucleic Acids Research,*** **37, W652-660.**