Grid-based Analysis of Tandem Mass Spectrometry Data in Clinical Proteomics

Andreas Quandt [a,b,1], Patricia Hernandez [a,2], Peter Kunzst [b], Cesare Pautasso [d], Marc Tuloup [a], Celine Hernandez [a], and Ron D. Appel [a,b]

[a] *Swiss Institute of Bioinformatics, Geneva, Switzerland*
[b] *University of Geneva, Geneva, Switzerland*
[c] *Swiss National Supercomputing Centre, Manno, Switzerland*
[d] *IBM Research GmbH, Zurich Research Laboratory, Switzerland*

Abstract. Biomarker detection is one of the greatest challenges in Clinical Proteomics. Today, great hopes are placed into tandem mass spectrometry (MS/MS) to discover potential biomarkers. MS/MS is a technique that allows large scale data analysis, including the identification, characterization, and quantification of molecules. Especially the identification process, that implies to compare experimental spectra with theoretical amino acid sequences stored in specialized databases, has been subject for extensive research in bioinformatics since many years. Dozens of identification programs have been developed addressing different aspects of the identification process but in general, clinicians are only using a single tools for their data analysis along with a single set of specific parameters. Hence, a significant proportion of the experimental spectra do not lead to a confident identification score due to inappropriate parameters or scoring schemes of the applied analysis software. The swissPIT (Swiss Protein Identification Toolbox) project was initiated to provide the scientific community with an expandable multi-tool platform for automated and in-depth analysis of mass spectrometry data. The swissPIT uses multiple identification tools to automatic analyze mass spectra. The tools are concatenated as analysis workflows. In order to realize these calculation-intensive workflows we are using the Swiss Bio Grid infrastructure. A first version of the web-based front-end is available (http://www.swisspit.cscs.ch) and can be freely accessed after requesting an account. The source code of the project will be also made available in near future.

Keywords. cyclic workflows, proteomics, tandem mass spectrometry, workflow manager

1. Introduction

Proteomics can be defined as the systematic study of the protein content (called the proteome) of a given cell, tissue or organism, at a given time and under specific conditions [18]. Proteomic research encompasses the identification, characterization and quantification of proteins. It involves various techniques, including single and tandem mass spectrometry.

[1] To whom correspondence should be addressed: andreas.quandt@isb-sib.ch
[2] Contributed to the paper equally as the first author

Studying proteins is a particularly challenging task. Unlike genes, proteins are neither homogenous nor static. Genes are confined in the cell nucleus, while proteins are present in all cell compartments, on the cell surface and even in extra-cellular fluids. Gene patterns are fairly constant while protein patterns mirror even small changes in their environmental changes including pathological and physiological developments. Great hopes are placed in protein separation techniques associated with mass spectrometry in order to discover potential diagnostic markers and therapeutic targets. They allow the rapid detection of proteomic biomarkers by comparing case and control samples such as cancer and non-cancer samples. Examples for biomarkers at the genomic level are single nucleotide polymorphisms (SNP) [10] or DNA methylation [4], and at the proteomic level, protein abundance or protein post-translational modifications. Proteomic biomarkers additionally provide the possibility to monitor the evolution of diseases or drug treatment. This is especially important for the correct treatment of cancer where over-treatment in earlier stages of the disease is a well-known phenomenon. For example, breast cancer patients often display significantly different clinical phenotypes and responses to a specific therapy [12]. Medical doctors can use the additional information provided by the analysis of proteomic biomarkers to identify the best method for treatment and the correct doses to be applied. For example, the treatment with hormonal therapy [16] has fewer side effects than restrain treatments with more noxious effects like radiation [1].

In clinical proteomics, biological samples are analyzed in high-throughput mode, producing thousands of data files each day. The analysis, storage, and distribution of this huge data volume requires the development of new bioinformatics tools capable to use distributed computing resources such as a Grid infrastructure. Grid Computing has been postulated as a modern paradigm to gain access to a large number of computational resources. It enables individual resource owners to share their infrastructure with each other, providing a large overall infrastructure for many different projects. The biggest benefit of a shared infrastructure comes from the optimized usage of resources. Individually, each project would most probably under-utilize dedicated resources or would not get access to enough resources. The Grid takes care of distributing the load on many resources, maximizing the utilization of the resources.

This paper describes the Swiss Protein Identification Toolbox (swissPIT) a new proteomics platform managing several data analysis tools by combining them in user-defined workflows (Figure 1). In order to be able to execute complex analysis workflows, swissPIT uses the infrastructure of the Swiss Bio Grid project. The Swiss Bio Grid (http://www.swissbiogrid.org/) provides a large-enough Grid infrastructure that can be used to perform identification workflows and parameter studies with swissPIT. Currently, five partner institutions are providing resources: the Swiss Institute of Bioinformatics (SIB), the Swiss National Supercomputing Centre (CSCS), the Biozentrum at Basel University, the Friedrich Miescher Institute in Basel and the Functional Genomics Centre Zurich. The infrastructure makes use of the NorduGrid Advanced Resource Connector ARC middleware (http://www.nordugrid.org/) to manage the submission of individual Grid jobs for swissPIT (i.e. the execution of the identification tools with a given parameter set). Two main advantages of grid computing for proteomic experiments are job distribution (e.g. when a great number of users want, at the same time, to analyze their datasets), and job parallelization (e.g. in case of exploration of parameter combinations).

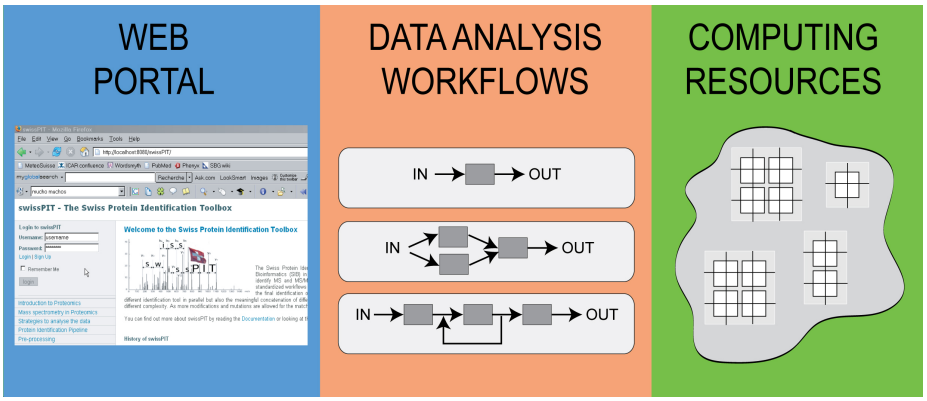

Figure 1. The three aspects of the swissPIT platform: a) the web interface that allows the user to login, choose the programs or workflow to run, upload the data, set the parameters and visualize the results; b) the data analysis workflows; c) the grid environment.

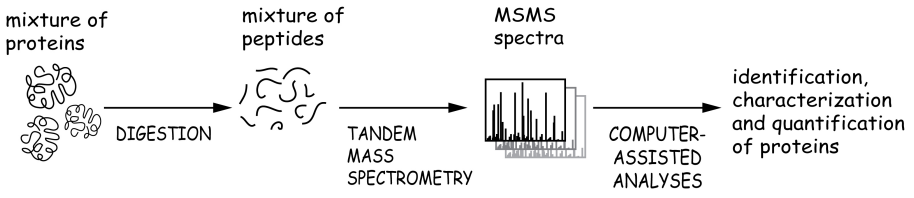

Figure 2. Whole proteins are purified and cleaved into peptides. During the mass spectrometric analysis, peptides are selected based on their m/z values, fragmented into pieces, and the m/z of the obtained fragments are reported in an MS/MS spectrum. Interpretation of the obtained mass spectra is done using specialized software and protein or genomic sequence databases.

1.1. Background

In clinical proteomics, both single and tandem mass spectrometry techniques are used to screen biological samples in high-throughput mode to identify and quantify interesting biomarkers. Unlike single mass spectrometry, tandem mass spectrometry provides information on the protein sequence and is additionally capable to identify proteins in mixtures. Tandem mass spectrometry (MS/MS) requires two mass analyzers in series. The first analyzer separates the peptides according to their mass-to-charge ratio (m/z). Then, selected peptides (called precursors) undergo fragmentation and the m/z ratios of the produced fragments are measured by the second analyzer. This yields MS/MS spectra composed of a precursor peptide mass and of fragment peaks (m/z and intensity). The number of peaks varies from about ten to several hundreds depending on factors like the precursor peptide length, the fragmentation quality, the mass spectrometer type and the peak detection process. The obtained mass spectra can then be used to carry out various analyses. In proteomics studies, the three major topics of interest are the identification, characterization and quantification of the proteins present in a sample. Figure 2 summarizes key steps of a typical MS/MS-based experiment.

To identify peptide patterns within the spectra, bioinformatics tools are applied to screen genomic and protein databases (such as the UniProtKB/swissProt and UniPro-

tKB/TrEMBL [19]) for high scoring matches. Databases are usually composed of tens of thousands to several millions of biological sequences. The most common approach to identify peptides is called Peptide Fragment Fingerprinting (PFF). This approach computes a similarity score between the experimental MS/MS spectrum and theoretical MS/MS spectra constructed from the theoretical digestion of protein sequences to peptides and *in silico* fragmentation of the peptides. The calculation of these similarity scores is often difficult because of the variation between observed peptides and their corresponding database entries. For example, the peptide may carry a post-translational modification that is not documented in the database. The sequence can also be mutated, or sequence rearrangements may have occurred during the digestion process (transpeptidation). All these modifications cause peak shifts in the experimental spectra which then differ from the theoretical spectra in the database. On the other hand, the database may also contain errors, i.e. when the protein sequences are automatically translated from genomic or transcriptomic sequences (e.g. EST contigs). Traditionally, identification tools, such as Sequest [9], Mascot [14] and Phenyx [2], require the user to specify in advance a list of possible modifications taken into account during the matching process. These tools perform what we call a "classical search". Their main advantage is the production of results in a reasonable amount of time, but they cannot identify peptides which carry unexpected modifications or mutations. “Open-modification search” tools, such as Popitam [7], OpenSea [15], GutenTag [6] and InsPecT [17], have been specifically designed to handle unexpected amino acid modifications, but they often need preliminary filtering steps. *De novo* sequencing is another approach, which infers sequence information from the experimental MS/MS spectrum. The identification is then performed by matching the obtained *de novo* sequence with the database peptides. *De novo* sequencing methods require spectra of higher quality with smaller fragment errors and a more or less continuous signal, or at least high-quality signal for several adjacent amino acids. Despite these disadvantages, *de novo* methods may surpass PFF methods, notably when searching genomic databases subjected to sequencing errors, when searching databases composed of homologous sequences or when analyzing a spectrum that originates from a mutated protein or variant [11].

1.2. MS/MS identifications workflows

For many years the combination of different search tools and analysis strategies has been proposed to improve the result quality of the identification process [8]. Hence, the development of a platform for MS/MS identification workflows is one of the most interesting topics in bioinformatics for proteomics. The swissPIT project aims at providing a protein identification platform from MS/MS data. It is being developed on a Grid infrastructure in order to perform complex identification workflows. It is anticipated that the meaningful combination of several analysis tools will allow on the one hand, to identify more interesting biomarkers, and on the other hand to increase the overall reliability of the results. Advantages of such a platform include (1) reduced human intervention by automated execution of software workflows (2) simple data and results sharing between different user groups by user-controlled data storage, (3) and the feasibility of large studies on parameter optimization as well as more complex identification processes by the use of distributed computing resources.

2. Methods

The heart of the swissPIT project is to implement meaningful identification workflows such as to improve the identification quality for given data sets. In order to develop such an individual identification strategy, the same input data has to be analyzed with different parameter settings using the various identification tools. Depending on the number of adjustable parameters, these parameter sweeps may be extensive, i.e. there may be several hundreds or thousands of different parameter sets for which the tools have to be executed and results have to be evaluated. Consequently, we expect thousands of individual executions for each of the identification tools, in order to discover the most adapted parameter combination for a specific data set. So as to be able to perform these calculations within a reasonable amount of time, a high-throughput computing infrastructure has to be used.

2.1. Typical scenario

Due to the advantages and disadvantages of the different search strategies, a typical scenario for an identification process is to apply a classical search followed by an open-modification search. In a classical search, an identification algorithm generally produces a list of accession numbers (ACs) corresponding to peptide sequences matching the identified spectra. In an open-modification search, these ACs can then be used to improve the identification by searching for unexpected modifications. The unexpected modifications can then be used again to perform a new classical search with an updated list of potential modifications.

2.2. Use case workflow

The use case describes an identification process with three searches performed sequentially, where the output of each search is the input of the following search. In swissPIT, the workflow of this use case contains three modules: two classical search modules and one open-modification search module. Modules in swissPIT are defined as workflow parts which have to be executed in a predefined order. For each of the three modules multiple algorithms exist that are executed in parallel to reduce to overall analysis time. To test the first workflows, we decided to implement interfaces to two classical search approaches and two open-modification search tools. Phenyx and X!Tandem [3] are well-known tools for classical search. Phenyx (http://phenyx.vital-it.ch/) is currently the most advanced tool available while X!Tandem being an open-source tool is often used as a reference implementation. Popitam and InsPecT are two more recent tools specialized for detecting unexpected modifications.

Using all four tools, the described use case example can be implemented as follows (see Figure 3): Phenyx and X!Tandem can be started in parallel to perform a first classical search on a specific data set. Both programs will identify a subset of all spectra and produce an AC list. This list is then used by Popitam in the following open-modification search, with Inspect running in parallel. The produced list of possible modifications (modres.xml) can then be used again by a repeated run of Phenyx and X!Tandem in order to identify spectra of the original data set that have not been identified during the first run. Besides the computational needs, a major problem is the handling of the different input and output files. The lack of standards for input and output formats, parameter de-

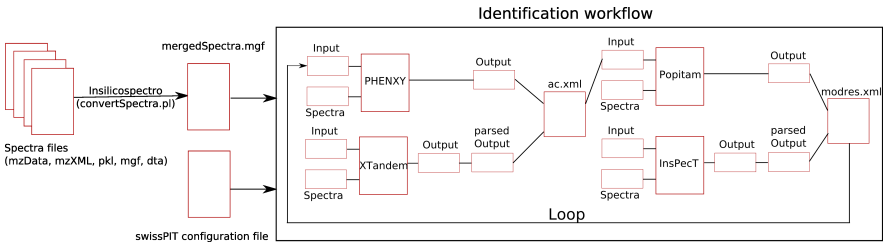

Figure 3. Use case workflow

finitions and presentation of the information content of the results requires many parsing steps in order to use standardized parameters and to merge results for the continued use within the workflow.

2.3. Workflow configuration

We implemented a basic workflow manager that can be configured by a single XML file. Figure 4 shows an example configuration file. It contains three parts necessary to describe a complete workflow for the swissPIT engine: the workflow (A), the modules (B), and the programs (C). In general, a workflow (A) contains different modules (B) with one or more programs (C). Each module and each program has a unique identifier (ID) which is used to execute the workflow with the correct parameter set. Unique ID's are necessary because a workflow can contain modules and programs of the same type but with different parameter sets.

In section A of the configuration file, the modules of the workflow are specified by using the <moduleid> tag. The module ID is unique and points to a specific module in section B of the configuration file. The order of the module ID's is important because the modules are performed in the same sequence as their ID's are specified. If no module matches the id specified in the workflow, swissPIT stops with an error message. The example configuration shown in Figure 4 defines a workflow with two modules (IDs "m01" and "m02"). The IDs are pointing to two modules (B) which are of the type "strictSearch" (m01) and "openModificationSearch" (m02). In addition to the module definition, section B also contains a parameter section. That section specifies parameters which are shared between different module types of a workflow. A module is defined by the parameters specific to this module (shared parameters between the programs of the module) and the programs executed within this module. The program IDs named within the <module> tag are pointing to a specific program definition (C). Depending on the program type, the program definition contains parameters which are not shared with other programs of the workflow. During the execution of the workflow, the various parameters are used to create the individual input files for each program.

2.4. Information flow within a workflow

Figure 3 illustrates the second major problem in realizing identification workflows. In addition to the intensive computational resources needed to execute these workflows, many parsing and converting steps are required to transport information between the tools. The lack of standardized formats has been recognized within the

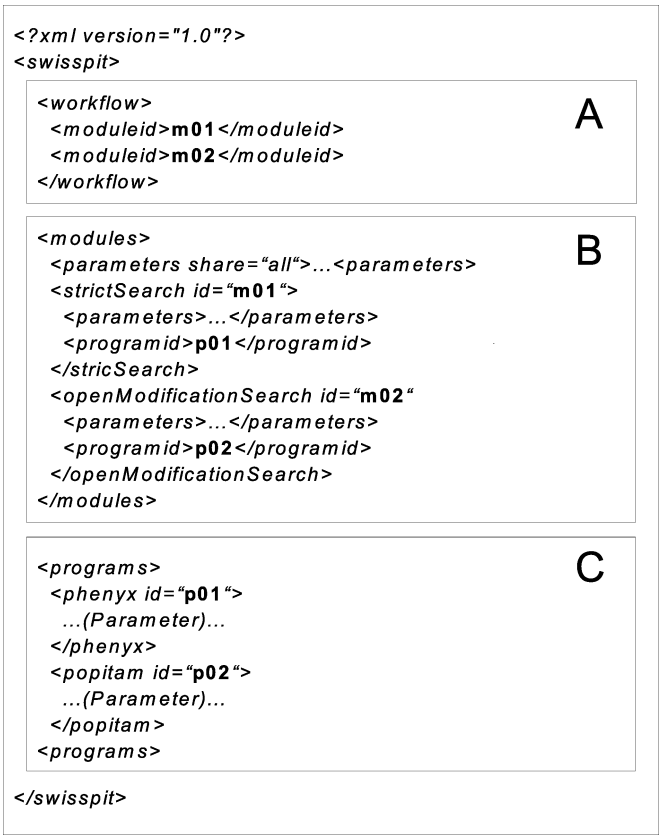

Figure 4. Example scheme of the swissPIT configuration file

scientific community. The Proteomics Standards Initiative (PSI) has been founded in 2002 (http://psidev.sourceforge.net/) to develop, among others, standard formats for raw data/peak lists (mzData) and for analysis results (analysisXML). At the time this article was written, analysisXML was still in development. Therefore, it was necessary to create internal formats for the information transport within the presented workflow. First, we developed the initial configuration file containing parameters which are shared by all programs, shared between programs of a specific search type (strict search (or classical search) and open modification search), and program specific parameters (see Figure 4). These parameters are the common basis to create the individual parameter files for each program. In a second step, we developed parsers to have a common output format for programs of the same module (search strategy). The use of a common output format allows not only the simplification of the information transport between the modules (ac.xml, modres.xml), it is also the first step of the development of a visualization tool displaying the results of different applications in a standardized way, in order to enhance result comparison.

3. Results

In this article we have described swissPIT, the Swiss Protein Identification Toolbox, a new workflow-based platform for the analysis of mass spectrometry data. By giving access to several analysis tools, swissPIT aims to mix different search strategies to increase the overall identification of spectra. We also argue for the need of high-throughput computing infrastructures like Grids in modern mass spectrometry-based proteomics to perform the in-depth analysis with our identification workflows.

The swissPIT platform has been installed on the CSCS node of the Swiss Bio Grid. Access to this installation is given by request to the authors. This command line version of swissPIT provides the execution of basic identification workflows with parallel program execution and sequential combination of different search strategies. The user can create new workflows by easily modifying the configuration file of swissPIT. He/she also can perform a parameter exploration by specifying different parameter sets which are executed in parallel. A web-based interface allows for the easy control of swissPIT (http://swisspit.cscs.ch). The web interface does not provide the execution of workflows yet, but it is already a useful tool for biologist to use several identification programs with a unified interface. It also allows an improved result comparison for the different tools by providing standardized parameter sets across the programs. The available programs are executed in parallel on the Grid infrastructure for faster analysis of the data. Submissions are handled with a swissPIT server certificate and the user authentication is managed by the login procedure. After the successful submission of a Grid job, the user can monitor its progress through the job monitor. Once all runs are finished, the user can browse the job folder within his/her personal user space to retrieve the individual result files. The results are presented to in their original view and can also be downloaded locally.

Grid Interaction Details

The Swiss Bio Grid aims to create a showcase of a basic infrastructure capable of serving a variety of bioinformatics calculations. There are three projects currently supported on this platform, one of which is the swissPIT proteomics project described in this paper. This is a showcase infrastructure, with the aim of producing real results, initiating the establishment of a more sustainable Swiss national Grid. It is based on two complementary technologies: the NorduGrid ARC middleware (which itself is based on Globus) and the United Devices GridMP desktop Grid. The proteomics project achieves Grid parallelization by parametrization, i.e. the execution of the same programs with different input parameters. Each program is submitted by the swissPIT end-user interface through the NorduGrid ARC submission interface. The GridMP services are not used for this project, as the executables are not all available on the Windows platform. In the current setup, the swissPIT interface acts as a portal to the users, i.e. the actual jobs are all submitted as a single 'swisspit' user to the Grid. The ARC monitoring clients are used to track the progress of each job, and when successful execution has been reported, again the ARC client tools are used to retrieve the output of the job.

4. Outlook

The swissPIT addresses complex problems that have not been tackled on this scale before. The most challenging aspect is to provide high-throughput computing resources to

the user by hiding the complexities of the system behind an intuitive interface. In the future, we have to automate processes such as individual user authentication and authorization. We also have to continue working on solutions for distributed data storage across all nodes of the Grid infrastructure, and to implement mechanisms for data access to allow data sharing between user groups. A further aspect will be the link between these data and shared databases such as Peptide Atlas [5] or Pride [13] which allow the application of statistics on a larger scale. We also have to work on an automatic distribution of the several biological databanks such as Uniprot/Swissprot or Uniprot/TrEMBL in their different versions within the Grid infrastructure, so that results of the distributed parameter sweeps are consistently making use of the same datasets.

In the short term, we are mainly working on improvements of the web interface, as well as on a common visualization interface for the results retrieved from the different search tools. A tool for graphically building and executing workflows is planned, as well as support for an enlarged set of identification algorithms. We also plan to use analysisXML as the internal file format for the information transport to unify the information flow and to simplify the connection to other programs interacting with swissPIT.

5. Acknowledgements

We would like to thank Sergio Maffioletti (CSCS) and Alexandre Masselot (Genebio S.A.) for intensive help with the Swiss Bio Grid infrastructure and the file format parsing.

References

[1] B. S. Bloom, N. de Pouvourville, S. Chhatre, R. Jayadevappa, and D. Weinberg. Breast cancer treatment in clinical practice compared to best evidence and practice guidelines. *Br J Cancer*, 90(1):26–30, Jan 2004.

[2] Jacques Colinge, Alexandre Masselot, Marc Giron, Thierry Dessingy, and Jérôme Magnin. Olav: towards high-throughput tandem mass spectrometry data identification. *Proteomics*, 3(8):1454–1463, Aug 2003.

[3] Robertson Craig and Ronald C Beavis. Tandem: matching proteins with tandem mass spectra. *Bioinformatics*, 20(9):1466–1467, Jun 2004.

[4] Partha M Das and Rakesh Singal. Dna methylation and cancer. *J Clin Oncol*, 22(22):4632–4642, Nov 2004.

[5] Frank Desiere, Eric W Deutsch, Nichole L King, Alexey I Nesvizhskii, Parag Mallick, Jimmy Eng, Sharon Chen, James Eddes, Sandra N Loevenich, and Ruedi Aebersold. The peptideatlas project. *Nucleic Acids Res*, 34(Database issue):D655–D658, Jan 2006.

[6] Ari Frank, Stephen Tanner, Vineet Bafna, and Pavel Pevzner. Peptide sequence tags for fast database search in mass-spectrometry. *J Proteome Res*, 4(4):1287–1295, 2005.

[7] Patricia Hernandez, Robin Gras, Julien Frey, and Ron D Appel. Popitam: towards new heuristic strategies to improve protein identification from tandem mass spectrometry data. *Proteomics*, 3(6):870–878, Jun 2003.

[8] Patricia Hernandez, Markus Müller, and Ron D Appel. Automated protein identification by tandem mass spectrometry: Issues and strategies. *Mass Spectrom Rev*, Nov 2005.

[9] Ashley L. McCormack Jimmy K. Eng and III John R. Yates. 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*, Volume 5, Issue 11:Pages 976–989, November 1994.

[10] Bao li Chang, Siqun L Zheng, Sarah D Isaacs, Aubrey Turner, Gregory A Hawkins, Kathy E Wiley, Eugene R Bleecker, Patrick C Walsh, Deborah A Meyers, William B Isaacs, and Jianfeng Xu. Polymorphisms in the cyp1a1 gene are associated with prostate cancer risk. *Int J Cancer*, 106(3):375–378, Sep 2003.

[11] Adam J Liska and Andrej Shevchenko. Expanding the organismal scope of proteomics: cross-species protein identification by mass spectrometry and its implications. *Proteomics*, 3(1):19–28, Jan 2003.

[12] Edison T Liu. Classification of cancers by expression profiling. *Curr Opin Genet Dev*, 13(1):97–103, Feb 2003.

[13] Lennart Martens, Alexey I Nesvizhskii, Henning Hermjakob, Marcin Adamski, Gilbert S Omenn, Joël Vandekerckhove, and Kris Gevaert. Do we want our data raw? Including binary mass spectrometry data in public proteomics data repositories. *Proteomics*, 5(13):3501–3505, Aug 2005.

[14] D. N. Perkins, D. J. Pappin, D. M. Creasy, and J. S. Cottrell. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, 20(18):3551–3567, Dec 1999.

[15] Brian C Searle, Surendra Dasari, Mark Turner, Ashok P Reddy, Dongseok Choi, Phillip A Wilmarth, Ashley L McCormack, Larry L David, and Srinivasa R Nagalla. High-throughput identification of proteins and unanticipated sequence modifications using a mass-based alignment algorithm for ms/ms de novo sequencing results. *Anal Chem*, 76(8):2220–2230, Apr 2004.

[16] Richard I Somiari, Stella Somiari, Stephen Russell, and Craig D Shriver. Proteomics of breast carcinoma. *J Chromatogr B Analyt Technol Biomed Life Sci*, 815(1-2):215–225, Feb 2005.

[17] Stephen Tanner, Hongjun Shu, Ari Frank, Ling-Chi Wang, Ebrahim Zandi, Marc Mumby, Pavel A Pevzner, and Vineet Bafna. Inspect: identification of posttranslationally modified peptides from tandem mass spectra. *Anal Chem*, 77(14):4626–4639, Jul 2005.

[18] M. R. Wilkins, C. Pasquali, R. D. Appel, K. Ou, O. Golaz, J. C. Sanchez, J. X. Yan, A. A. Gooley, G. Hughes, I. Humphery-Smith, K. L. Williams, and D. F. Hochstrasser. From proteins to proteomes: large scale protein identification by two-dimensional electrophoresis and amino acid analysis. *Biotechnology (N Y)*, 14(1):61–65, Jan 1996.

[19] Cathy H Wu, Rolf Apweiler, Amos Bairoch, Darren A Natale, Winona C Barker, Brigitte Boeckmann, Serenella Ferro, Elisabeth Gasteiger, Hongzhan Huang, Rodrigo Lopez, Michele Magrane, Maria J Martin, Raja Mazumder, Claire O'Donovan, Nicole Redaschi, and Baris Suzek. The universal protein resource (uniprot): an expanding universe of protein information. *Nucleic Acids Res*, 34(Database issue):D187–D191, Jan 2006.