

SWISS-TANDEM: A Web-based Workspace for MS/MS Protein Identification on PC Grids

Dominique Zosso, Konstantin Arnold, Torsten Schwede and Michael Podvinec
Swiss Institute of Bioinformatics and Biozentrum University of Basel, Switzerland
dominique.zosso@epfl.ch, {konstantin.arnold, torsten.schwede, michael.podvinec}@unibas.ch

Abstract

Tandem mass spectrometry is a powerful technique for the identification of proteins contained in a biological sample. We present a grid-based application service for the analysis of tandem mass spectra, based on the open-source X!Tandem algorithm. The implemented parallelization strategy allows the distribution of analysis jobs to a grid of desktop PCs, utilizing their idle capacity. For this application service, a user-friendly project-oriented front-end was developed, implemented as a web-based workspace. After analysis, results are visualized and presented to the user along with a set of powerful and novel tools to allow further interpretation and evaluation of search results.

1. Introduction

Proteomics is a cornerstone of the emerging field of systems biology, as it allows the systematic study of sets of proteins found in particular cell types and states or in response to particular endogenous or exogenous stimuli. In this context, protein mass spectrometry is used as a sensitive detection tool [1]. Recent years have seen the successful application of proteomics to a wide variety of questions of clinical and biological research interest, and its value in the development of new drugs and the discovery of biomarkers has become widely recognized [2-4]. This surge was made possible in part through advances in instrumentation and automatization which allow the design and analysis of large-scale experiments. The large sets of data created in this fashion bring forward the need for sophisticated data management, analysis and visualization tools and workflows. At the same time, the analysis of protein MS/MS data is computationally intensive: At its core are peptide identification algorithms that match observed peptide fragment masses against mass fingerprints calculated from theoretical digests of protein databases, requiring considerable compute power. Moreover, refined search strategies that allow the specific discovery or detection of protein modifications and mutations are gaining relevance [5]. These features quickly lead to combinatorial explosion of the search space, and therefore cause massive increase of needed compute power. To satisfy the high computational demands of large data sets, high-performance compute clusters dedicated to the analysis of MS/MS results are commonplace nowadays. While more compute power is certainly not the only way of improving protein identification quality, it allows the researcher to consider a larger number of protein modifications that would otherwise go undetected.

In essence, matching peptide fragment spectra and theoretical spectra derived from a protein database is a data-parallel problem that lends itself well to distribution to many loosely connected computers, such as a grid of regular desktop PCs providing idle capacity. In a first approximation, sets of spectra, as well as protein databases, can be split into smaller fractions to be analyzed on individual computers without information exchange during calculations. Interdependency of the results is mainly introduced by the statistical evaluation of peptide matches (expectation values), as well as by the inference of proteins from cumulated peptide

evidence. The latter operation is generally not time-limiting and need not be parallelized. We have previously investigated the effect of splitting sets of spectra and/or protein databases on peptide and protein expectation values, finding no significant difference in results to analyses carried out on a single processor [6]. In the same work, the gain in performance from different parallelization strategies was assessed, suggesting that basic principles of economy of scales and data distribution overheads must be respected.

In this paper, we briefly recapitulate how we have parallelized and grid-enabled the open-source peptide identification algorithm X!Tandem [6, 7]. An application service was developed that allows the parallel execution of peptide searches on a university desktop PC grid consisting of hundreds of PCs, providing the computational back-end of SWISS-TANDEM. Here, we focus on the integration of project-based job submission and results retrieval and visualization into a user-friendly workspace. Users interact with the application service through a web-based environment adapted from our SWISS-MODEL Workspace for protein homology modeling [8]. Finally, we report on a number of visualization and reporting tools developed within this framework, which provide deeper insight into protein identification results not only for mass spectrometry experts, but also to biologists, who very often are domain experts, but only superficially involved in the analysis process.

2. Materials and methods

2.1. X!Tandem parallelization and grid-enabling

In the present study, the open-source X!Tandem software was used to match tandem MS spectra with a protein database [7]. Main reasons for its selection were its free availability and demonstrated performance. The principles, however, of our method should apply analogously to other commercial and free search programs.

X!Tandem expands a protein database into a list of peptides, following an expansion model of user-defined complexity. In a two-step approach, the software first selects candidate proteins from a large database, and subsequently performs a refined search against these, taking into account potential amino acid mass modifications, as well as partial cleavages and point mutations. Peptide and protein expectation values account for spectra and database complexity.

The parallelization scheme employed here includes splitting of both tandem spectra and protein database, and reproduces the X!Tandem native two-step approach [6]. First, both spectra and proteins are each re-bundled into (m,n) subdivisions, and a set of work-units is formed by the cross-product of these subdivisions. The $m \times n$ sub-jobs run separately on the PC grid nodes and their non-refined candidate protein identifications are collected. At refinement, all spectra subdivisions are matched against this protein candidate database. Results are merged into a single output file and consolidation is run to calculate the final protein e-values. We have previously shown that subdividing the set of tandem spectra has no tangible impact on the e-value outcome. Moreover, the e-value shift after the non-refined step, caused by smaller input peptide lists, can be anticipated by lowering the e-value threshold accordingly. While the consolidation step, necessary for complete protein e-values, introduces some e-value fluctuations, this has no significant impact on the overall performance of peptide and protein identifications [6].

2.2. Web-based front-end, application service, and the PC grid

To facilitate the handling of a protein identification job, a web-based front-end has been derived from the SWISS-MODEL workspace framework [8]. This PHP/Perl-based framework

offers user-based project and data management. Registered users have access to their previous projects and data, and can submit new jobs to specific tools. The user interacts with the workspace through a dynamic PHP-driven front-end running on an Apache web server. Data and computations are provided by a Perl server, communicating through a XML-RPC interface with the front-end. Calculations can be run online, or are submitted as batch jobs to a local resource, while the web front-end lets users monitor job progress and view completed results.

Protein identification jobs are registered with a central application service daemon, which implements the abovementioned parallelization scheme. The service takes care of all necessary data pre-processing, including conversion of tandem MS input files into dta-peaklists as a prerequisite before the data sets are split. Work-unit deployment, monitoring, and result retrieval is managed through a local resource management system (LRMS) abstraction layer.

This layer, derived from ProtoGRID [9], allows to address underlying compute resources and management systems through a simple and unified API for job and data handling. The application service daemon saves and provides state information for all protein identification jobs, allowing recovery of stalled or crashed jobs.

2.3. Visualization and reporting tools

As the X!Tandem native XML results output file is not suitable for direct human interpretation, a number of common, as well as some novel report-types and charts are available in the workspace to assist interpretation and evaluation of protein identification results. Information contained in the results file and, where possible, the original MS dataset is structured into 4 bottom-up hierarchical layers: 1) peptide inspection, 2) protein level, 3) sample overview, and 4) an optional cross-sample level, requiring an additional summary creation step. The main features of these views are pointed out in the following paragraphs.

The peptide inspection layer describes the details of a spectrum-peptide assignment, providing the essential information to qualify the peptide identification. This includes available data about the peptide mass and charge state, the corresponding MS values, the theoretical peptide product ion-table, as well as a superposition of measured and simulated peptide mass fingerprints.

Next, the protein layer contains all information available for one specific protein identification, i.e. its cumulated peptide evidence and the resulting protein e-value. Coverage graphs provide details on which portions of the sequence are covered by identified peptides (color-coded by peptide e-value) and highlight detected mass modifications by vertical bars. Based on the protein description in the FASTA database, pointers to external data sources are provided. In particular, an interface to the SWISS-MODEL Repository [10] allows to project peptide coverage and mass modification data onto a 3D structure model of the specific protein.

The sample overview layer collects information about all inferred proteins of a given protein identification project. As often several proteins in a database share portions of their sequence, both within and across species, a single identified peptide may evoke multiple proteins. A novel evidence chart has been introduced to disambiguate such multiple protein inferences: Proteins sharing at least one identified peptide are clustered together using graph exploration.

Finally, several analyses may be combined in a summary. This project type provides visualization of protein presence and coverage across different protein identification runs (e.g. fractions of a sample), and offers a consolidated list of proteins identified in a measurement series. In this way, the shortcomings incurred by the analytical incompleteness of single runs can be overcome by combining the results of multiple biological and/or analytical replicates of an experiment, as suggested in [11].

3. Results and discussion

In this section, we describe how the SWISS-TANDEM workspace has been used to analyze the protein content of a known 17-proteins mixture, of which LC-MS and MS/MS data are publicly available as a mzXML file [12].

3.1. Project submission

The 25.8 MB mzXML file has been uploaded through the web-interface. Visualization of the LC-MS dataset allows to reasonably restrict the considered scan time and m/z range of MS/MS data. UniProtKB/Swiss-Prot release of November 17th 2006 has been selected out of the list of preloaded reference protein databases, comprising several recent releases of the IPI collections [13] and the UniProtKB/Swiss-Prot and UniProtKB/trEMBL databases [14]. Alternatively, a user-specified protein database in FASTA format could have been uploaded. X!Tandem analysis parameters are configured for a tryptic peptide expansion model with a maximum of 3 missed cleavage sites. Cysteine residue mass was modified by +57 Da considering protein sample preparation. Methionine oxidation (+16 Da), asparagine/glutamine deamidation (+1 Da), and serine/threonine phosphorylation (+80 Da) in refinement, were added as frequently observed potential modifications.

3.2. Results retrieval and visualization

Out of 2'093 considered tandem mass spectra, PC grid X!Tandem analysis is able to assign 341 spectra to a peptide, resulting in 166 different implied proteins, grouped into 35 clusters. Although no protein within a cluster can directly be excluded, attention should focus on top-ranked proteins (leaders) with a small protein e-value and a relatively large number of high-quality peptide matches. Secondary identifications of non-leader proteins can be made if unique high-quality peptide identifications exist.

All of the top 10 leaders are correct assignments, and all 16 true positives are found within the 21 best hits. Rabbit myosin is detected in three versions (MYH4, MLRS and MLE3), whereas three proteins presumed present in the mixture (CASB_BOVIN, LCA_BOVIN, MYG_HORSE) could not be detected at all with the chosen settings, an instance of analytical incompleteness. Although human serum albumin is not a leader, ranked second in its cluster behind the bovine homologue, it still is easily spotted as a valid identification based on the high number of high-quality unique peptides found. For several proteins, the species of origin could not be disambiguated, as multiple species share the leader position; e.g. the only identified actinic peptide “QEYDEAGPSIVHR” was simultaneously assigned to 35 different species.

As a graphical illustration of the implemented visualization tools and their utility for protein identification, we show four complementary representations of the *E. coli* alkaline phosphatase precursor peptide “AAGLATGNVSTAELQ(+1)DATPAALVAHV”^T, exhibiting glutamine deamidation at residue Q174, correctly identified from MS/MS spectrum #1983. Figure 1A and B show the doubly charged precursor ion peak in the LC-MS chromatogram and the corresponding MS/MS mass fingerprint with identified product ion peaks. The error bars above the spectrum indicate very low deviation between measured and modeled fragment m/z ratios. The hypothetical deamidation is confirmed by another spectrum assigned to the same peptide sequence, as shown by the vertical bar in the protein coverage graph (Figure 1C). The overall coverage of the identified protein is 37%. The high number of assigned peptides, several of which are unique, make PPB_ECOLI a protein lead within its homology group (Figure 1D).

4. Conclusions and outlook

In this paper, we have presented the development of a workspace for the analysis of protein MS data. It uses idle CPUs of desktop PCs to perform protein identification from MS spectra using X!Tandem, and makes results accessible through a web server. Development of this system is ongoing, and a pre-production version is currently being evaluated by our institute's mass spectrometry facility and its users. Future directions of this work will be to improve the features, and to relieve some of the drawbacks of the service's implementation as a web server. Spectra files, for instance, are currently uploaded through a web form, which is a cumbersome and lengthy process. We are also investigating alternate data staging strategies. Moreover, display of spectra and analysis data is based on static data only; efficient dynamic representation (including user annotations) will be enabled by a database back-end, currently undergoing specification. Lastly, we are continuously optimizing performance of the underlying application service/LRMS interface to enable high-throughput processing of large numbers of jobs.

The workspace integrates a set of tools to assess the validity of peptide and protein matches. Further research in this direction is the implementation of complementary identification tools (a-posteriori models), scoring functions, and plausibility tests. These orthogonal criteria may improve the discriminative power of protein identification beyond unidimensional score statistics-based measures.

In summary, the SWISS-TANDEM workspace promotes the de-centralization of both the computation and the interpretation of protein searches. Tools for interpretation are put into the hands of the research scientists, and the necessary compute power is obtained through federation of the idle capacity of available PCs. Clearly, this system will not make the opinion of mass spectrometry experts redundant. Instead, the workspace encourages biologist "customers" to examine and interpret the outcome of their experiments by themselves, and allows them to contribute domain-specific knowledge.

Acknowledgments

The authors would like to thank Jürgen Kopp for his work with the interface to the SWISS-MODEL repository and are gratefully for valuable feedback from Paul Jenö and Suzette Moes.

References

- [1] S. D. Patterson and R. H. Aebersold, *Nat Genet*, vol. 33 Suppl, pp. 311-23, 2003.
- [2] C. R. Cho, M. Labow, M. Reinhardt, et al., *Curr Opin Chem Biol*, vol. 10, pp. 294-302, 2006.
- [3] S. Ciordia, V. de Los Rios, and J. P. Albar, *Clin Transl Oncol*, vol. 8, pp. 566-80, 2006.
- [4] A. Schmidt and R. Aebersold, *Genome Biol*, vol. 7, pp. 242, 2006.
- [5] R. Aebersold and M. Mann, *Nature*, vol. 422, pp. 198-207, 2003.
- [6] D. Zosso, M. Podvinec, M. Müller, et al., presented at HealthGrid 2007, Geneva, 2007.
- [7] R. Craig and R. C. Beavis, *Bioinformatics*, vol. 20, pp. 1466-7, 2004.
- [8] K. Arnold, L. Bordoli, J. Kopp, et al., *Bioinformatics*, vol. 22, pp. 195-201, 2006.
- [9] M. Podvinec, S. Maffioletti, P. Kunszt, et al., presented at 2nd IEEE International Conference on e-Science and Grid Computing, Amsterdam, The Netherlands, 2006.
- [10] J. Kopp and T. Schwede, *Nucleic Acids Res*, vol. 34, pp. D315-8, 2006.
- [11] M. R. Wilkins, R. D. Appel, J. E. Van Eyk, et al., *Proteomics*, vol. 6, pp. 4-8, 2006.
- [12] Anonymous, <http://sashimi.sourceforge.net/repository.html>, Dec 18, 2006.
- [13] P. J. Kersey, J. Duarte, A. Williams, et al., *Proteomics*, vol. 4, pp. 1985-8, 2004.
- [14] C. H. Wu, R. Apweiler, A. Bairoch, et al., *Nucleic Acids Res*, vol. 34, pp. D187-91, 2006.