BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 16 2013, pages 2064–2065
doi:10.1093/bioinformatics/btt329

Systems biology

Advance Access publication June 21, 2013

Sipros/ProRata: a versatile informatics system for quantitative community proteomics

Yingfeng Wang[1], Tae-Hyuk Ahn[1], Zhou Li[2] and Chongle Pan[1,2,*]

[1]Computer Science and Mathematics Division, Oak Ridge National Laboratory and [2]Graduate School of Genome Science and Technology, University of Tennessee-Oak Ridge National Laboratory, Oak Ridge, 37831 TN, USA

Associate Editor: Olga Troyanskaya

ABSTRACT

Summary: Sipros/ProRata is an open-source software package for end-to-end data analysis in a wide variety of community proteomics measurements. A database-searching program, Sipros 3.0, was developed for accurate general-purpose protein identification and broad-range post-translational modification searches. Hybrid Message Passing Interface/OpenMP parallelism of the new Sipros architecture allowed its computation to be scalable from desktops to supercomputers. The upgraded ProRata 3.0 performs label-free quantification and isobaric chemical labeling quantification in addition to metabolic labeling quantification. Sipros/ProRata is a versatile informatics system that enables identification and quantification of proteins and their variants in many types of community proteomics studies.

Availability: Both programs are freely available under the GNU GPL license at Sipros.omicsbio.org and ProRata.omicsbio.org.

Contact: panc@ornl.gov

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on February 27, 2013; revised on May 30, 2013; accepted on May 31, 2013

1 INTRODUCTION

Community proteomics (metaproteomics) aims to characterize the whole protein complement of a microbial community. In a shotgun approach, proteins extracted from environmental samples of microbial communities are digested with trypsin into a more complex peptide mixture. Peptides are then separated by liquid chromatography and analyzed by tandem mass spectrometry (MS/MS). MS/MS data are searched against translated metagenomic sequences to identify proteins from different members of a microbial community. Post-translational modifications (PTMs) of proteins can be identified computationally by searching an expanded peptide space. Abundance changes of identified proteins can be measured between different environmental samples of a microbial community by quantitative proteomics. Common quantification methods include label-free approaches, ^{15}N metabolic labeling and isobaric chemical labeling with TMT and iTRAQ. Here we present a versatile software package for a wide range of community proteomics analysis. The package integrates a redesigned and reimplemented Sipros 3.0 program for highly scalable database searching and an upgraded ProRata 3.0 program for handling all common quantification methods.

*To whom correspondence should be addressed.

2 SOFTWARE FRAMEWORK

The input MS/MS data for Sipros/ProRata are FT1 and FT2 text files. The Raxport program is freely available at Raxport.omicsbio.org for extracting FT1 and FT2 files from raw files generated by Thermo Scientific mass spectrometers or from mzML files. The Sipros/ProRata package integrates modular programs into a variety of workflows (Supplementary Fig. S1). The input and output of these programs are all simple tab-delimited text files that can be streamed from one program to the next. Three programming languages—C++, Python and R—were used in the package. The computing-intensive tasks of database searching and selected ion chromatogram analysis were implemented in C++ for computational efficiency and scalability. Peptide filtering and assembly, spectral counting and isobaric chemical labeling quantification were implemented in Python for ease of development and maintenance. Statistical analysis was developed in R to take advantage of its wide selection of statistical tools. Different types of proteomics measurements were handled using different workflows that mix and match these modules in appropriate configurations. Owing to the extendibility of this software framework, we believe Sipros/ProRata is an open and flexible platform for further development of new bioinformatics tools for emerging proteomics technologies.

3 IDENTIFICATION BY SIPROS

Sipros was previously used for two niche applications in community proteomics: stable isotope probing (Pan *et al.*, 2011) and amino acid mutation identification (Hyatt and Pan, 2012) using high-resolution MS/MS. Here, the capability of Sipros 3.0 was expanded to perform general-purpose database searches and broad-range PTM searches. PTMs that undergo neutral loss during peptide fragmentation, such as phosphorylation, were searched by calculating precursor masses with intact modified peptides and computing fragment masses with neutral loss. Sipros has two scoring functions, one optimized for high-resolution MS/MS data and the other optimized for low-resolution MS/MS data. MS/MS data can be acquired by both CID and HCD. Database-searching results are filtered at a given peptide false discovery rate (FDR) level estimated using concatenated reverse sequences in the protein database.

Sipros 3.0 was developed from the ground up in a new code base to achieve excellent scalability from desktops to supercomputers. It is based on a hybrid parallelism that combines Message Passing Interface (MPI) and OpenMP. OpenMP is used to create shared memory multithreaded processes for multicore central

processing units (CPUs) on desktops and compute nodes of a supercomputer. Tens of thousands of MS/MS scans are loaded to the shared memory and compared with indexed candidate peptides by multiple threads concurrently. Extensive code optimization achieved a 2-fold speedup from the previous version by eliminating redundant computation and maximizing shared memory efficiency. Sipros 3.0 can finish database searching of a typical 24-hour low-resolution MS/MS dataset of a bacterial proteome in ~10 minutes on a quad-core desktop with 4 GB of memory. When running on a supercomputer, MPI was used to distribute many processes across compute nodes. Each process was responsible for a subset of MS/MS scans and protein sequences. Load balancing was performed dynamically at both the OpenMP level and MPI level. Supplementary Figure S2 shows a near-linear scalability of Sipros on a top open-science supercomputer, Titan, up to 24 000 CPU cores on 1500 compute nodes.

As the majority of published metaproteomics studies used Sequest (Eng *et al.*, 1994) for database searching, the performance of Sipros was compared with Sequest for peptide identification. In a benchmark 2-hour *Pseudomonas putida* proteome dataset with low-resolution MS/MS, Sipros identified ~10% more peptides than Sequest (5744 peptides by Sipros and 5154 peptides by Sequest). The peptide level FDR was controlled at 1% for both programs. To demonstrate its capability for broad-range PTM searches, Sipros was used to search a 22-hour metaproteome run of the acid mine drainage (AMD) community with high-resolution MS/MS. Sipros considered 34 PTMs (Supplementary Table S1) for 57 370 target protein sequences in the database. The search was completed in ~1 hour using 24 000 CPU cores on Titan. In total, 38 374 peptides were identified at 1% peptide FDR, which produced 3597 protein identifications with at least two peptides and a unique peptide (Supplementary Table S2). As a PTM on a position of a protein sequence can be identified by multiple peptides, peptides were mapped onto protein sequences and PTMs were tabulated by protein positions (Supplementary Table S3). A total of 9761 position-specific PTMs were identified in the AMD community. Many of these modifications stemmed from proteome sample preparation.

4 QUANTIFICATION BY PRORATA

ProRata was originally developed for metabolic labeling-based quantitative proteomics (Pan *et al.*, 2006) using identification results from Sequest/DTASelect. Here we integrated ProRata with Sipros and upgraded ProRata for label-free quantification and isobaric chemical labeling quantification. The label-free quantification was based on spectral counting. Spectral counts of proteins were normalized and compared between different samples using exact Poisson tests or Student's *t*-tests. Significantly changed proteins can be identified based on *P*-values with multiple comparison correction and spectral count differences. The FDR of detecting significantly changed proteins can be estimated with a permutation test, in which replicates are randomly shuffled among different conditions to estimate the number of false-positive hits.

ProRata 3.0 also performs label-free quantification based on the peak height of identified peptides in their selected ion chromatograms. Supplementary Table S4 lists the quantification results based on the identification results in Supplementary Table S2. This was used to estimate the relative abundance of a modified version of a peptide compared with its unmodified version based on the total peak height of their chromatography peaks (Supplementary Table S3).

In isobaric chemical labeling analysis, TMT- or iTRAQ-labeled peptides were quantified by normalized intensities of reporter ions with isotopic impurity correction. Relative abundances of proteins in different samples were estimated by summing up their peptide abundances. Supplementary Table S5 shows the iTRAQ analysis result from an AMD benchmark run at the expected abundance ratio of 4:1 for all proteins. The measured abundance ratios of proteins were close to the expected value of 4 (quartiles: $Q_1 = 3.9$, $Q_2 = 4.0$, $Q_3 = 4.2$). In a biological comparison, significantly changed proteins can be detected by fold changes and *P*-values estimated with the rank product test (Breitling *et al.*, 2004). FDR of protein quantification can be similarly estimated with a permutation test.

5 DISCUSSION

Sipros/ProRata provides a number of advantages in comparison with many existing proteomics packages. First, Sipros can perform large-scale database searching with hybrid parallelism. Its scalability allows searches against large metagenomic databases and searches for a large number of PTMs. Second, ProRata can be applied to all common quantification methods used in community proteomics, including spectral counting, metabolic labeling and isobaric chemical labeling. Third, the integration of Sipros/ProRata provides a comprehensive and flexible informatics system that can accomplish identification and quantification of proteins, PTMs, mutations and stable isotope probing.

ACKNOWLEDGEMENTS

We would like to thank Doug Hyatt for technical assistance with the Sipros upgrade and Robert Hettich, Gregory Hurst and Jill Banfield for helpful discussions.

Funding: This work was funded by the U.S. Department of Energy, Office of Biological and Environmental Research, Genomic Science Program. This research used resources of the Oak Ridge Leadership Computing Facility. Oak Ridge National Laboratory is managed by UT-Battelle, LLC, for the U.S. Department of Energy under contract DE-AC05-00OR22725.

Conflict of Interest: none declared.

REFERENCES

Breitling,R. *et al.* (2004) Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Lett.*, **573**, 83–92.

Eng,J.K. *et al.* (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.*, **5**, 976–989.

Hyatt,D. and Pan,C. (2012) Exhaustive database searching for amino acid mutations in proteomes. *Bioinformatics*, **28**, 1895–1901.

Pan,C. *et al.* (2006) ProRata: a quantitative proteomics program for accurate protein abundance ratio estimation with confidence interval evaluation. *Anal. Chem.*, **78**, 7121–7131.

Pan,C. *et al.* (2011) Quantitative tracking of isotope flows in proteomes of microbial communities. *Mol. Cell. Proteomics*, **10**, M110.006049.