CoreFlow: A computational platform for integration, analysis and modeling of complex biological data

Pasculescu, Adrian; Schoof, Erwin; Creixell, Pau; Zheng, Yong; Olhovsky, Marina; Tian, Ruijun; So, Jonathan; Vanderlaan, Rachel D.; Pawson, Tony; Linding, Rune

Published in: Journal of Proteomics

Link to article, DOI: 10.1016/j.jprot.2014.01.023

Publication date: 2014

Document Version
Publisher's PDF, also known as Version of record

Link back to DTU Orbit

Citation (APA):
Technical note

CoreFlow: A computational platform for integration, analysis and modeling of complex biological data

Adrian Pasculescu, Erwin M. Schoof, Pau Creixell, Yong Zheng, Marina Olhovska, Ruijun Tian, Jonathan So, Rachel D. Vanderlaan, Tony Pawson, Rune Linding, Karen Colwill

ARTICLE INFO
Available online 3 February 2014

Keywords:
Computational pipeline
Mass spectrometry
Data analysis
Statistical analysis
Workflow

ABSTRACT
A major challenge in mass spectrometry and other large-scale applications is how to handle, integrate, and model the data that is produced. Given the speed at which technology advances and the need to keep pace with biological experiments, we designed a computational platform, CoreFlow, which provides programmers with a framework to manage data in real-time. It allows users to upload data into a relational database (MySQL), and to create custom scripts in high-level languages such as R, Python, or Perl for processing, correcting and modeling this data. CoreFlow organizes these scripts into project-specific pipelines, tracks interdependencies between related tasks, and enables the generation of summary reports as well as publication-quality images. As a result, the gap between experimental and computational components of a typical large-scale biology project is reduced, decreasing the time between data generation, analysis and manuscript writing. CoreFlow is being released to the scientific community as an open-sourced software package complete with proteomics-specific examples, which include corrections for incomplete isotopic labeling of peptides (SILAC) or arginine-to-proline conversion, and modeling of multiple/selected reaction monitoring (MRM/SRM) results.

Biological significance
CoreFlow was purposely designed as an environment for programmers to rapidly perform
Mass spectrometry is a computationally intense experimental procedure. A wide variety of software applications have been developed to facilitate analysis, including search engines that interpret the spectra [1–4], databases that hold raw and processed results [5–7], and sophisticated algorithms that calculate statistical significance [8–10]. Even with the availability of all these programs, the ever-changing nature of research often requires scientists to develop their own scripts to handle a specific analysis and, to integrate mass spectrometry data with other experimental data. To address this need, we have built a software application, CoreFlow, which provides an organized framework for programmers to rapidly perform, document and share data analysis steps.

CoreFlow is designed to manage all data manipulation steps within a project (Fig. 1A). It is divided into two main sections: database management and analysis (Fig. 1B). The database management section is used for loading, storing and handling of experimental data in a highly efficient yet flexible manner (Supplementary Figs. 1–5). Here, one can perform basic database operations such as creating new tables, modifying existing ones, adding key indexes, and filtering, joining or aggregating data. The analysis section assembles and then tracks computational pipelines using a hierarchical organization of owners, projects, threads and tasks (Supplementary Fig. 6); a thread represents a particular analysis (e.g., Multiple Reaction Monitoring) and, within each thread, a task holds the code for a specific function (e.g., normalization to bait). These pipelines provide a permanent record of all the computational work performed inside each specific project. The standardized format reveals the complete data analysis pipeline to all users, thereby promoting sharing of data, scripts, and results. Within a pipeline, newer tasks are often built using information extracted from previous tasks via application programming interface (API) calls. These calls allow access to a library of data processing scripts for common functions such as data quality checks and normalization, enable re-use of code for data extraction from tables, and provide the ability to obtain content from attachments linked to tasks (Supplementary Fig. 7). To help track task interdependencies, we created a built-in data provenance feature that identifies these calls (Supplementary Fig. 8). Similarly, data within a task is often extracted from multiple database tables, and CoreFlow maps the relationships between the original tables and the resulting temporary and final tables (Supplementary Fig. 9). At task completion, a detailed analysis report can be generated that includes the task description, code, results and data provenance features (Supplementary Text).

New tasks within a thread can easily be added using the blank template provided or by duplicating and editing an existing task. If an existing task is copied, the user only needs to change the parameters of any API calls (in particular, the unique task identifier) and update table names in the SQL scripts. Each task includes administrative features for version tracking, quality control, priority setting, and file attachment (Supplementary Fig. 10). Icons can be added to the task label to serve as visual cues to the task’s purpose. Three integrated playgrounds (similar to Google Code Playgrounds) accelerate analysis by allowing for a rapid cycle of script development and testing. The first playground with an embedded Wiki page enables the addition of meta information about the analysis and its purpose, assumptions and caveats (Supplementary Fig. 11). It can also be used to indicate where and how certain parameters can be changed in the code (e.g., cutoff values, number of data points to be included). This is important, as CoreFlow is designed for rapid analysis and programming where developing user-specific interfaces for each task to accommodate parameter entries is not desirable as it would slow the pace of analysis. A second playground allows for complex database queries that pre-process different data types and sources. The final processing occurs in the third playground where users write scripts in various programming languages to handle the data. These scripts can be readily transferred, without any changes, either to a local programming environment for debugging or a high performance system for speed. They may vary in terms of their complexity and level of abstraction, from bioinformatic analysis in Python or Perl to more detailed statistical analysis in R. For each of these programming languages, CoreFlow allows the import and use of their specific libraries and packages as well as creation of new libraries. This permits, for example, large string processing or manipulation and
transformation of sequences (e.g., in silico digestion of proteins into peptides) using BioPython or BioPerl. Libraries in R can also be employed for statistical analysis, data visualization, and immediate rendering of images in SVG, PDF or other graphic formats for inclusion in manuscripts. Results can be exported in compatible formats for visualization or modeling tools such as Cytoscape [11] or DataRail [12].

Due to its flexibility, CoreFlow can be applied to practically any type of data analysis. Our labs have used it in over fifty very diverse projects that include large-scale mass spectrometry, RNAi, next-generation sequencing, imaging and animal models [13-16]. We typically use it in conjunction with complementary tools targeted to a particular experimental application (e.g., MaxQuant [4] and Proteome Discoverer (ThermoScientific) for mass spectrometry quantitation). We quickly move data (in XML, flat or binary file formats) from such applications to the CoreFlow database where we extract relevant data and model it (Figs. 1a, 2, Supplementary Fig. 12).

In a recent paper involving MRM [16], CoreFlow was used to normalize all samples within a time course to the bait, examine reproducibility between samples, and cluster proteins with similar profiles (Fig. 2). It has helped identify systematic errors that are often introduced in experiments and to correct them computationally, leading to cleaner data and improving the accuracy of results. For example, we have estimated and corrected labeling incorporation deficiencies and arginine-to-proline conversions that affect mass spectrometry measurements when using Stable Isotope Labeling by Amino acids in Cell culture (SILAC) (Fig. 3, Supplementary Fig. 13, Supplementary Text).

CoreFlow implements a client–server architecture, supported by the Apache web server on the Linux operating system, with a web interface written in PHP and Perl, and an underlying MySQL database. This implementation provides the advantages of centralized access, resource distribution, back-up capabilities, and easy upgrading. It can be deployed

Fig. 1 – The CoreFlow pipeline. (A) CoreFlow is designed to integrate different types of experimental input data as well as computational scripts and database functionalities in order to correct and analyze data before publication-quality figures are produced. (B) From the main menu, a user can read CoreFlow’s documentation, deploy the different database functionalities (create, browse or query database tables as well as import data from files using the ‘file to DB function’) or access analytical pipelines. Within the Analysis module, a project-specific pipeline is created as data inputs and analytic output expand with research progress.
on a local computer, an application server, or even on high-performance computing systems to improve speed. The work-load can be distributed across larger computational infrastructures through parallelization of R jobs and other scripts. Any size limits on data are imposed only by the operating environment (Linux, Apache, MySQL, PHP) and not by CoreFlow itself. Issues of browser timeouts are avoided by forking long tasks into child processes that maintain communication with the browser.

CoreFlow is now being released as an open-source application to assist other research scientists in managing complex data, correcting systematic errors, assessing the quality of
measurements, finding hidden patterns in experimental data, and ultimately, in modeling biological systems. The source code, database archive, documentation and demonstration are available at http://coreflow.mshri.on.ca, as is a link to a GitHub repository for code sharing and to a Google group for online discussions and assistance. The standard installation package and the on-line public version include a number of small and well-documented scripts (see Supplementary Table 1 for a list of scripts) outlining the workflow process. If desired, CoreFlow can be downloaded to run in a VirtualBox Linux environment.

We anticipate that CoreFlow will become a valuable resource for researchers who require a streamlined, flexible framework for analyzing and sharing biological data. Moving forward, we plan to release all code related to our publications in this format and we encourage other potential users to share their workflows too. By promoting the public release of code, CoreFlow can help eliminate the 'black box' issue [17–19] of unpublished code and reduce the use of unsuitable software [20] in publications. In this manner, we expect that the content of CoreFlow will become as valuable as the software itself.

Author contributions


Competing financial interests

The authors declare no competing financial interests.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.jprot.2014.01.023.

Acknowledgments

Thanks to Naveed Mohammed for systems administration support, and Sören Brunak, Jinho Kim and Evangelia Petsalakis for comments on the manuscript. This work was supported by funds to T.P. from Genome Canada through the Ontario Genomics Institute, the Ontario Research Fund GI2 program, and the Canadian Institutes of Health Research (CIHR, MOP-6849, MOP-13466). R.L. is supported by the Lundbeck Foundation, the Human Frontier Science Program (HFSP), the Danish Council for Independent Research (FSS) and the European Research Council (ERC). R.D.V., J.S. and R.T. received fellowship support from CIHR.

REFERENCES

A. Illustration of labeling efficiency and conversion:

- **Complete labeling efficiency**:
  - Unlabeled sample: [sequence]  
  - Labeled sample: [sequence]  
  - True ratio: 6:3
  - Measured ratio: 6:3

- **Incomplete labeling efficiency**:
  - Unlabeled sample: [sequence]  
  - Labeled sample: [sequence]  
  - True ratio: 6:3
  - Measured ratio: 5:4

- **Arginine-to-Proline conversion**:
  - Unlabeled sample: [sequence]  
  - Labeled sample: [sequence]  
  - True ratio: 6:3
  - Measured ratio: 4:3

B. Graph showing measured ratio vs. true ratio:

- Incorporation percentages: 0.9, 0.95, 0.98, 0.99, 1

C. Graph showing measured ratio vs. true ratio:

- Nr. of Prolines: 0, 1, 2, 3

D. Steps for correcting labeling:

1. Assessment of labeling efficiency
2. Assessment of Arg-Pro conversion
3. Computational correction

   - MR = Measured Ratio
   - TR = True Ratio
   - nP = number of Prolines
   - e = Labeling Efficiency (e.g., 0.96)
   - RP = Arg/Pro Conversion Rate (e.g., 0.08)

   \[ TR = \frac{1}{e^{nP(RP)(e-1)}} \]

E. Diagrams showing correcting steps:

- Measured ratio correction
- Correcting Arg -> Pro labeling errors


Fig. 3 – Computational correction of SILAC inaccuracies using CoreFlow. (A) Incomplete labeling (center) and arginine-to-proline conversion (right) are two phenomena that can make SILAC ratios inaccurate and far from their true ideal values (left). (B) Incomplete labeling deviates the measured ratio from the true ratio in a manner that is non-linearly correlated with the SILAC incorporation percentage. For severely unlabeled samples (e.g. 0.9 incorporation), it becomes impossible to measure ratios that are above a threshold. (C) Arginine-to-proline conversion is, unsurprisingly, dependent on the number of prolines in the MS peptide; peptides with more prolines are more prone to wrong measurements. (D) After assessing labeling deficiency and arginine-to-proline conversion rates in the sample of interest (please refer to the Supplementary Text for further details), the measured ratios can be corrected using the formula provided. (E) By using this formula, measured SILAC ratios affected by both labeling and arginine-to-proline conversion inaccuracies can be converted to correct ratios that match unaffected samples.