Assessing the problem of doing proteomics with unsequenced organisms*

Róger Bermúdez-Chacón†
Supervisor: Katja Bärenfaller‡

June 10, 2013

Introduction

When dealing with unannotated or partially annotated proteomes, protein identification is often carried out by searching a database containing protein sequences from other species. Identification of a protein from a different species is often assumed as enough evidence to claim that this particular protein was identified in a protein sample. This work attempts to systematically evaluate at what extent this holds true, by comparing search results of unannotated proteins extracted from the cassava (*Manihot esculenta*) root against different annotated databases, including the reference species *Arabidopsis thaliana*, the Viridiplantae database (containing proteins from a large number of green plants), and the annotated proteome of cassava itself.

Materials, Tools and Methods

Data sources

Cassava root sample

A sample from the cassava root was analyzed with an LTQ-Orbitrap mass spectrometer. The resulting ion spectra were used in the queries.

Proteome databases

For queries against *Arabidopsis thaliana*, the database `fgcz_3702d_TAIR10`[1] (from the Functional Genomics Center Zurich (FGCZ), 29.042.854 residues in 71.033 sequences from *Arabidopsis thaliana* only) was used. For queries against Viridiplantae, the database `fgcz_viridi_d` (from the FGCZ, containing 2.062.037 sequences, 693.434.028 residues from annotated proteomes of many species of ''green plants'' –including *Manihot esculenta*–) was used. For queries against cassava itself, the database `P764_db2_d` [2] (entire annotated proteome of *Manihot esculenta* (cassava), 68.888 sequences from 26.812.366 residues) was used. A copy of the cassava database, in `Fasta` format, was also used during the data processing.

Tools

The queries against all proteome databases were performed using `Mascot v2.4.1` [3]. Additionally, `BlastP` [4, 5] queries were performed to search for additional homologs when required.

The results were processed and the data were analyzed with `R` [6]. The auxiliar package `seqinr` [7] was used to read `fasta` files.

Methods

The ion spectra read from the cassava root protein sample was compared in `Mascot`, by performing MS/MS ions search queries against each of the three databases listed above.

The occurrence of carbamidomethyl groups and oxidation were used in all queries as fixed and variable modifications, respectively.

*This project was held as a Lab Rotation in Bioinformatics, as required by the Master program in Computational Biology and Bioinformatics - ETH Zürich

†D-INFK - Computational Biology and Bioinformatics Master program
‡D-BIOL - Plant Biotechnology Group

The queries were performed over all entries in each database. *Trypsin* was used as the cleaving enzyme (up to 1 missed cleavages allowed). The peptide and fragment mass tolerances used were $10\ ppm$ and $0.8\ Da$, respectively.

The score reported for ranking the matches found was $-10 \log P_m$, being P_m the probability that a match m happened as a random event.

Search against the *Arabidopsis thaliana* database
: The query was restricted to results with a false-positive tolerance (significance threshold) of at most 0.05, and ion scores of at least 24.

Search against the Viridiplantae database
: The query was restricted to results with a significance threshold of at most 0.05, and ion scores of at least 38.

Search against the *Manihot esculenta* database
: The query was restricted to results with a significance threshold of at most 0.05, and ion scores of at least 25.

The query results were exported from `Mascot` for further processing as comma-separated text files with a formatted header including the search parameters and metadata.

The data processing was coded as an `R` script. The implementation relies heavily upon regular expression transformations and set operations. The data sets were processed as follows:

- Each of the query result files was parsed into a data structure of the form $[header, data]$, with $header$ being a set of $name, value$ pairs (*dictionary*) of variables read from the file header, and $data$ the actual values read from the comma-separated file section.
- The cassava database, provided as a `Fasta` file, was read into an `R` object with sequence information by using the `R` package `seqinr`.
- Removal of contaminant references and alternative splice specifications was done by regular expression matching against the protein identifiers. Here, proteins with identifiers prefixed with contaminant markers (`REV_, ZZ_, zz|ZZ_, rr|REV_`) were filtered out of the results. Alternative splice specifications (suffixes matching '`.[0-9]+`' in the protein identifier) were ignored in Arabidopsis protein descriptors by using only the suffix-free identifier in protein comparisons.
- The mapping between protein identifiers of cassava and Arabidopsis, to infer homology, was done by parsing the `Fasta` descriptors over the whole cassava database. Entries for which homology information between cassava and Arabidopsis has been validated comply with the format

  $>$ [**cassava_id**] | [`PACid`] | [`Annot`] | [...] | [**arabidopsis_id**] | [...] | [`protein_description`]

  and as such, count with both species identifiers.

In particular, the following data were extracted:

Query results of the protein sample against the *Arabidopsis thaliana* database

1. Different proteins identified by the query. Performed by listing the unique protein identifiers after contaminant removal (ignoring alternative splicing specification suffixes).
2. Identified peptides, and proteins from the cassava database that contain them. Performed by listing the unique peptides that satisfy the scoring, significance and rank conditions defined in the search parameters. These peptides are compared in an 'all-against-all' affix (substring) search between each peptide and all the entries in the cassava database.
3. Cassava homologs of the Arabidopsis search results. Performed by listing the proteins from the cassava database whose `Fasta` descriptor refers to any cassava protein returned from the results in 1.
4. Cassava proteins both containing high-scoring peptides from the query against the Arabidopsis database and being labeled as homologs of the high-scoring Arabidopsis proteins. Performed by a set intersection between the results from 2. and 3.

Query results of the protein sample against the Viridiplantae database

5. Different proteins and protein groups identified by the query. Performed by listing unique protein identifiers after contaminant removal (ignoring alternative splicing specification).

6. Identified cassava proteins. Performed by listing the unique returned proteins from the Viridiplantae database whose protein identifier corresponds to a cassava (*Manihot esculenta*) protein (i.e. protein identifiers with suffix `_MANES`). Since the Viridiplantae database does not return protein identifiers in the format the cassava database uses, the mapping between results from Arabidopsis and cassava databases had to be done by sequence and/or protein description comparison; conflicts were resolved manually when ambiguities were found.
7. High-scoring protein groups with no proteins in the cassava proteome. Performed by identifying the protein groups with cassava proteins, and removing them from the list of returned protein groups.
8. Cassava homologs of the proteins in the highest-scoring protein group with no cassava proteins. Performed by ranking the proteins returned by 7. and searching for homologs of the highest-scoring protein in `Blastp`. If no cassava homologs are listed, a search by protein description is performed on the cassava database.

Query results of the protein sample against the cassava proteome

9. Different proteins identified by the query. Performed by listing the unique protein identifiers after contaminant removal.
10. Cassava proteins occurring in the cassava database search and in the homologs inferred from Arabidopsis and Viridiplantae database search. Performed by a set intersection between the results from 4., 8., and 9.

Results

The results, summarized in Figure 1, were as follows:

Query results of the protein sample against the *Arabidopsis thaliana* proteome

1. The query returned entries for `29253` different protein/peptide combinations. `1442` different proteins returned significant matches. After contaminant removal, and ignoring alternative splicing, `840` proteins from Arabidopsis were found in the query results.
2. As many as `13735` protein/peptide entries that satisfy the score, significance and rank conditions given by the search parameters were found. These entries correspond to `1677` different peptides from the Arabidopsis database. These peptides were found in `1087` different proteins in the cassava database.
3. `1297` out of the `1442` Arabidopsis proteins found in 1. have been mapped to cassava proteins, and are thus assumed as confirmed homologs.
4. The cassava proteins found both by high-scoring Arabidopsis peptide search (2.) and annotated homology (3.) were `523`

Query results of the protein sample against the Viridiplantae database

5. `4990` different proteins (excluding contaminants) from `652` different protein groups were found.
6. Only `24` cassava proteins were returned by the query. These cassava protein identifiers from the Viridiplantae database were mapped to the actual cassava database by sequence and/or description comparison. Matches with only `20` proteins from the cassava database were found.
7. `632` out of the `652` protein groups found in 5. do not have any cassava protein.
8. From these groups without cassava proteins, the highest-scoring proteins found correspond to phosphopyruvate hydratases (*enolases*). The protein with the best match (entry name `E4MVZ0_THEHA`) is described in the database as `mRNA, clone: RTFL01-03-H18 OS=Thellungiella halophila`. An additional query against the UniProt[8] database revealed that this protein is part of the phosphopyruvate hydratase complex, and is also related to the enolases in the corresponding family and domain databases.

   Since no cassava homologs to this top-score protein were identified by `BlastP`, a search by description using both `enolase` and `phosphopyruvate hydratase` was performed on the cassava database. `5` proteins from the cassava database were found.

Query results of the protein sample against the cassava proteome

9. `1412` different proteins from the cassava database were found in the query.
10. None of the proteins occurred in all query results.

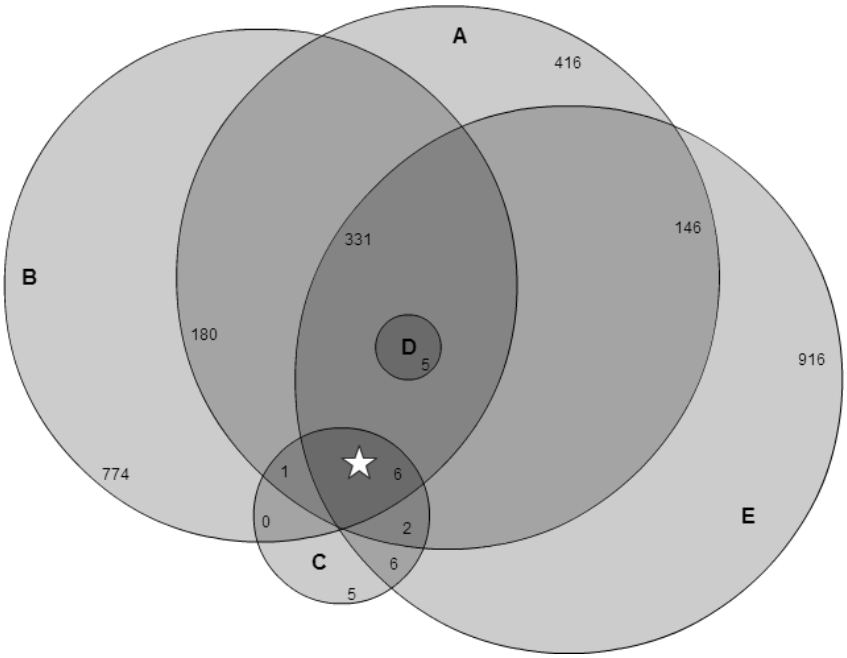

Figure 1. Protein distribution for different queries. Each region is labeled with the number of different proteins found on such region exclusively. A star denotes the region where the homology assumption is supported by most of the queries.
A Cassava proteins containing high-scoring peptides found on the query against the Arabidopsis database.
B Annotated cassava homologs of the high-scoring Arabidopsis proteins.
C Identified cassava proteins found on the query against the Viridiplantae database.
D Cassava homologs of the proteins whose group does not contain cassava proteins.
E Cassava proteins found on the query against the cassava database.

Discussion

It is worth noticing from the results above, that no single query returned results supported by all of the others. The region denoted by a star in Figure 1. shows an agreement between the possible homologs inferred from the queries against the Arabidopsis database, the cassava proteins found on the Viridiplantae database, and the results from the cassava database itself, and is thus the region with the best candidates to annotate the sample.

The global overlap between results from the Arabidopsis high-scoring peptides (**A**), Arabidopsis-cassava annotated homologs (**B**), and cassava database results (**E**), consists of only up to a third of the results returned by each query separately. This number narrows further down when the identified cassava proteins in the Viridiplantae query (**C**) are considered. Reliable identification by any of the queries alone is hence not possible.

In the results from the Viridiplantae database query, protein candidates to label the sample were clustered together in protein groups or families given by protein similarity. Some of the high-scoring protein families did not have any cassava protein annotations. Cassava proteins shown as **D** in Figure 1. were found –via `BlastP`– by similarity search against the highest-scoring protein from those protein groups. This best-scoring protein was an *enolase*, and none of the 24 cassava proteins found against the Arabidopsis database corresponded to neither *enolases* nor *phosphopyruvate hydratases*.

One obvious next step will be to expand the results from **D** to include more results, for example, from cassava proteins similar not to the best-scoring hit only, but rather to a large number of high-scoring proteins. The identity and function of cassava protein matches not identified by the query against the Viridiplantae database might also be of interest for a future work.

Conclusions

Since agreement on candidate proteins for sample annotation only occurs on a fraction of the query results, the assumption of correct identification by using a single reference database can definitely not be taken for granted. There is overwhelming evidence that protein identification strongly depends on the database chosen for queries. Furthermore, because this approach only tries to find protein correspondences, the annotation of parts of the protome specific to a given organism still poses a problem.

When possible, it is adviseable to work with fully-annotated proteomes and avoid such assumptions altogether. However, when this is not an option, one must be very cautious when assuming real interspecies protein correspondences with unannotated proteomes.

Bibliography

[1] P. Lamesch, T. Z. Berardini, D. Li, D. Swarbreck, C. Wilks, R. Sasidharan, R. Muller, K. Dreher, D. L. Alexander, M. Garcia-Hernandez, *et al.*, ''The arabidopsis information resource (tair): improved gene annotation and new tools,'' *Nucleic acids research*, vol. 40, no. D1, pp. D1202--D1210, 2012.

[2] S. Prochnik, P. Marri, B. Desany, P. Rabinowicz, C. Kodira, M. Mohiuddin, F. Rodriguez, C. Fauquet, J. Tohme, T. Harkins, D. Rokhsar, and S. Rounsley, ''The cassava genome: Current progress, future directions,'' *Tropical Plant Biology*, vol. 5, no. 1, pp. 88--94, 2012.

[3] M. Hirosawa, M. Hoshida, M. Ishikawa, and T. Toya, ''Mascot: multiple alignment system for protein sequences based on three-way dynamic programming,'' *Computer applications in the biosciences: CABIOS*, vol. 9, no. 2, pp. 161--167, 1993.

[4] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, ''Gapped blast and psi-blast: a new generation of protein database search programs,'' *Nucleic acids research*, vol. 25, no. 17, pp. 3389--3402, 1997.

[5] S. F. Altschul, J. C. Wootton, E. M. Gertz, R. Agarwala, A. Morgulis, A. A. Schäffer, and Y.-K. Yu, ''Protein database searches using compositionally adjusted substitution matrices,'' *Febs Journal*, vol. 272, no. 20, pp. 5101--5109, 2005.

[6] R Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013.

[7] D. Charif and J. Lobry, ''SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis.,'' in *Structural approaches to sequence evolution: Molecules, networks, populations* (U. Bastolla, M. Porto, H. Roman, and M. Vendruscolo, eds.), Biological and Medical Physics, Biomedical Engineering, pp. 207--232, New York: Springer Verlag, 2007. ISBN : 978-3-540-35305-8.

[8] T. U. Consortium, ''Reorganizing the protein space at the universal protein resource (uniprot),'' *Nucleic Acids Research*, vol. 40, no. D1, pp. D71--D75, 2012.