Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Applications Note

OXFORD

Sequence analysis

Rabifier2: an improved bioinformatic classifier of Rab GTPases

Jaroslaw Surkont *, Yoan Diekmann † and José B. Pereira-Leal *

Instituto Gulbenkian de Ciência, Oeiras, 2780-156, Portugal

*To whom correspondence should be addressed.

†Current address is MACE-Lab, Research Department of Genetics, Evolution and Environment University College, London, UK

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Summary: The Rab family of small GTPases regulates and provides specificity to the endomembrane trafficking system; each Rab subfamily is associated with specific pathways. Thus, characterization of Rab repertoires provides functional information about organisms and evolution of the eukaryotic cell. Yet, the complex structure of the Rab family limits the application of existing methods for protein classification. Here, we present a major redesign of the Rabifier, a bioinformatic pipeline for detection and classification of Rab GTPases. It is more accurate, significantly faster than the original version and is now open source, both the code and the data, allowing for community participation.

Availability and Implementation: Rabifier and RabDB are freely available through the web at http://rabdb.org. The Rabifier package can be downloaded from the Python Package Index at https://pypi.python.org/pypi/rabifier, the source code is available at Github https://github.com/evocell/rabifier.

Contact: jsurkont@igc.gulbenkian.pt, jleal@igc.gulbenkian.pt

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

1 Introduction

The Rab family, the largest member of the Ras superfamily of small guanine nucleotide-binding proteins, is a key regulator of vesicle trafficking in eukaryotic cells. This highly paralogous family can be further divided into subfamilies associated with specific trafficking pathways. Rab function tends to be conserved across species, for example, Rab1 in mouse can functionally replace its orthologue in yeast (Haubruck *et al.*, 1989). Hence, Rabs are one of the 'diagnostic' gene families whose annotation informs about presence and evolution of trafficking functions and pathways in Eukaryotes. However, classification into subfamilies is complicated as paralogues are very similar to each other, so it has traditionally been done manually using bespoke approaches (e.g., Pereira-Leal, 2008; Ackers *et al.*, 2005; Elias *et al.*, 2012). Previously we developed a bioinformatic method to automatically classify Rabs that uses multiple decision steps and a manually curated reference set of Rab subfamilies (Diekmann *et al.*, 2011). We also created a web-accessible database (RabDB) where we display Rab annotation for all genomes available as a part of Superfamily 1.75 (Gough *et al.*, 2001) at the time, alongside a web tool that allows users to annotate submitted sequences.

Rabifier and RabDB have provided means to the community to explore the Rab universe. Yet, the availability of substantially updated third-party programs used as part of our annotation pipeline (e.g., HMMER3) prompted us to recode and partially redesign our pipeline, providing both better and faster annotations. The latter is especially important given the ever increasing amount of genomic data. The new version of the pipeline adds new features and improves on both accuracy and speed of sequence classification. Rabifier has been released as an open-source software to facilitate the further community-driven development of the classifier.

2 Rabifier2 & RabDB2

The Rabifier pipeline (fig. S1) has two main parts: an input protein sequence is classified whether or not it belongs to the Rab family (phase 1), and if it is a Rab, which subfamily it most likely belongs to (phase 2). Rab family assignment is based on satisfying three conditions: (1) presence of the G domain, (2) the top hit against the reference database is a Rab, (3) at least one RabF motif (Pereira-Leal and Seabra, 2000) is present. In the second phase, Rabifier measures similarities between the query protein and the reference Rab subfamily datasets to assign a confidence score to

each subfamily prediction. Alternatively, if the sequence is only marginally similar to any of the subfamilies, it is classified as RabX (unknown/new Rab). Both phases rely on manually curated sets of protein sequences that include Rabs, representatives of each Rab subfamily, and other small GTPases of the Ras superfamily.

Rabifier updates include changes to both the reference databases and the pipeline. Among numerous modifications to the original pipeline (text S1), two are the most noticeable: (1) HMMER3 replaces BLAST in the majority of similarity searches, (2) subfamily classification system is now based on sequence score comparison against a model of each subfamily, which is subsequently used as input for the naive Bayes classifier.

Improvements – performance. We measured classification performance of the new Rabifier pipeline by performing a 10-fold cross-validation analysis on the training dataset (https://github.com/evocell/rabifier-data). The original sequences defining each Rab subfamily and non-Rab GTPases were randomly partitioned into 10 equally-sized subsamples. A single subsample was used as the validation data and the remaining subsamples were used to train the classifier. The process was repeated 10 times with each subsample being used once as the validation data. We observed very high classification accuracy for both phases of the Rabifier2 pipeline (fig. 1A, S2).

We further tested the performance of Rabifier2 on an external set of more than 400 manually curated eukaryotic Rabs (Elias *et al.*, 2012), which were not used to train the classifier (Dataset S1), and compared it with the original Rabifier. Although we do not expect substantial errors in the careful phylogenetic annotation by Elias *et al.* (2012), we note that the Rabifier's classifications repeating errors by Elias *et al.* (2012) would be considered true positives. The two phases are considered separately. The high performance of the original classifier in phase 1 left little space for improvement, we did not observe any significant difference between the implementations (fig. S3). Yet, in phase 2 the difference is substantial; Rabifier2 provides correct subfamily assignments for many sequences misclassified by Rabifier1 (7% increase in correct annotations, 53% decrease in incorrect annotations; fig. 1B). We also compared the performance of Rabifier2 with a method developed by Klöpper *et al.* (2012), the only other available method for automatic Rab classification, which, however, only allows analysing one sequence at the time through the web; Rabifier2 provides much higher specificity at a small cost in sensitivity (table S1, S2). Finally, we tested if the high performance of Rabifier2 is consistent across different Rab subfamilies and eukaryotic supergroups. For the majority of the subfamilies present in the test dataset Rabifier2 provides correct annotations (fig. S4). Similarly, the classification performance is consistent across the eukaryotes (fig. S5).

The Rabifier2 codebase has been redesigned and rewritten, the third party software used in the pipeline has been updated to the most recent versions. This resulted in a major speed improvement (up to 10 fold, fig. 1C). Rabifier2 makes better use of parallel processing: the total computation time increases very slowly with the number of simultaneously classified sequences, compared to the old implementation. This major speedup allows for fast annotation of hundreds of genomes.

Improvements – access. Rabifier source code and the reference database are now freely available, which allows running Rabifier locally. In addition, precomputed Rab annotations for all (244) eukaryotic species present in Ensembl databases (Flicek *et al.*, 2014) are available in RabDB, which will remain up-to-date with new Ensembl database releases, providing Rab annotation to newly sequenced species. The improved web interface enables interactive exploration of the Rab family in selected species, including the navigation of taxonomy, and drawing phylogenetic profiles of presence/absence of Rab subfamilies in chosen taxa. Each protein contains a detailed annotation and is linked to the corresponding entry in the Ensembl database. The website also allows submitting protein sequences

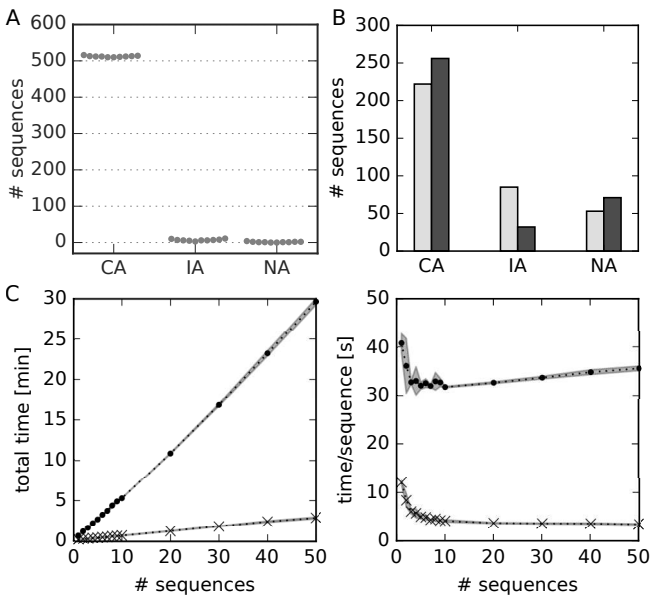

Fig. 1. (A) Rabifier2 (phase 2) 10-fold cross-validation, each dot represents a result of one cross-validation repeat. (B) Phase 2 performance comparison between Rabifier1 (light gray) and Rabifier2 (dark gray) using Dataset S1. (C) Speed comparison between Rabifier1 (•) and Rabifier2 (×), the total time for a given number of sequences (left) and time per sequence (right) using up to 4 CPU cores. Abbreviations: CA (Correct Annotation), IA (Incorrect Annotation), NA (No Annotation, RabX).

for classification; due to the performance increase, user can now upload hundreds of sequences at a time, compared to 5 sequences in the original version. It is also now possible to change several parameters used by the classifier and view a detailed output for each annotation.

3 Conclusions

Rabifier2 provides major improvements in Rab annotation, both in terms of speed and accuracy, over the initial version. It is the only publicly available tool for large-scale annotation of Rab GTPases, which can be used in genome annotation pipelines. We used it to annotate Rab diversity across Eukaryotes, which can be explored through the web. We have also released the source code of Rabifier to facilitate further development of the pipeline, enabling, for example, its extension to other protein families.

Acknowledgements

The authors thank all members of the Computational Genomics Laboratory for helpful discussions. In particular, we wish to thank Marc Gouw for help with the implementation of the Rabifier and RabDB interfaces. We would also like to thank the Bioinformatics Unit of the Instituto Gulbenkian de Ciência for hosting RabDB.

Funding

This work has been supported by the Fundação para a Ciência e a Tecnologia, under the grant PTDC/EBB-BIO/119006/2010, and PhD fellowship SFRH/BD/51880/2012 to JS.

References

Ackers, J. P., *et al.* (2005). A bioinformatic analysis of the RAB genes of Trypanosoma brucei. *Molecular and Biochemical Parasitology*, **141**(1), 89 – 97.

Diekmann, Y., *et al.* (2011). Thousands of Rab GTPases for the cell biologist. *PLoS Comput. Biol.*, **7**(10), e1002217.

Elias, M., *et al.* (2012). Sculpting the endomembrane system in deep time: high resolution phylogenetics of Rab GTPases. *J. Cell Sci.*, **125**(10), 2500–2508.

Flicek, P., *et al.* (2014). Ensembl 2014. *Nucleic Acids Res.*, **42**(Database issue), D749–55.

Gough, J., *et al.* (2001). Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J. Mol. Biol.*, **313**(4), 903–19.

Haubruck, H., *et al.* (1989). The ras-related mouse ypt1 protein can functionally replace the YPT1 gene product in yeast. *The EMBO journal*, **8**(5), 1427.

Klöpper, T. H., *et al.* (2012). Untangling the evolution of Rab G proteins: implications of a comprehensive genomic analysis. *BMC Biology*, **10**(1), 71.

Pereira-Leal, J. B. (2008). The Ypt/Rab family and the evolution of trafficking in fungi. *Traffic*, **9**(1), 27–38.

Pereira-Leal, J. B. and Seabra, M. C. (2000). The mammalian Rab family of small GTPases: definition of family and subfamily sequence motifs suggests a mechanism for functional specificity in the Ras superfamily. *J. Mol. Biol.*, **301**(4), 1077–1087.