The Journal of Open Source Software

Pynteny: a Python package to perform synteny-aware, profile HMM-based searches in sequence databases

Semidán Robaina-Estévez [1] **and José M. González** [1]

1 Department of Microbiology. University of La Laguna. Spain.

DOI: 10.21105/joss.05289

Software

- Review
- Repository
- Archive

Editor: Kevin M. Moerman

Reviewers:

- @Kevin-Mattheus-Moerman

Submitted: 20 March 2023
Published: 22 March 2023

Summary

With a growing number of available sequence data, automated function annotation of sequences has become a key subfield of Bioinformatics. In most cases, annotation methods rely on sequence similarity to peptides with known functions to assign functional labels. This approach assumes that similarity implies homology, i.e., shared ancestry. Sequence similarity is most commonly assessed by either alignment-based methods, such as BLAST (Boratyn et al., 2019), or sequence profile-based methods, such as HMMER3 (Mistry et al., 2013). In the first case, query sequences are aligned and compared to a reference sequence database. For the latter, however, query sequences are compared to a profile Hidden Markov Model (HMM), a probabilistic model of the sequence space which is obtained from a collection of representative sequences with the same annotated function. Therefore, profile-based methods are particularly well-suited when query sequences are not sufficiently represented in reference databases, as it facilitates the search of distant homologs due to the sequence variability encoded in the profile HMM (Eddy, 2004; Johnson et al., 2010).

While function is generally conserved among sequence orthologs, i.e., homologs that are the result of a speciation event, this is not the general case of paralogs, that is, homologs that are the result of a gene duplication event, which typically undergo functional diversification. Due to the existence of paralogs, it is impossible to assess orthology solely based on sequence similarity, and additional sources of information, such as phylogenetics and genomic context are necessary to resolve paralogous from orthologous sequences. The consideration of genomic context, such as *synteny*—the physical co-location of genes within the same chromosome across different species—during function annotation is particularly useful in prokaryotes, where genes tend to cluster together into operons and gene organizations above operons. In these cases, syntenic information can reduce annotation uncertainty by providing additional, co-localization constraints to the homology search. Therefore, constraining profile-based searches with syntenic information could markedly benefit annotation pipelines of prokaryotic sequences, particularly those originating in metagenomic samples, which typically are poorly represented in reference databases.

Statement of need

Here we introduce `Pynteny`, a Python tool designed to conduct synteny-aware, profile HMM searches in prokaryotic sequence databases. `Pynteny` facilitates querying sequence databases with arrangements of profile HMMs that reflect a target syntenic block. To this end, it enables encoding positional information, such as gene order, maximum in-between gene distances, and strand specificity, into the search query.

For instance, consider the following syntenic block:

Figure 1: Example of a syntenic structure.

The syntenic block shown above is composed of four genes, `A`, `B`, `C`, and `D`. Genes A-C locate consecutively in the positive strand and are followed by three untargeted genes and by gene D, which is located in the negative strand. `Pynteny` allows searching for the syntenic block above with the following query string:

$$> HMM_A\ 0\ > (HMM_{B1}|HMM_{B2})\ 0\ > HMM_C\ 3\ < HMM_D,$$

where HMM_A represents the name of the HMM modeling gene A, each integer represents the maximum number of untargeted genes between consecutive HMMs, and < and > indicate the strand in which to search for the HMM pattern, antisense and sense, respectively. Alternative HMMs can be used for a single gene, as shown in the HMM group $(HMM_{B1}|HMM_{B2})$ above, in which case the search will be performed with both HMMs. Additionally, gene symbols can be used directly in the query string when the PGAP HMM database (Li et al., 2020) is employed, which is the default database used by `Pynteny`.

`Pynteny` was designed to be used by researchers working with large, unannotated sequence databases, such as those typically encountered in metagenomic analyses. It can be accessed through a command line interface or easily integrated into pipelines as a Python package. It can directly handle assembled nucleotide sequence data, however, it also accepts annotated genomes in GenBank format as input data. In both cases, the package provides all the necessary functionality to preprocess the sequence database.

`Pynteny` relies on Prodigal (Hyatt et al., 2010) to translate and add positional tags to individual genes, and on HMMER3, (Mistry et al., 2013) to search sequence databases for homologs through profile HMMs. Usage information as well as examples in the form of Jupyter Notebooks for both the command line interface and the Python package are available in the documentation.

State of the field

Several existing tools are dedicated to the exploration, analysis, and visualization of synteny blocks among genomes. In these tools, users typically input several annotated genomes and obtain a collection of syntenic relations of shared gene sets among the genomes. Examples of these tools are MCScan (Tang et al., 2008) and MCScanX (Wang et al., 2012), Clinker (Gilchrist & Chooi, 2021), pyGenomeViz (Shimoyama, 2022), genePlotR (Guy et al., 2010), gggenomes (Hackl & Ankenbrand, 2022), GENESPACE (Lovell et al., 2022) and Mology (Ahrens et al., 2021). These tools are excellent resources for the identification and analysis of syntenic relations among genomes, and they are functionally complementary to `Pynteny`. Specifically, rather than exploring syntenic blocks within annotated genomes, `Pynteny`'s objective is to search for specific syntenic structures within unannotated (assembled) sequence data, as well as to leverage syntenic information to reduce uncertainty due to paralogs during function annotation. Therefore, `Pynteny` requires a previous identification of conserved syntenic structures, which can be obtained from existing tools such as the ones previously indicated.

Availability

`Pynteny` is available as a Python package for Linux and MacOS under the Apache 2.0 license and hosted on bioconda, a Docker image is also available. Source code can be found on GitHub, and features a GitHub CodeSpace environment and a Continuous Integration workflow

to run integration tests on changes. Documentation is hosted on GitHub pages and is built for each new release.

Acknowledgements

We acknowledge constructive feedback from Pynteny users as well as editors and reviewers Alex Batisse, C. Thoben, David Nicholson, Ariane Sasso, and Leah Wasser at PyOpenSci who have tremendously helped to improve the package. This study was funded by project PID2019-110011RB-C32 (Spanish Ministry of Science and Innovation, Spanish State Research Agency, doi: 10.13039/501100011033).

References

Ahrens, J. B., Wade, K. J., & Pollock, D. D. (2021). A fast, general synteny detection engine. *bioRxiv*. https://doi.org/10.1101/2021.06.03.446950

Boratyn, G. M., Thierry-Mieg, J., Thierry-Mieg, D., Busby, B., & Madden, T. L. (2019). Magic-BLAST, an accurate RNA-seq aligner for long and short reads. *BMC Bioinformatics*, *20*(1), 405. https://doi.org/10.1186/s12859-019-2996-x

Eddy, S. R. (2004). What is a hidden Markov model? *Nature Biotechnology*, *22*(10), 1315–1316. https://doi.org/10.1038/nbt1004-1315

Gilchrist, C. L. M., & Chooi, Y.-H. (2021). Clinker & clustermap.js: Automatic generation of gene cluster comparison figures. *Bioinformatics*, *37*(16), 2473–2475. https://doi.org/10.1093/bioinformatics/btab007

Guy, L., Roat Kultima, J., & Andersson, S. G. E. (2010). genoPlotR: comparative gene and genome visualization in R. *Bioinformatics*, *26*(18), 2334–2335. https://doi.org/10.1093/bioinformatics/btq413

Hackl, T., & Ankenbrand, M. J. (2022). *Gggenomes: A grammar of graphics for comparative genomics*. https://github.com/thackl/gggenomes

Hyatt, D., Chen, G.-L., LoCascio, P. F., Land, M. L., Larimer, F. W., & Hauser, L. J. (2010). Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, *11*(1), 119. https://doi.org/10.1186/1471-2105-11-119

Johnson, L. S., Eddy, S. R., & Portugaly, E. (2010). Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics*, *11*(1), 431. https://doi.org/10.1186/1471-2105-11-431

Li, W., O'Neill, K. R., Haft, D. H., DiCuccio, M., Chetvernin, V., Badretdin, A., Coulouris, G., Chitsaz, F., Derbyshire, M. K., Durkin, A. S., Gonzales, N. R., Gwadz, M., Lanczycki, C. J., Song, J. S., Thanki, N., Wang, J., Yamashita, R. A., Yang, M., Zheng, C., … Thibaud-Nissen, F. (2020). RefSeq: expanding the Prokaryotic Genome Annotation Pipeline reach with protein family model curation. *Nucleic Acids Research*, *49*(D1), D1020–D1028. https://doi.org/10.1093/nar/gkaa1105

Lovell, J. T., Sreedasyam, A., Schranz, M. E., Wilson, M., Carlson, J. W., Harkess, A., Emms, D., Goodstein, D. M., & Schmutz, J. (2022). GENESPACE tracks regions of interest and gene copy number variation across multiple genomes. *eLife*, *11*, e78526. https://doi.org/10.7554/eLife.78526

Mistry, J., Finn, R. D., Eddy, S. R., Bateman, A., & Punta, M. (2013). Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Research*, *41*(12), e121–e121. https://doi.org/10.1093/nar/gkt263

Shimoyama, Y. (2022). *pyGenomeViz: A genome visualization python package for comparative genomics*. https://github.com/moshi4/pyGenomeViz

Tang, H., Bowers, J. E., Wang, X., Ming, R., Alam, M., & Paterson, A. H. (2008). Synteny and Collinearity in Plant Genomes. *Science*, *320*(5875), 486–488. https://doi.org/10.1126/science.1153917

Wang, Y., Tang, H., DeBarry, J. D., Tan, X., Li, J., Wang, X., Lee, T., Jin, H., Marler, B., Guo, H., Kissinger, J. C., & Paterson, A. H. (2012). MCScanX: A toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Research*, *40*(7), e49. https://doi.org/10.1093/nar/gkr1293