META-CLADE : a new tool to identify domains and functionally annotate metagenomic and metatranscriptomic sequences.

Ari Ugarte, Juliana Bernardes, Alessandra Carbone. Laboratory of Computational and Quantitative Biology (LCQB) UMR 7238 CNRS - UniversitéPierre et Marie Curie, 75006, Paris, France.

The improvements of next-generation sequencing have allowed researchers to study the genomic diversity in microbial communities. The increased complexity of metagenomics data poses computational challenges in assembling, annotating, and classifying genomic fragments from multiple organisms. Domain identification provides insights of the biological function of a protein. Hence, domain annotation is a crucial step to identify and quantify the genes in a microbial community that are known and those that are completely new. Traditional protein annotation methods describe known domains with probabilistic models representing the consensus among homologous domain sequences. When relevant signals become too weak to be identified by consensus, attempts for annotation fails. CLADE [1], a new method for protein domain identification which achieves highly accurate predictions for single genomes compared to HMMER methodology [2] is based on the observation that many structural and functional protein constraints are not globally conserved through all species but might be locally conserved in separate clades. CLADE uses an extension of the probabilistic model library in order to characterize local models to improve signal detection. CLADE has been used to develop META-CLADE [3], a new protein domain annotation tool for metagenomics and metatranscriptomics. In order to evaluate META-CLADE performance, we simulated a dataset containing 500,000 reads of Roche's 454 FLX titanium sequencer. We built this data set from 40 marine bacterial and archaeal complete genome sequences assuming equal abundance. Genes predicted by FragGeneScan [4] in simulated reads were translated to proteins and annotated with META-CLADE and HMMER using Pfam27 [5] domain database. META-CLADE identifies substantially more domains than HMMER in simulated reads (~30% more detected domains). Besides the improvement in domain recognition, META-CLADE agrees with 99,5% of HMMER domain predictions and reinforces the signal of agreed domains. To prove that this new method is suitable for real data, it was applied to 5 data sets collected from 5 different ocean stations containing unicellular marine eukaryotic metatranscriptomic sequences [6]. META-CLADE outperforms HMMER methodology in domain recognition, and signal detection in agreed domains for all data sets. Domains identified by eachmethods were mapped for functional annotation using Pfam2GO [7] and a list of GO Terms [8,9] for each sample was obtained. META-CLADE allows extending the list of significant GO Terms. Moreover, it permits to have a better resolution of significant GO Terms and highlights the functional characteristics of each sample. In conclusion, our results show that META-CLADE is suitable not only for domain recognition but also to improve functional annotation in metagenomicsandmetatranscriptomics studies [10].

[1]High performance domain identification in proteinsreachedwith the agreement of many profiles and domainco-occurrence. J. Bernardes, G. Zaverucha, C. Vaquero, A. Carbone. Submitted. (2015)

[2] HMMER web server: interactive sequence similarity searching. R.D. Finn, J. Clements, S.R. Eddy. Nucleic Acids Research (2011) Web Server Issue 39:W29-W37.

[3] Meta-clade: a highlyprecise annotation method for metagenomicsamples, A.Ugarte, J.Bernardes, A. Carbone, in preparation. (2015)

[4] FragGeneScan: Predicting Genes in Short and Error-prone Reads. Mina Rho, Haixu Tang, and Yuzhen Ye. Nucleic Acids Research (2010)

[5] The Pfam protein families database. R.D. Finn, A. Bateman, J. Clements, P. Coggill, R.Y. Eberhardt, S.R. Eddy, A. Heger, K. Hetherington, L. Holm, J. Mistry, E.L.L. Sonnhammer, J. Tate, M. Punta.Nucleic Acids Research (2014) Database Issue 42:D222-D230

[6] The impact of temperature on marine phytoplankton resource allocation and metabolism.Toseland A., Daines S. J., Clark J. R., Kirkham A., Strauss J., Uhlig C., Lenton T. M., Valentin K., Pearson G. A., Moulton V., Mock T. (2013). Nature Climate Change 3, 979–984

[7] Pfam2GO.Mitchell et al. (2015) Nucleic Acids Research.43 :D213-D221

[8] Ashburner et al. Gene ontology: tool for the unification of biology (2000) Nat Genet 25(1):25-9.

[9] The Gene Ontology Consortium. Gene Ontology Consortium: going forward. (2015) Nucl Acids Res 43 Database

issue D1049–D1056.
[10] A new approach to the functional annotation of metagenomicsamples,A.Ugarte, T.Mock, A.Falciatore, A.Carbone, in preparation(2015).