BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 6 2013, pages 790–791
doi:10.1093/bioinformatics/bts721

Genome analysis

Advance Access publication January 9, 2013

Fragment recruitment on metabolic pathways: comparative metabolic profiling of metagenomes and metatranscriptomes

Dhwani K. Desai[1,2,*,†], Harald Schunck[1,†], Johannes W. Löser[1,3] and Julie LaRoche[1,2]

[1]Helmholtz Centre for Ocean Research Kiel (GEOMAR), Düsternbrooker Weg 20, Kiel 24105, Germany, [2]Department of Biology, Dalhousie University, 1355 Oxford Street, Halifax B3H3Y8, Canada and [3]Institute of Computer Science, Christian-Albrechts University Kiel, Ludewig-Meyn Strasse 4, Kiel 24118, Germany

Associate Editor: Michael Brudno

ABSTRACT

Motivation: The sheer scale of the metagenomic and metatranscriptomic datasets that are now available warrants the development of automated protocols for organizing, annotating and comparing the samples in terms of their metabolic profiles. We describe a user-friendly java program FROMP (Fragment Recruitment on Metabolic Pathways) for mapping and visualizing enzyme annotations onto the Kyoto Encyclopedia of Genes and Genomes (KEGG) metabolic pathways or custom-made pathways and comparing the samples in terms of their Pathway Completeness Scores, their relative Activity Scores or enzyme enrichment odds ratios. This program along with our fully configurable PERL-based annotation organization pipeline Meta2Pro (METAbolic PROfiling of META-omic data) offers a quick and accurate standalone solution for metabolic profiling of environmental samples or cultures from different treatments. Apart from pictorial comparisons, FROMP can also generate score matrices for multiple meta-omics samples, which can be used directly by other statistical programs.

Availability: The source code and documentation for FROMP can be downloaded from https://sites.google.com/site/dhwanidesai/home/software along with the Meta2Pro collection of PERL scripts. Supplementary data are available at https://sites.google.com/site/dhwanidesai/home/fromp_suppl.

Contact: Dhwani.Desai@Dal.Ca

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

Received on May 18, 2012; revised on November 7, 2012; accepted on December 20, 2012

1 INTRODUCTION

The rapidly accumulating environmental meta-omic (metagenomic and metatranscriptomic) projects resulting from high-throughput sequencing techniques warrant the development of new protocols that can provide a quick overview of the microbial metabolic potential or activity. There has been some effort towards management of such data (Sun *et al.*, 2011), its taxonomic and metabolic profiling (Arumugam *et al.*, 2010; Huson *et al.*, 2007; Meyer *et al.*, 2008; Yamada *et al.*, 2011), visualization of metabolic pathways and statistical analyses of community differences (Parks and Beiko, 2010). In most cases, the tools are web-based and the primary method for annotation is BLAST (Altschul *et al.*, 1990). We describe here a standalone set of tools to get a rapid and accurate overview of the metabolic functions of the resident microbial community. The enzyme identification component of this pipeline, based on the ModEnzA Enzyme Commission (EC) numbers (Desai *et al.*, 2011) and Pfam (Punta *et al.*, 2012) profile hidden Markov models (HMMs), provides a quick and accurate EC number identification. The standout feature is the FROMP (Fragment Recruitment on Metabolic Pathways) pathway mapping and comparative visualization tool, which maps EC numbers and Pfam annotations onto the Kyoto Encyclopedia of Genes and Genomes (KEGG) reference (Kanehisa *et al.*, 2012) or custom-made metabolic pathways. Multiple meta-omic samples can be compared with each other based on a Pathway Completeness Score modified from Inskeep *et al.* (2010), a Pathway Activity Score or an odds ratio for enzyme enrichment (Gill *et al.*, 2006). The increased accuracy of HMM-based annotation and an ability to compare multiple meta-omic samples at once are attributes that improve on currently available metabolic profiling tools such as Megan (Huson *et al.*, 2007) and MG-RAST (Meyer *et al.*, 2008).

2 METHODS AND FEATURES

The java program FROMP is a part of the Meta2Pro (METAbolic PROfiling of META-omic data) pipeline (Supplementary Fig. S1). It maps the EC numbers from ModEnzA directly onto the KEGG pathways or user-defined custom-made pathways. The Pfam hits are first mapped to the corresponding Gene Ontology IDs (Ashburner *et al.*, 2000) (using the conversion files pfam2go, kegg2go and ec2go downloaded from http://www.geneontology.org/external2go/), which are then mapped to KEGG reaction IDs or EC numbers.

Pathway Completeness Score: We have modified the weighing scheme for EC numbers described in (Inskeep *et al.*, 2010) by adding a term for the presence of continuous unbranched chains of reactions.

For each EC *i*, the weight

$$W_i = \left[\left(N_{(T,i)}/N_{(U,i)}\right)/N_{(P,i)}\right] * \sqrt{L_{(UBC,r)}}$$

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

where $N_{(T,i)}$ is the total number of ECs in all pathways that have EC i, $N_{(U,i)}$ is the number of unique ECs in all the pathways that have EC i and $N_{(P,i)}$ is the total number of pathways where EC i is present and $L_{(UBC,r)}$ is the total edge-length of the unbranched chain containing EC i in the reference pathway.

The pathway completeness score for a pathway p is then

$$C_p = \left(\frac{\sum_{i \in EC_p} W_i * I_i * \sqrt{L_{(UBC,s)}}}{\sum_{i \in EC_p} W_i} \right)$$

where W_i is the specificity weight of each EC i in pathway p, and I_i is 1 if the EC number is detected in the sample and $L_{(UBC,s)}$ is the edge-length of the unbranched chain containing EC i in the sample.

Odds ratio for gene enrichment: As described in Gill *et al.* (2006), if A and C are the occurrence counts of a given EC in sample i and all other comparison samples j, respectively, and B and D are occurrence counts of all other ECs in sample i and comparison samples j, respectively, then the odds ratio for the given EC in sample i is (A/B)/(C/D).

Pathway Activity Score: This is simply the sum of counts for the ECs in a given pathway multiplied by the EC weight.

Equalization of sequencing effort: To remove bias introduced by differences in sequencing effort, the user can equalize unequal sample sizes to the smallest sample by randomly selecting equal numbers from the other samples.

Custom-designed pathways: In addition to the KEGG pathways, the users can design their own pathways in the Pathway Designer. Chemical species and EC numbers can be placed on a grid and linked with lines. The customized pathways can then be added to any project in FROMP.

Input: Apart from reading the output of the *hmmscan* program (Eddy, 1998), FROMP can also read in tab- or comma-separated list of EC numbers and Pfam accession numbers (one column), ECs and Pfams with counts (two column) and ECs and Pfams with counts and sequence IDs (three column) of the meta-omic sequences. It also accepts a matrix file of EC counts, with the samples arranged in columns and the EC numbers in the rows.

Output: The comparative recruitment of various samples on the reference pathways can be exported as PNG image files. The various score matrices (including the EC count matrix) for the samples and the sequence IDs of the fragments mapping onto each EC or pathway can also be exported as text files.

3 COMPARATIVE ANALYSIS OF METATRANSCRIPTOMES FROM THE OXYGEN MINIMUM ZONE OFF PERU

Three metatranscriptomic samples from a depth profile (oxic, oxycline and anoxic), collected from one station in the Peruvian oxygen minimum zone, were analysed with FROMP and mapped onto a custom-designed pathway that included biological reactions that are thought to be carried out in oxygen-depleted environments (Supplementary Figs S2–S4). The oxic sample metabolism was dominated by sequences affiliated to oxic respiration (cytochrome-c oxidase, EC 1.9.3.1). It also had elevated levels of nitrite reduction transcripts (EC 1.7.2.1), while sequences from the oxycline and the anoxic samples exclusively mapped onto nitrate reductases (EC 1.7.99.4). In the sulfur cycle, sequences similar to hydrogen sulfite reductase (EC 1.8.99.3) mostly originated from the anoxic depth, while the sulfate adenylyltransferase (EC 2.7.7.4) was most abundant in the oxic surface. These visual observations are also supported by the odds ratio or enrichment factors calculated for the ECs in these samples (Supplementary Fig. S5).

4 CONCLUSION

We present here a set of tools for accurate standalone metabolic profiling of meta-omic data. The java program FROMP and the Meta2Pro collection of PERL scripts along with the relevant documentation are available from https://sites.google.com/site/dhwanidesai/home/software.

Funding: This work is funded by the WGL-PAKT project 'REAL' (Leibniz Association) and is a contribution of the Collaborative Research Centre 754 'Climate—Biogeochemistry Interactions in the Tropical Oceans' (www.sfb754.de), which is supported by the German Research Association.

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Arumugam,M. *et al.* (2010) SmashCommunity: a metagenomic annotation and analysis tool. *Bioinformatics*, **26**, 2977–2978.

Ashburner,M. *et al.* (2000) Gene Ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.

Desai,D.K. *et al.* (2011) ModEnzA: accurate identification of metabolic enzymes using function specific profile HMMs with optimized discrimination threshold and modified emission probabilities. *Adv. Bioinformatics*, **2011**, Article ID 743782.

Eddy,S. (2010) HMMER User's Guide: Biological sequence analysis using profile hidden Markov models. http://hmmer.janelia.org (5 February 2013, date last accessed).

Gill,S.R. *et al.* (2006) Metagenomic analysis of the human distal gut microbiome. *Science*, **312**, 1355–1359.

Huson,D.H. *et al.* (2007) MEGAN analysis of metagenomic data. *Genome Res.*, **17**, 377–386.

Inskeep,W.P. *et al.* (2010) Metagenomes from high-temperature chemotrophic systems reveal geochemical controls on microbial community structure and function. *Plos One*, **5**, e9773.

Kanehisa,M. *et al.* (2012) KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.*, **40**, D109–D114.

Meyer,F. *et al.* (2008) The metagenomics RAST server—a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, **9**, 386.

Parks,D.H. and Beiko,R.G. (2010) Identifying biologically relevant differences between metagenomic communities. *Bioinformatics*, **26**, 715–721.

Punta,M. *et al.* (2012) The Pfam protein families database. *Nucleic Acids Res.*, **40**, D290–D301.

Sun,S. *et al.* (2011) Community cyberinfrastructure for Advanced Microbial Ecology Research and Analysis: the CAMERA resource. *Nucleic Acids Res.*, **39**, D546–D551.

Yamada,T. *et al.* (2011) iPath2.0: interactive pathway explorer. *Nucleic Acids Res*, **39**, W412–W415.