BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 13 2014, pages 1926–1927
doi:10.1093/bioinformatics/btu130

Sequence analysis Advance Access publication March 10, 2014

The Bio-Community Perl toolkit for microbial ecology

Florent E. Angly[1,*], Christopher J. Fields[2] and Gene W. Tyson[1]

[1]Australian Centre for Ecogenomics, School of Chemistry and Molecular Biosciences, Level 5, Molecular Biosciences Building (76), The University of Queensland, Brisbane St Lucia, QLD 4072, Australia and [2]HPCBio, Carver Biotechnology Center, Institute for Genomic Biology, 1206 West Gregory Drive | MC-195, Urbana, IL 61801, USA

Associate Editor: John Hancock

ABSTRACT

Summary: The development of bioinformatic solutions for microbial ecology in Perl is limited by the lack of modules to represent and manipulate microbial community profiles from amplicon and meta-omics studies. Here we introduce Bio-Community, an open-source, collaborative toolkit that extends BioPerl. Bio-Community interfaces with commonly used programs using various file formats, including BIOM, and provides operations such as rarefaction and taxonomic summaries. Bio-Community will help bioinformaticians to quickly piece together custom analysis pipelines and develop novel software.

Availability an implementation: Bio-Community is cross-platform Perl code available from http://search.cpan.org/dist/Bio-Community under the Perl license. A readme file describes software installation and how to contribute.

Contact: f.angly@uq.edu.au

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

Received on November 28, 2013; revised on January 29, 2014; accepted on March 2, 2014

1 INTRODUCTION

Sequencing is common in most fields of biological research, and the throughput of modern platforms is orders of magnitudes higher than traditional Sanger sequencing (Metzker, 2010). The BioPerl bioinformatic toolkit (Stajich *et al*., 2002) has attracted a large community of users and developers and has become critical in many sequencing projects by allowing quick code development and interaction between programs using incompatible file formats. In microbial ecology, sequencing is used routinely for 16S rRNA gene amplicon surveys (Tringe and Hugenholtz, 2008), metagenomics (Handelsman, 2004) and metatranscriptomics (Frias-Lopez *et al*., 2008). Because most microorganisms remain uncultivated (Rappé and Giovannoni, 2003), culture-independent molecular surveys are essential for the characterization of environmental microbial communities. However, they require large computational resources, novel bioinformatic tools and elaborate pipelines. Many tools have been developed to analyze the resulting sequence data. For example, libraries written in Python (Knight *et al*., 2007) and R (Dixon, 2003; Kembel *et al*., 2010) provide blocks for building bioinformatic software. QIIME (Caporaso *et al*., 2010) and mothur (Schloss *et al*., 2009) are dedicated packages with scripts to build complete analysis pipelines, but they use incompatible file formats. Here, we introduce Bio-Community, a set of format-agnostic modules and scripts to parse and manipulate taxonomic or functional microbial community profiles.

2 FEATURES

2.1 Object model

Bio-Community is a Perl object-oriented toolkit that extends BioPerl. It is centered around the `Community` object, which contains a group of entities from the same geographic area (Fig. 1).

These entities are `Member` objects, representing individual genomes, genes, taxa or operational taxonomic units from amplicon and meta-omic surveys. `Member` objects store attributes such as an identifier, a taxon or a sequence and can be given weights to account for the fact that there is no one-to-one relationship between a sequencing read and a microbial cell. The relative abundance or abundance rank of a `Member` can be calculated based on this `Member`'s count, weight and the total count in the `Community` (Fig. 2). Similarly, absolute abundance is based on total microbial abundance in the community, quantifiable by epifluorescence microscopy, qPCR or flow cytometry (Rinsoz *et al*., 2008).

2.2 Diversity metrics

Bio-Community quantifies community α, β and γ diversity (Whittaker, 1972) using a range of metrics [reviewed by Magurran (2004)]. The diversity of a single `Community` object, α diversity, is represented by metrics of richness, evenness, dominance and indices (Supplementary Table S1). Several `Community` objects can be grouped into a `Meta` object, representing a metacommunity (Leibold *et al*., 2004). This object provides methods to measure γ diversity, i.e. the collective diversity of its communities, and β diversity, i.e. their dissimilarity. The γ metrics are the same as those available for α diversity, whereas those for β diversity include qualitative and quantitative forms (Supplementary Table S1).

2.3 Data input and output

Community profiles (e.g. a site-by-species table) describe the distribution of members in biological samples. Operations to read and write these files are handled by the `IO` module and are important for exchanging data between programs using different formats. We have implemented parsers for five common file types (Supplementary Table S2), including the BIOM standard (McDonald *et al*., 2012). Examples of these file types are given in the t/data folder of the Bio-Community package. The parsers automatically detect file format based on its content using the

*To whom correspondence should be addressed.

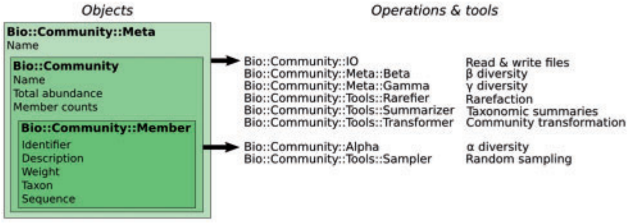


Fig. 1. Main objects, their attributes and operation modules

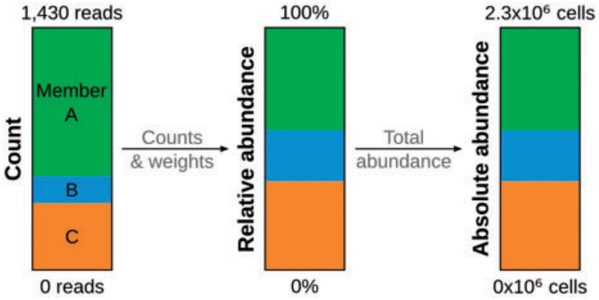
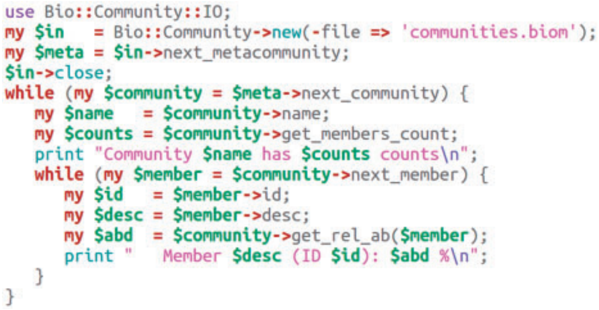


Fig. 2. Relation between abundance types. Relative abundance depends on member counts and weights, whereas absolute abundance is further derived from a total abundance measure



```
use Bio::Community::IO;
my $in   = Bio::Community->new(-file => 'communities.biom');
my $meta = $in->next_metacommunity;
$in->close;
while (my $community = $meta->next_community) {
   my $name   = $community->name;
   my $counts = $community->get_members_count;
   print "Community $name has $counts counts\n";
   while (my $member = $community->next_member) {
      my $id   = $member->id;
      my $desc = $member->desc;
      my $abd  = $community->get_rel_ab($member);
      print "   Member $desc (ID $id): $abd %\n";
   }
}
```

Fig. 3. Vignette illustrating the use of Bio-Community to read a BIOM community profile and report member information

FormatGuesser module, and iteratively record member identifier, taxonomy and abundance.

2.4 Tools

Tool modules can perform operations such as community transformation, rarefaction and taxonomic summaries (Fig. 1). Utility scripts using these modules are available in Bio-Community (Supplementary Table S3). They allow biologists to perform specific operations on community profiles, but they do not form an entire microbial analysis pipeline. These scripts can also be regarded as examples of integration of Bio-Community into bioinformatic scripts (Fig. 3). This integration can also leverage external modules to rapidly develop powerful custom scripts, e.g. Getopt::Euclid for handling command-line arguments, BioPerl modules for reading sequences or running external programs (e.g. BLAST) (Camacho *et al*., 2009) and Statistics::R for using R libraries or visualization capabilities.

3 CONCLUSIONS

Bio-Community provides several file formats to interface with popular programs and will help bioinformaticians quickly construct custom analysis pipelines or novel software for microbial ecology. The integration of relative and absolute abundance with diversity metrics permits holistic microbial studies (Dinsdale *et al*., 2008; Dove *et al*., 2013; Nathani *et al*., 2013), while weights can be added to account for gene copy number (Kembel *et al*., 2012) or genome length (Angly *et al*., 2009; Beszteri *et al*., 2010) bias. We encourage programmers to join the development of Bio-Community at https://github.com/bioperl/Bio-Community and to add support for new file formats, diversity metrics or tools.

Funding: Australian Research Council DE120101213 to FEA and DP1093175 to GWT.

Conflict of interest: none declared.

REFERENCES

Angly,F.E. *et al.* (2009) The GAAS metagenomic tool and its estimations of viral and microbial average genome size in four major biomes. *PLoS Comput. Biol.*, **5**, e1000593.

Beszteri,B. *et al.* (2010) Average genome size: a potential source of bias in comparative metagenomics. *ISME J.*, **4**, 1075–1077.

Camacho,C. *et al.* (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.

Caporaso,J.G. *et al.* (2010) QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods*, **7**, 335–336.

Dinsdale,E.A. *et al.* (2008) Microbial ecology of four coral atolls in the northern Line Islands. *PLoS One*, **3**, e1584.

Dixon,P. (2003) VEGAN, a package of R functions for community ecology. *J. Veg. Sci.*, **14**, 927–930.

Dove,S.G. *et al.* (2013) Future reef decalcification under a business-as-usual CO2 emission scenario. *Proc. Natl Acad. Sci. USA*, **110**, 15342–15347.

Frias-Lopez,J. *et al.* (2008) Microbial community gene expression in ocean surface waters. *Proc. Natl Acad. Sci. USA*, **105**, 3805–3810.

Handelsman,J. (2004) Metagenomics: application of genomics to uncultured microorganisms. *Microbiol. Mol. Biol. Rev.*, **68**, 669–685.

Kembel,S.W. *et al.* (2012) Incorporating 16S gene copy number information improves estimates of microbial diversity and abundance. *PLoS Comput. Biol.*, **8**, e1002743.

Kembel,S.W. *et al.* (2010) Picante: R tools for integrating phylogenies and ecology. *Bioinformatics*, **26**, 1463–1464.

Knight,R. *et al.* (2007) PyCogent: a toolkit for making sense from sequence. *Genome Biol.*, **8**, R171.

Leibold,M.A. *et al.* (2004) The metacommunity concept: a framework for multi-scale community ecology. *Ecol. Lett.*, **7**, 601–613.

Magurran,A.E. (2004) *Measuring biological diversity*. Blackwell, Oxford, United Kingdom.

McDonald,D. *et al.* (2012) The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome. *GigaScience*, **1**, 7.

Metzker,M.L. (2010) Sequencing technologies—the next generation. *Nat. Rev. Genet.*, **11**, 31–46.

Nathani,N.M. *et al.* (2013) Comparative evaluation of rumen metagenome community using qPCR and MG-RAST. *AMB Express*, **3**, 55.

Rappé,M.S. and Giovannoni,S.J. (2003) The uncultured microbial majority. *Annu. Rev. Microbiol.*, **57**, 369–394.

Rinsoz,T. *et al.* (2008) Application of real-time PCR for total airborne bacterial assessment: comparison with epifluorescence microscopy and culture-dependent methods. *Atmos. Environ.*, **42**, 6767–6774.

Schloss,P.D. *et al.* (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Env. Microbiol.*, **75**, 7537–7541.

Stajich,J.E. *et al.* (2002) The Bioperl toolkit: perl modules for the life sciences. *Genome Res.*, **12**, 1611–1618.

Tringe,S.G. and Hugenholtz,P. (2008) A renaissance for the pioneering 16S rRNA gene. *Curr. Opin. Microbiol.*, **11**, 442–446.

Whittaker,R.H. (1972) Evolution and measurement of species diversity. *Taxon*, **21**, 213–251.