BRIEFINGS IN BIOINFORMATICS. VOL 13. NO 6. 743–746
Advance Access published on 8 September 2012
doi:10.1093/bib/bbs020

Metagenomic analysis: the challenge of the data bonanza

Chris I. Hunter, Alex Mitchell, Philip Jones, Craig McAnulla, Sebastien Pesseat, Maxim Scheremetjew and Sarah Hunter

Submitted: 4th November 2011; Received (in revised form): 27th January 2012

Abstract

Several thousand metagenomes have already been sequenced, and this number is set to grow rapidly in the forthcoming years as the uptake of high-throughput sequencing technologies continues. Hand-in-hand with this data bonanza comes the computationally overwhelming task of analysis. Herein, we describe some of the bioinformatic approaches currently used by metagenomics researchers to analyze their data, the issues they face and the steps that could be taken to help overcome these challenges.

Keywords: *metagenomics; next-generation sequencing (NGS); high-throughput sequencing (HTS); functional analysis; environmental bioinformatics*

METAGENOMICS: A BROAD FIELD

The discipline of metagenomics is the study of the genetic material present in a given environment (for a detailed review of the field, see [1, 2]). However, the term 'metagenomics' applies to a very broad range of technical activities, including the collection of environmental samples [3], the extraction of deoxyribonucleic acid/ribonucleic acid (RNA)/protein from those samples, the ever-increasing variety of technologies used for sequencing [4] and the subsequent analysis and interpretation of the resulting data. In this article, we briefly review the current practices in metagenomic sequence analysis and describe potential future developments that may impact on them.

TAXONOMIC ANALYSIS AND METAGENOMICS

The taxonomic classification of living things has long been a central theme in biology; this is particularly true of metagenomics. Amplicon-based taxonomic studies currently dominate the field, and, at the time of writing, more than 80% of the publicly available data sets within the MG-RAST service [5] are taxonomic analyses of the 16S RNA marker gene.

Corresponding author. Chris I. Hunter, EMBL Outstation European Bioinformatics Institute (EBI), Wellcome Trust Genome Campus, Hinxton, CB10 1SD, Cambridge, UK. Tel.: +44 (0) 1223 494 444; Fax: +44 (0)1223 494 468; E-mail: chrish@ebi.ac.uk

Chris Hunter is a curator and bioinformatician for the European Metagenomics portal at the European Bioinformatics Institute in Cambridge, UK. A post-doctorate qualified and technically competent biologist, with over 10 years experience in genomic and genetic research.

Alex Mitchell is a curation coordinator for the InterPro database at the European Bioinformatics Institute in Cambridge, UK. He joined EMBL-EBI in 2011. He has a DPhil in pharmacology and has over 10 years of experience in protein sequence analysis and classification.

Philip Jones is the software development coordinator for the InterPro database at the European Bioinformatics Institute in Cambridge, UK. He joined the EBI in 2004, initially working on PRIDE, the Proteomics Identifications Database. He holds an MSc in Software Engineering from the Open University and has over 10 years experience in bioinformatics software development.

Craig McAnulla is a bioinformatician for the InterPro database at the European Bioinformatics Institute in Cambridge, UK. He has a PhD in Microbiology and several years of experience in running large-scale sequence analyses.

Sebastien Pesseat is a web developer and graphic designer for the European Bioinformatics Institute in Cambridge, UK. He joined the EMBL-EBI in 2010 after 7 years as a consultant for the UN. He holds an MSc in digital image manipulation, web technologies and multimedia from the University of Nice Sophia-Antipolis.

Maxim Scheremetjew is a software developer for the InterPro database at the European Bioinformatics Institute in Cambridge, UK; he has worked there for almost 1 year. Prior to this, he was a bioinformatician and developer at Entelechon GmbH in Germany

Sarah Hunter is the InterPro team leader at EMBL-EBI, a post she has held since 2007. She previously worked in the pharmaceutical and biotech industries and holds an MSc in Bioinformatics from the University of Manchester.

Other phylogenetic classification approaches, such as those offered by Phymm [6] and PhyloPythia [7], are also being used more extensively.

Such analyses are highly valuable, as particular phylogenetic groupings can be associated with important functions, and the diversity of a microbial community is thought to provide an indication of the resilience of the system (i.e. its ability to carry on functioning when conditions change). However, taxonomic studies may not necessarily reflect the complex biological processes that exist in an environment, as microbial genes can move horizontally between unrelated species. Consequently, the same functional gene can be present in a variety of backgrounds. Furthermore, these approaches do not take account of intra-species diversity (where organisms may gain or lose function as they adapt to a specific environment) or situations where organisms may be actively engaged in only a subset of their functional repertoire.

FUNCTIONAL ANALYSIS OF METAGENOMIC SAMPLES

A complementary approach is to analyze the putative functional entities (such as protein coding sequences) within the genomic and/or transcriptomic sequences from an environmental sample. This has become an increasingly realistic proposition with the increasing power and reducing cost of high-throughput sequencing; it is now feasible to sequence a representative proportion of an entire metagenome at reasonable price. The remaining challenge is to process the massive volumes of data produced by such approaches.

Analysis of putative protein coding sequences typically begins with the identification and translation of open reading frames within nucleotide sequences. A minimum size constraint is usually applied, as prediction of function for very short sequences is not reliable. Frequently, pairwise sequence alignment methods, such as BLAST [8], are then used to infer function by searching for similarity to other sequences in a reference database.

One of the original design specifications for BLAST was to provide a tool for fast comparison of sequences. Despite having been developed over 20 years ago, it is still one of the fastest sequence comparison algorithms available. Nevertheless, the sheer volume of sequence data produced during metagenomic studies means that BLAST-based analyses represent significant bottlenecks, which are unlikely to be addressed simply by scaling up computational resources [9].

PROTEIN SIGNATURE-BASED ANALYSES

An alternative protein sequence analysis approach is to use computational models, known as protein signatures, of the type housed in the InterPro [10] consortium of databases, such as Pfam [11], PROSITE [12], PRINTS [13], CATH-Gene3D [14] and TIGRFAMs [15]. These signatures draw on multiple sequence alignments of protein families, domains and functionally important sites. By using such alignments, protein signatures are able to model the (often few) amino acid residues that are conserved in distantly related proteins that are essential for stability and function. Identifying such residues is not possible with pairwise alignment techniques, and consequently protein signatures are usually more sensitive at detecting divergent homologs [16, 17].

Protein signature-based sequence analysis methods offer two further important advantages over their pairwise alignment-based counterparts. As they are built to recognize specific functional entities, such as individual protein families or particular functional domains, matches to signatures are highly accurate predictors of function. This is in contrast to pairwise alignment approaches, where the only significant matches are often to other uncharacterized sequences, meaning that no functional information can be inferred. Furthermore, recent technological advances, such as the development of the HMMER3 algorithm [18], have led to substantial performance increases in a number of protein signature-based analysis techniques, so that they can now offer fast, as well as accurate and sensitive, alternatives to BLAST.

A number of metagenomic analysis pipelines already use protein signatures to predict the functional characteristics of metagenomics data sets. For example, both CAMERA [19] and WebMGA [20] use Pfam and TIGRFAMs alongside BLAST-based approaches for functional sequence analysis. CARMA [21] and CoMet [22] also draw on Pfam for their analyses.

EMBL-EBIs recently launched resource (http://www.ebi.ac.uk/metagenomics) uses InterPro for functional characterization of metagenomic sequences. InterPro combines different types of

protein signature from multiple diverse databases, providing extensive sequence coverage and fine-grained functional analyses. It also provides additional benefits, such as the association of Gene Ontology terms [23] with signatures and inference of potential involvement in biological pathways, further augmenting the annotation of protein sequences. InterPro's utility is expected to grow in the future as investigations into over-represented amino acid sequences in metagenomic data lead to the *in silico* identification of novel protein families and domains, which will in turn be modeled and incorporated into the InterPro Consortium's member databases.

COMPUTATIONAL ADVANCES IN METAGENOMIC ANALYSIS— THE NEED FOR SPEED

Even if protein signature-based methods are used, the time taken to analyze metagenomic data currently far outweighs the length of time taken to produce the sequences in the first place. It is anticipated that new paradigms, such as the use of graphical processing unit (GPU) computing and cloud computing, may help to mitigate this bottleneck in the future.

Promising work has already begun in this area. For example, the developers of Parallel-META [24] have reported a 10–15-fold increase in analysis speeds using GPU over central processing unit. CloVR [25], meanwhile, provides a virtualized machine containing multiple microbial sequence analysis pipelines, including one for metagenomics. It gives the user the option to run their analysis locally or using a commercial or academic cloud.

The use of GPUs and other hardware-based approaches is limited by the specialist programming required to adapt software to run on these architectures. Indeed, the number of general bioinformatics applications that can be run on GPUs is still restricted because of this. Cloud computing facilities should eventually revolutionize the way metagenomics researchers work, potentially allowing even small laboratories access to vast amounts of compute power. However, there remain some drawbacks with this approach, including the relative expense of the compute (running a fully utilized compute farm is cheaper than purchasing time on a commercial cloud [26]) and potential security issues related to transferring data into the cloud environment.

METADATA PROVIDES CONTEXT TO ANALYSIS

Speed is not the only important consideration in metagenomics analysis. Critical to any metagenomic study is the extent and quality of the associated metadata, as this provides context to the experiments and allows meaningful comparisons to be made between studies.

This is exemplified by the Western English Channel study [27], where multiple samples have been meaningfully compared across a large time series. The collection of detailed metadata for each sample allowed the researchers to hypothesize which factors affected the species and functional variety at that site the most.

In recognition of its importance, there has recently been a community-driven shift toward a greater degree of sample contextual metadata being archived with study data, which has been largely facilitated by the Genomic Standards Consortium (GSC) [28]. The mission statement of the GSC is to work toward the implementation of new genomic standards for metadata and methods of capturing and exchanging that metadata. It is immensely valuable to store standards-compliant metadata and the raw sequence data they describe in public repositories, as it allows future reuse and reinterpretation of these data by other scientists. For this reason, researchers are encouraged to submit metadata and raw sequence reads to the INSDC Nucleotide Archives either directly or by the EMBL-EBI metagenomics portal.

CONCLUSION: THE NEED FOR A CONSOLIDATED APPROACH TO METAGENOMICS

Multiple public resources already exist that allow users to view and analyze metagenomics data; however, the field still faces several challenges. It is vital that the metagenomics service providers adopt consistent policy toward metadata, metadata standards and user access to associated raw data, so that metagenomes can be interpreted appropriately by researchers. Despite improvements to functional analysis methods (including the adoption of protein signatures for increased search performance and the optimization of algorithms such as HMMER), the expense of compute remains a barrier to the full realization of metagenomics' potential. It is hoped that collaboration between analysis providers

will lead to better exploitation of new computing paradigms to solve some of these issues.

Key Points

- Metagenomics has historically been dominated by the taxonomic diversity approach, but next generation sequencing is changing this, with more people beginning to investigate the functional potential of an environmental sample.
- Protein signatures are a sensitive way to identify protein families, domains and functionally important sites within protein sequence fragments.
- High-quality contextual data are essential to allow meaningful comparisons to be made between environmental samples.
- The EMBL-EBI metagenomics portal has recently been launched in beta. It facilitates InterPro-driven functional analysis of metagenome sequences and combines this with a metadata-rich archive of metagenomics experiments.

Acknowledgements
The authors thank Penny Hirsch for helpful discussions about this manuscript.

References

1. Wooley JC, Godzik A, Friedberg I. A primer on metagenomics. *PLoS Comput Biol* 2010;**6**(2):e1000667.
2. Gilbert J, Dupont C. Microbial metagenomics: beyond the genome. *Annu Rev Marine Sci* 2011;**3**:347–71.
3. Hildebrandt A, Lacorte S, Barceló D. Sampling of water, soil and sediment to trace organic pollutants at a river-basin scale. *Anal Bioanal Chem* 2006;**386**(4):1075–88.
4. Zhou X, Ren L, Li Y, *et al*. The next-generation sequencing technology: a technology review and future perspective. *Sci China Life Sci* 2010;**53**(1):44–57.
5. Meyer F, Paarmann D, D'Souza M, *et al*. The metagenomics RAST server—a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* 2008;**9**:386.
6. Brady A, Salzberg SL. Phymm and PhymmBL: metagenomic phylogentic classification with interpolated Markov models. *Nat Methods* 2009;**6**(9):673–8.
7. McHardy AC, Martín HG, Tsirigos A, *et al*. Accurate phylogenetic classification of variable-length DNA fragments. *Nat Methods* 2007;**4**(1):63–72.
8. Altschul SF, Gish W, Miller W, *et al*. Basic local alignment search tool. *J Mol Biol* 1990;**215**:403–10.
9. Angiuoli SV, White JR, Matalka M, *et al*. Resources and costs for microbial sequence analysis evaluated using virtual machines and cloud computing. *PLoS One* 2011;**6**(10):e26624.
10. Hunter S, Apweiler R, Attwood TK, *et al*. InterPro: the integrative protein signature database. *Nucleic Acids Res* 2009;**37**:D211–5.
11. Finn RD, Mistry J, Tate J, *et al*. The Pfam protein families database. *Nucleic Acids Res* 2010;**38**:D211–22.
12. Sigrist CJA, Cerutti L, de Castro E, *et al*. PROSITE, a protein domain database for functional characterization and annotation. *Nucleic Acids Res* 2010;**38**:D161–6.
13. Attwood TK, Bradley P, Flower DR, *et al*. PRINTS and its automatic supplement, prePRINTS. *Nucleic Acids Res* 2003;**31**:400–2.
14. Lees J, Yeats C, Redfern O, *et al*. Gene3D: merging structure and function for a thousand genomes. *Nucleic Acids Res* 2010;**38**:D296–300.
15. Selengut JD, Haft DH, Davidsen T, *et al*. TIGRFAMs and genome properties: tools for the assignment of molecular function and biological process in prokaryotic genomes. *Nucleic Acids Res* 2007;**35**:D260–4.
16. Gribskov M, McLachlan AD, Eisenberg D. Profile analysis: detection of distantly related proteins. *Proc Natl Acad Sci USA* 1987;**84**(13):4355–8.
17. Eddy SR. Profile hidden Markov models. *Bioinformatics* 1998;**14**:755–63.
18. Eddy SR. Accelerated profile HMM searches. *PLoS Comput Biol* 2011;**7**(10):e1002195.
19. Sun S, Chen J, Li W, *et al*. Community cyberinfrastructure for advanced microbial ecology research and analysis: the CAMERA resource. *Nucleic Acids Res* 2011;**39**:D546–51.
20. Wu S, Zhu Z, Fu L, *et al*. WebMGA: a customizable web server for fast metagenomic sequence analysis. *BMC Genomics* 2011;**12**:444.
21. Gerlach W, Jünemann S, Tille F, *et al*. WebCARMA: a web application for the functional and taxonomic classification of unassembled metagenomic reads. *BMC Bioinformatics* 2009;**10**:430.
22. Lingner T, Asshauer KP, Schreiber F, Meinicke P. CoMet—a web server for comparative functional profiling of metagenomes. *Nucleic Acids Res* 2011;**39**:W518–23.
23. The Gene Ontology Consortium The Gene Ontology in 2010: extensions and refinements. *Nucleic Acids Res* 2010;**38**:D331–5.
24. Su X, Xu J, Ning K. Parallel-META: efficient metagenomic data analysis based on high-performance computation. *BMC Systems Biology* 2012;**6(Suppl 1)**:S16.
25. Angiuoli SV, Matalka M, Gussman G, *et al*. CloVR: a virtual machine for automated and portable sequence analysis from the desktop using cloud computing. *BMC Bioinformatics* 2011;**12**:356.
26. Wilkening J, Wilke A, Desai N, Meyer F. Using clouds for metagenomics: a case study. In: *IEEE Cluster 2009*. 2009New Orleans, LA.
27. Gilbert J, Field D, Swift P, *et al*. The taxonomic and functional diversity of microbes at a temperate coastal site: a 'Multi-Omic' study of seasonal and diel temporal variation. *PLoS One* 2010;**5**(11):e15545.
28. Field D, Amaral-Zettler L, Cochrane G, *et al*. The genomic standards consortium. *PLoS Biol* 2011;**9**(6):e1001088.