A

A 123 of Metagenomics

Torsten Thomas[1], Jack Gilbert[2] and Folker Meyer[3]
[1]School of Biotechnology and Biomolecular Sciences & Centre for Marine Bio-Innovation, University of New South Wales, Sydney, NSW, Australia
[2]Department of Ecology & Evolution, University of Chicago, Chicago, IL, USA
[3]Institute of Genomic and Systems Biology, Argonne National Laboratory, Argonne, IL, USA

Introduction

Microbial ecology aims to comprehensively describe the diversity and function of microorganisms in the environment. Culturing, microscopy, and chemical or biological assays were not too long ago the main tools in this field. Molecular methods, such as 16S rRNA gene sequencing, were applied to environmental systems in the 1990s and started to uncover a remarkable diversity of organisms (Barns et al. 1994). Soon, the thirst for describing microbial systems was no longer satisfied by the knowledge of the diversity of just one or a few genes. Thus, approaches were developed to describe the total genetic diversity of a given environment (Riesenfeld et al. 2004). One such approach is metagenomics, which involves sequencing the total DNA extracted from environmental samples. Arguably, metagenomics has been the fastest growing field of microbiology in the last few years and has almost become a routine practice. The learning curve in the field has been steep, and many obstacles still need to be overcome to make metagenomics a reliable and standard process. It is timely to reflect on what has been learned over the past few years from metagenome projects and to predict future needs and developments.

This brief primer gives an overview for the current status and practices as well as limitations of metagenomics. We present an introduction to sampling design, DNA extraction, sequencing technology, assembly, annotation, data sharing, and storage.

Sampling Design and DNA Processing

Metagenomic studies of single habitats, for example, acid mine drainage (Tyson et al. 2004), termite hindgut (Warnecke et al. 2007), cow rumen (Hess et al. 2011), and the human gastrointestinal tract (Gill et al. 2006), have provided an insight into the basic diversity and ecology of these environments. Moreover, comparative studies have explored the ecological distribution of genes and the functional adaptations of different microbial communities to specific ecosystems (Tringe et al. 2005; Dinsdale et al. 2008; Delmont et al. 2011). These pioneering studies were predominately designed to develop

K.E. Nelson (ed.), *Genes, Genomes and Metagenomes: Basics, Methods, Databases and Tools*,
DOI 10.1007/978-1-4899-7478-5,

and prove the general metagenomic approach and were often limited by the high cost of sequencing. Hence, desirable scientific methodology, including biological replication, could not be adopted, a situation that precluded appropriate statistical analyses and comparison (Prosser 2010).

The significant reduction, and indeed continuing fall, in sequencing costs (see below) now means that the central tenants of scientific investigation can be adhered to. Rigorous experimental design will help researchers explore the complexity of microbial interactions and will lead to improved catalogs of proteins and genetic elements. Individual ecosystems can now be studied with appropriate cross-sectional and temporal approaches designed to identify the frequency and distribution of variance in community interaction and development (Knight et al. 2012). Such studies should also pay close attention to the collection of comprehensive physical, chemical, and biological data (see below). This will enable scientists to elucidate the emergent properties of even the most complex biological system. This capability will provide the potential to identify drivers at multiple spatial, temporal, taxonomic, phylogenetic, functional, and evolutionary levels and to define the feedback mechanisms that mediate equilibrium.

The frequency and distribution of variance within a microbial ecosystem are basic factors that must be ascertained by rigorous experimental design and analysis. For example, to analyze the microbial community structure from 1 l of seawater in a coastal pelagic ecosystem, one must also ideally define how representative this will be for the ecosystem as a whole and what the bounds of that ecosystem are. Numerous studies of marine systems have shown how community structure can vary between water masses and over time (e.g., Gilbert et al. 2012; Fuhrman 2009; Fuhrman et al. 2006, 2008; Martiny et al. 2006), and metagenomics currently helps further define how community structure varies in these environments (Ottesen et al. 2011; DeLong et al. 2006; Rusch et al. 2007; Gilbert et al. 2010a). In contrast, in soil systems variance in space appears to be far larger than in time (Mackelprang et al. 2011; Barberan et al. 2012; Bergmann et al. 2011; Nemergut et al. 2011; Bates et al. 2011). Considerable work still is needed in order to determine spatial heterogeneity, for example, how representative a 0.1 mg sample of soil is with respect to the larger environment from which it was taken.

The design of a sampling strategy is implicit in the scientific questions asked and the hypotheses tested, and standard rules outside of replication and frequency of observation are hard to define. However, the question of "depth of observation" is prudent to address because researchers now can sequence microbiomes of individual environments with exceptional depth or breadth. By enabling either deep characterization of the taxonomic, phylogenetic, and functional potential of a given ecosystem or a shallow investigation of these elements across hundreds or thousands of samples, current sequencing technology (see below) is changing the way microbial surveys are being performed (Knight et al. 2012).

DNA handling and processing play a major role in exploring microbial communities through metagenomics (see also DNA extraction methods for human studies, "Extraction Methods, DNA" and "Extraction Methods, Variability Encountered in"). Specifically, it is well known that the type of DNA extraction used for a sample will affect the community profile obtained (e.g., Delmont et al. 2012). Therefore, with projects like the Earth Microbiome Project that aim to compare a large number of samples, efforts have been made to standardize DNA extraction protocols for every physical sample. Clearly, no single protocol will be suitable for every sample type (Gilbert 2011, 2010b). For example, a particular extraction protocol might yield only very low DNA concentrations for a particular sample type, making it necessary to explore other protocols in order to improve efficiency. However, differences among DNA extraction protocols may limit comparability of data. Therefore, researchers need to further define in qualitative and quantitative terms how different DNA extraction methodologies affect microbial community structure.

Sequencing Technology and Quality Control

The rapid development of sequencing technologies over the past few years has arguably been one of the driving forces in the field of metagenomics. While shotgun metagenomic studies initially relied on hardware-intensive and costly Sanger sequencing technology (Tyson et al. 2004; Venter et al. 2004) available only to large research institutes, the advent and continuous release of several next-generation sequencing (NGS) platforms has democratized the sequencing market and has given individual laboratories or research teams access to affordable sequencing data. Among the available NGS options, the Roche (Margulies et al. 2005), Illumina (Bentley et al. 2008), Ion Torrent (Rothberg et al. 2011), and SOLiD (Life Technologies) platforms have been applied to metagenomic samples, with the former two being more intensively used than the latter. The features of these sequencing technologies have been extensively reviewed – see, for example, Metzker (2010) and Quail et al. (2012) – and are therefore only briefly summarized here (Table 1).

Roche's platform utilizes pyrosequencing (also often referred to as 454 sequencing because of the name of the company that initially developed the platform) as its underlying molecular principle. Pyrosequencing involves the binding of a primer to a template and the sequential addition of all four nucleoside triphosphates in the presence of a DNA polymerase. If the offered nucleoside triphosphate matches the next position after the primer, then its incorporation results in the release of diphosphate (pyrophosphate, or PPi). PPi production is coupled by an enzymatic reaction involving an ATP sulfurylase and a luciferase to the production of a light signal that is detected through a charge-coupled device. The Ion Torrent sequencing platform uses a related approach; however, here, protons that are released during nucleoside incorporation are detected through semiconductor technology. In both cases, the production of light or charge signals relates to the incorporation of the sequentially offered nucleoside, which can be used to deduce the sequence downstream of the primer. Homopolymer sequences create signals proportional to the number of positions; however, the linearity of this relationship is limited by enzymatic and engineering factors leading to well-investigated insertion and deletion (Indel) sequencing errors (Prabakaran et al. 2011; McElroy et al. 2012).

Illumina sequencing is based on the incorporation and detection of fluorescently labeled nucleoside triphosphates to extend a primer bound to a template. The key feature of the nucleoside triphosphates is a chemically modified 3′ position that does not allow for further chain extension ("terminator"). Thus, the primer gets extended by only one position, whose identity is detected by different fluorescent colors for each of the four nucleosides. Through a chemical reaction, the fluorescent label is then removed, and the 3′ position is converted into a hydroxyl group

A 123 of Metagenomics, Table 1 Next-generation sequencing technologies and their throughput, errors, and application to metagenomics

Machine (manufacturer)	Throughput (per machine run)	Reported errors	Error/metagenomic example references
GLX Titanium (454/Roche)	~1 M reads @ ~500 nt	0.56 % indels; up to 0.12 % substitution	(McElroy et al. 2012; Fan et al. 2012)
HiSeq 2000 (Illumina)	~3 G reads @ 100 nt	~0.001 % indels; up to 0.34 % substitution	(McElroy et al. 2012; Quail et al. 2012; Hess et al. 2011)
Ion Torrent PGM (Life Technologies)	~0.1–5 M reads @ ~200 nt	1.5 % indels	(Loman et al. 2012; Whiteley et al. 2012)
SOLiD (Life Technologies)	~120 M reads @ ~50 nt	Up to 3 %	(Salmela 2010; Zhou et al. 2011; Iverson et al. 2012)

allowing for another round of nucleoside incorporation. The use of a reversible terminator thus allows for a stepwise and detectable extension of the primer that results in the determination of the template sequence. In theory, this process could be repeated to generate very long sequences; in practice, however, misincorporation of nucleosides in the many clonal template strands results in the fluorescent signal getting out of phase, and thus reliable sequencing information is only obtained for about 200 positions (Quail et al. 2012).

SOLiD sequencing utilizes ligation, rather than polymerase-mediated chain extension, to determine the sequence of a template. Primers are extended through the ligation with fluorescently labeled oligonucleotides. The high specificity of the ligase ensures that only oligonucleotides matching the downstream sequence will be incorporated; and by encoding different oligonucleotides with different fluorophores, the sequence can be determined.

It is important to understand the features of the sequencing technology in terms of throughput, read length, and errors (see Table 1), because these will have a significant impact on downstream processing. For example, the relative high frequency of homopolymer errors for the pyrosequencing technology can impact ORF identification (Rho et al. 2010) but might still allow for reliable gene annotation, because of its comparatively long read length (Wommack et al. 2008). Conversely, the short read length of Illumina sequencing might reduce the rate of annotation of unassembled data, but the substantial throughput and data volume generated can facilitate assembly of entire draft genomes from metagenomic data (Hess et al. 2011). These considerations are also particularly relevant with new sequencing technologies coming online. These include single-molecule sequencing using zero-mode waveguide nanostructure arrays (Eid et al. 2009), which promises read lengths beyond 1,000 bp and has been shown to improve the hybrid assemblies of genomes (Koren et al. 2012), as well as nanopore sequencing (Schneider and Dekker 2012), which also promises long read lengths.

One important practical aspect to consider when analyzing raw sequencing data is the quality value assigned to reads. For a long time, the quality assessment provided by the technology vendor was the only available option for data consumers. Recently, however, a vendor-independent error detection and characterization has been described that relies on error estimate-based reads that are accidentally duplicated during the PCR stages (a fact described for Ion Torrent, 454, and Illumina sequencing technologies) (Trimble et al. 2012). Moreover, a significant number of publicly available metagenomic datasets contain sequence adaptors (apparently because quality control is often performed on the level of assembled sequences, not raw reads). Simple statistical analyses with tools such as FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) will rapidly detect most of these adapter contaminations. An important aspect of quality control is therefore that each individual dataset requires error profiling and that relying on general properties of the platform used is not sufficient.

Assembly

Assembly of shotgun sequencing data can in general follow two strategies: the overlap-layout-consensus (OLC) and the de Bruijn graph approach (see also "▶ A De Novo Metagenomic Assembly Program for Shotgun DNA Reads"). These two strategies are employed by a number of different genome assemblers, and this topic has been reviewed recently (Miller et al. 2010). Basically, the OLC assembly involves the pairwise comparison of sequence reads and the ordering of matching pairs into an overlap graph. These overlapping sequences are then merged into a consensus sequence. Assembly with the de Bruijn strategy involves representing each sequence's reads in a graph of all possible k-mers. Two k-mers are connected when the sequence reads have them in sequential, overlapping positions. Thus, all reads of a dataset are represented by the connection within the de Bruijn graph, and assembled contigs are generated by traversing these connections to yield a sequence of k-mers.

The OLC assembly has the advantage that pairwise comparison can be performed to allow for a defined degree of dissimilarity between reads. This can compensate for sequencing errors and allows for the assembly of reads from heterogeneous populations (Tyson et al. 2004). However, memory requirement for pairwise comparisons increases exponentially with the numbers of reads in the dataset; hence, the OLC assembler often cannot deal with large datasets (e.g., Illumina data). Nevertheless, several OLCs, including the Celera Assembler (Miller et al. 2008), Phrap (de la Bastide and McCombie 2007), and Newbler (Roche), have been used to assemble partial or complete draft genomes from metagenomic data; see, for example, Tyson et al. (2004), Liu et al. (2011), and Brown et al. (2012).

In contrast, memory requirements of de Bruijn assemblers are largely determined by the k-mer size chosen to define the graph. Thus, these assemblers have been used successfully with large numbers of short reads. Initially, de Bruijn assemblers designed for clonal genomes, such as Velvet (Zerbino and Birney 2008), SOAP (Li et al. 2008), and ABySS (Simpson et al. 2009), were used to assemble metagenomic data. Because of the heterogeneous nature of microbial populations, however, assemblies often ended up fragmented. One reason is that every positional difference between two reads from the same region of two closely related genomes will create a "bubble" in the graph. Another reason is that sequence errors in low-abundance reads cause terminating branches. Traversing such a highly branched graph leads to large number of contigs. These problems have been partially overcome by modification of existing de Bruijn assemblers such as MetaVelvet (Namiki et al. 2012) or by newly designed de Bruijn-based algorithms such as Meta-IDBA (Peng et al. 2011; see also "Meta-IDBA, overview"). Conceptually, these solutions often include the identification of subgraphs that correspond to individual genomes or the abundance information of k-mers to find an optimal solution path through the graph.

These subdividing approaches are analogous to binning metagenomic reads or contigs, in order to identify groups of sequences that define a specific genome. These bins or even individual sequence reads can also be taxonomically classified by comparison with known reference sequences. Binning and classifying of sequences can be based on phylogeny, similarity, or composition (or combinations thereof), and a large number of algorithms and software is available. For recent comparisons and benchmarking of binning and classification software, please see Bazinet and Cummings (2012) and Droge and McHardy (2012). Obviously, care has to be taken with any automated process, since nonrelated sequences can be combined to produce genomic chimera bins or classes. It is thus advisable that any binning or classification strategy is thoroughly tested through appropriate *in vitro* and *in silico* simulations (Mavromatis et al. 2007; Morgan et al. 2010; McElroy et al. 2012). Also, manual curation of contigs and iterative assembly and mapping can produce improved genomes from metagenomic data (Dutilh et al. 2009). Through such carefully designed strategies and refined processes, nearly complete genomes can be assembled, even for low-abundance organisms from large numbers of short reads (Iverson et al. 2012).

Annotation

Initially, techniques developed for annotating clonal genomes were applied to metagenomic data, and several tools for metagenomic analysis, such as MG-RAST (Meyer et al. 2008) and IMG/M (Markowitz et al. 2008), were derived from existing software suites. For metagenomic projects, the principal challenges lie in the size of the dataset, the heterogeneity of the data, and the fact that sequences are frequently short, even if assembled prior to analysis.

The first step of the analysis (after extensive quality control; see above) involves identification of genes from a DNA sequence. Fundamentally, two approaches exist: the extrinsic approach, which relies on similarity comparison of an unknown sequence to existing databases, and the intrinsic (or *de novo*) approach, which applies statistical analysis of sequence properties, such as frequently used codon usage, to define likely open reading frames (ORFs). For metagenomic data, the extrinsic approach (e.g., running a similarity search with BLASTX) comes at a significant computational cost (Wilkening et al. 2009), rendering it less attractive. *De novo* approaches based on codon or nucleotide k-mer usage are thus more promising for large datasets. *De novo* gene-calling software for microbial genomes are trained on long contigs and assume clonal genomes. For metagenomic datasets this approach is often however unsuitable, because training data is lacking and multiple different codon usage (or k-mer) profiles are present due to the multiple, different genomes present.

However, several software packages have been designed to predict genes for short fragments or even reads (see Trimble et al. 2012 for a review). The most important finding of that review is the effect of errors on gene prediction performance, reducing the reading frame accuracy of most tools to well below 20 % at 3 % sequencing error. Only the software FragGeneScan (Rho et al. 2010; see also FragGeneScan, overview) accounted for the possibility that metagenomic sequences may contain errors, thus allowing it to clearly outperform its competitors.

Once identified, protein-coding genes require functional assignment. Here again, numerous tools and databases exist. Many researchers have found that performing BLAST analysis against the NCBI nonredundant database adds little value to their metagenomic datasets. Preferable are databases that contain high-level groupings of functions, for example, into metabolic pathways as in KEGG (Kanehisa 2002) or into subsystems as in SEED (Overbeek et al. 2005). Using such higher-level groupings allows for the generation of overviews and comparison between samples after statistical normalization.

The time and resources required to perform functional annotations are substantial, but approaches that project multiple results derived from a single sequence analysis into multiple namespaces can minimize these computational costs (Wilke et al. 2012). Numerous tools are also available to predict, for example, short RNAs and/or other genomic features, but these tools are frequently less useful for large metagenomic datasets that exhibit both low sequence quality and short reads.

Several integrations package annotation functionality into a single website. The CAMERA (Seshadri et al. 2007) website, for example, provides users with the ability to run a number of pipelines on metagenomic data. The Joint Genome Institute's IMG/M web service also provides an analysis for assembled metagenomic data, which has been used so far for over 300 metagenomic datasets. The European Bioinformatics Institute provides a service aimed at smaller, typically 454/pyrosequencing-derived metagenomes. The most popular service is the MG-RAST system (Meyer et al. 2008), used for over 50,000 metagenomes with over 140 billion base pairs of data. The system offers comprehensive quality control, tools for comparison of datasets, and data import and export tools to, for example, QIIME (Caporaso et al. 2010) using standard formats such as BIOM (McDonald et al. 2012).

Metadata, Standards, Sharing, and Storage

With over 50,000 metagenomes available, the scientific community has realized that standardized metadata ("data about data") and higher-level classification (e.g., a controlled vocabulary) will increase the usefulness of datasets for novel discoveries (see also ▶ Metagenomics, Metadata, and Meta-analysis). Through the efforts of the Genomic Standards Consortium (GSC) (Field et al. 2011), a set of minimal questionnaires has been developed and accepted by the community (Yilmaz et al. 2010) that allows effective communication of metadata for metagenomic samples of diverse types. While the "required" GSC metadata is purposefully minimal and thus provides only a rough description, several domain-specific environmental packages exist that contain more detailed information.

As the standards evolve to match the needs of the scientific community, the groups developing software and analysis services have begun to rely on the presence of GSC-compliant metadata, effectively turning them into essential data for any metagenome project. Furthermore, comparative analysis of metagenomic datasets is becoming a routine practice, and acquiring metadata for these comparisons has become a requirement for publication in several scientific journals. Since reanalysis of raw sequence reads is often computationally too costly, the sharing of analysis results is also advisable. Currently only the IMG/M and MG-RAST platforms are designed to provide cross-sample comparisons without the need to recompute analysis results. In the MG-RAST system, moreover, users can share data (after providing metadata) with other users or make data publicly available.

Metagenomic datasets continue to grow in size. Indeed the first multi-hundred gigabase pair of metagenomes already exists. Therefore, storage and curation of metagenomic data have become a central theme. The on-disk representation of raw data and analyses has led to massive storage issues for groups attempting meta-analyses. Currently there is no solution for accessing relevant subsets of data (e.g., only reads and analyses pertaining to a specific phylum or a specific species) without downloading the entire dataset. Cloud technologies may in the future provide attractive computational solutions for storage and computing problems. However, specific and metadata-enabled solutions are required for cloud systems to power the community-wide (re-)analysis efforts of the first 50,000 metagenomes.

Conclusion

Metagenomics has truly proven a valuable tool for analyzing microbial communities. Technological advances will continue to drive down the sequencing cost for metagenomic projects and, in fact, the flood of current datasets indicates that funding to obtain sequences is not a major limitation. Major bottlenecks are encountered, however, in terms of storage and computational processing of sequencing data. With community-wide efforts and standardized tools, the impact of these current limitations might be managed in the short term. In the long term, however, large standardized databases will be required (e.g., a MetaGeneBank) to give information access to the entire scientific community. Every metagenomic dataset contains many new and unexpected discoveries, and the efforts of microbiologists worldwide will be needed to ensure that nothing is being missed. As for the data, whether raw or processed, it is just data. Only its biological and ecological interpretation will further our understanding of the complex and wonderful diversity of the microbial world around us.

Government License

References

Barberan A, Bates ST, et al. Using network analysis to explore co-occurrence patterns in soil microbial communities. ISME J. 2012;6(2):343–51.

Barns SM, Fundyga RE, et al. Remarkable archaeal diversity detected in a Yellowstone National Park hot spring environment. Proc Natl Acad Sci U S A. 1994;91(5):1609–13.

Bates ST, Berg-Lyons D, et al. Examining the global distribution of dominant archaeal populations in soil. ISME J. 2011;5(5):908–17.

Bazinet AL, Cummings MP. A comparative evaluation of sequence classification programs. BMC Bioinforma. 2012;13(1):92.

Bentley DR, Balasubramanian S, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature. 2008;456(7218):53–9.

Bergmann GT, Bates ST, et al. The under-recognized dominance of Verrucomicrobia in soil bacterial communities. Soil Biol Biochem. 2011;43(7):1450–5.

Brown MV, Lauro FM, et al. Global biogeography of SAR11 marine bacteria. Mol Syst Biol. 2012;8:595.

Caporaso JG, Kuczynski J, et al. QIIME allows analysis of high-throughput community sequencing data. Nat Methods. 2010;7(5):335–6.

de la Bastide M, McCombie WR. Assembling genomic DNA sequences with PHRAP. Curr Protoc Bioinforma. 2007. Chapter 11: Unit11 14.

Delmont TO, Malandain C, et al. Metagenomic mining for microbiologists. ISME J. 2011;5(12):1837–43.

Delmont TO, Prestat E, et al. Structure, fluctuation and magnitude of a natural grassland soil metagenome. ISME J. 2012;6(9):1677–87.

DeLong EF, Preston CM, et al. Community genomics among stratified microbial assemblages in the ocean's interior. Science. 2006;311(5760):496–503.

Dinsdale EA, Edwards RA, et al. Functional metagenomic profiling of nine biomes. Nature. 2008;452(7187): 629–32.

Droge J, McHardy AC. Taxonomic binning of metagenome samples generated by next-generation sequencing technologies. Brief Bioinform. 2012;13(6):646–55.

Dutilh BE, Huynen MA, et al. Increasing the coverage of a metapopulation consensus genome by iterative read mapping and assembly. Bioinformatics. 2009;25(21): 2878–81.

Eid J, Fehr A, et al. Real-time DNA sequencing from single polymerase molecules. Science. 2009; 323(5910):133–8.

Fan L, Reynolds D, et al. Functional equivalence and evolutionary convergence in complex communities of microbial sponge symbionts. Proc Natl Acad Sci U S A. 2012;109(27):E1878–87.

Field D, Amaral-Zettler L, et al. The genomic standards consortium. PLoS Bio. 2011;9(6):e1001088.

Fuhrman JA. Microbial community structure and its functional implications. Nature. 2009;459(7244):193–9.

Fuhrman JA, Hewson I, et al. Annually reoccurring bacterial communities are predictable from ocean conditions. Proc Natl Acad Sci U S A. 2006; A103(35):13104–9.

Fuhrman JA, Steele JA, et al. A latitudinal diversity gradient in planktonic marine bacteria. Proc Natl Acad Sci U S A. 2008;A105(22):7774–8.

Gilbert JA, Field D, 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. 2010a;5(11):e15545.

Gilbert JA, Meyer F, et al. The earth microbiome project: meeting report of the "1 EMP meeting on sample selection and acquisition at Argonne National Laboratory October 6 2010". Stand Genomic Sci. 2010b;3(3):249–53.

Gilbert JA, Bailey M, et al. The earth microbiome project: the Meeting Report for the 1st International Earth Microbiome Project Conference, Shenzhen, China, June 13th-15th 2010. Stand Genomic Sci. 2011;5(2):243–7.

Gilbert JA, Steele JA, et al. Defining seasonal marine microbial community dynamics. ISME J. 2012;6: 298–308.

Gill SR, Pop M, et al. Metagenomic analysis of the human distal gut microbiome. Science. 2006;312(5778): 1355–9.

Hess M, Sczyrba A, et al. Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. Science. 2011;331(6016):463–7.

Iverson V, Morris RM, et al. Untangling genomes from metagenomes: revealing an uncultured class of marine Euryarchaeota. Science. 2012;335(6068):587–90.

Kanehisa M. The KEGG database. Novartis Found Symp. 2002;247:91–101. discussion 101–103, 119–128, 244–152.

Knight R, Jansson J, et al. Designing better metagenomic surveys: the role of experimental design and metadata capture in making useful metagenomic datasets for ecology and biotechnology. Nat Biotechnol. 2012;30(6):513–2.

Koren S, Schatz MC, et al. Hybrid error correction and *de novo* assembly of single-molecule sequencing reads. Nat Biotechnol. 2012;30(7):693–700.

Li R, Li Y, et al. SOAP: short oligonucleotide alignment program. Bioinformatics. 2008;24(5):713–4.

Liu MY, Kjelleberg S, et al. Functional genomic analysis of an uncultured delta-proteobacterium in the sponge *Cymbastela concentrica*. ISME J. 2011;5(3):427–35.

Loman NJ, Misra RV, et al. Performance comparison of benchtop high-throughput sequencing platforms. Nat Biotechnol. 2012;30(5):434–9.

Mackelprang R, Waldrop MP, et al. Metagenomic analysis of a permafrost microbial community reveals a rapid response to thaw. Nature. 2011;480(7377): 368–71.

Margulies M, Egholm M, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 2005;437(7057):376–80.

Markowitz VM, Ivanova NN, et al. IMG/M: a data management and analysis system for metagenomes. Nucleic Acids Res. 2008;36(Database issue):D534–8.

Martiny JB, Bohannan BJ, et al. Microbial biogeography: putting microorganisms on the map. Nat Rev Microbiol. 2006;4(2):102–12.

Mavromatis K, Ivanova N, et al. Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. Nat Methods. 2007;4(6): 495–500.
McDonald D, Clemente JC, et al. The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome. Gigascience. 2012;1(1):7.
McElroy KE, Luciani F, et al. GemSIM: general, error-model based simulator of next-generation sequencing data. BMC Genomics. 2012;13:74.
Metzker ML. Sequencing technologies – the next generation. Nat Rev Genet. 2010;11(1):31–46.
Meyer F, Paarmann D, et al. The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinforma. 2008;9:386.
Miller JR, Delcher AL, et al. Aggressive assembly of pyrosequencing reads with mates. Bioinformatics. 2008;24(24):2818–24.
Miller JR, Koren S, et al. Assembly algorithms for next-generation sequencing data. Genomics. 2010;95(6):315–27.
Morgan JL, Darling AE, et al. Metagenomic sequencing of an *in vitro*-simulated microbial community. PLoS One. 2010;5(4):e10209.
Namiki T, Hachiya T, et al. MetaVelvet: an extension of Velvet assembler to *de novo* metagenome assembly from short sequence reads. Nucleic Acids Res. 2012;40(20):e155.
Nemergut DR, Costello EK, et al. Global patterns in the biogeography of bacterial taxa. Environ Microbiol. 2011;13(1):135–44.
Ottesen EA, Marin R, et al. Metatranscriptomic analysis of autonomously collected and preserved marine bacterioplankton. ISME J. 2011;5(12):1881–95.
Overbeek R, Begley T, et al. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Res. 2005;33(17):5691–702.
Peng Y, Leung HC, et al. Meta-IDBA: a *de Novo* assembler for metagenomic data. Bioinformatics. 2011;27(13):i94–101.
Prabakaran P, Streaker E, et al. 454 antibody sequencing – error characterization and correction. BMC Res Notes. 2011;4:404.
Prosser JI. Replicate or lie. Environ Microbiol. 2010;12(7):1806–10.
Quail M, Smith ME, et al. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics. 2012;13(1):341.
Rho M, Tang H, et al. FragGeneScan: predicting genes in short and error-prone reads. Nucleic Acids Res. 2010;38(20):e191.
Riesenfeld CS, Schloss PD, et al. Metagenomics: genomic analysis of microbial communities. Annu Rev Genet. 2004;38:525–52.
Rothberg JM, Hinz W, et al. An integrated semiconductor device enabling non-optical genome sequencing. Nature. 2011;475(7356):348–52.
Rusch DB, Halpern AL, et al. The Sorcerer II global ocean sampling expedition: Northwest Atlantic through Eastern Tropical Pacific. PLoS Biol. 2007;5(3):e77.
Schneider GF, Dekker C. DNA sequencing with nanopores. Nat Biotechnol. 2012;30(4):326–8. doi: 10.1038/nbt.2181.
Salmela L. Correction of sequencing errors in a mixed set of reads. Bioinformatics. 2010;26(10):1284–90.
Seshadri R, Kravitz SA, et al. CAMERA: a community resource for metagenomics. PLoS Biol. 2007;5(3):e75.
Simpson JT, Wong K, et al. ABySS: a parallel assembler for short read sequence data. Genome Res. 2009;19(6):1117–23.
Trimble WL, Keegan KP, et al. Short-read reading-frame predictors are not created equal: sequence error causes loss of signal. BMC Bioinforma. 2012;13(1):183.
Tringe SG, von Mering C, et al. Comparative metagenomics of microbial communities. Science. 2005;308(5721):554–7.
Tyson GW, Chapman J, et al. Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature. 2004;428(6978):37–43.
Venter JC, Remington K, et al. Environmental genome shotgun sequencing of the Sargasso Sea. Science. 2004;304(5667):66–74.
Warnecke F, Luginbuhl P, et al. Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. Nature. 2007;450(7169): 560–5.
Whiteley AS, Jenkins S, et al. Microbial 16S rRNA Ion Tag and community metagenome sequencing using the Ion Torrent (PGM) platform. J Microbiol Methods. 2012;91(1):80–8.
Wilke A, Harrison T, et al. The M5nr: a novel non-redundant database containing protein sequences and annotations from multiple sources and associated tools. BMC Bioinforma. 2012;13:141.
Wilkening J, Wilke A, et al. Using clouds for metagenomics: a case study. IEEE Cluster 2009. 2009
Wommack KE, Bhavsar J, et al. Metagenomics: read length matters. Appl Environ Microbiol. 2008;74(5):1453–63.
Yilmaz P, Kottmann R, et al. The “Minimum Information about an ENvironmental Sequence” (MIENS) specification. Nat Biotechnol. 2010. in print.
Zerbino DR, Birney E. Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. Genome Res. 2008;18(5):821–9.
Zhou R, Ling S, et al. Population genetics in nonmodel organisms: II. Natural selection in marginal habitats revealed by deep sequencing on dual platforms. Mol Biol Evol. 2011;28(10):2833–42.

A De Novo Metagenomic Assembly Program for Shotgun DNA Reads

Huaiqiu Zhu
Department of Biomedical Engineering, and Center for Theoretical Biology, Peking University, Beijing, China

Synonyms

MAP: metagenomic assembly program

Definition

Contig: a set of overlapping DNA segments that together represent a consensus region of DNA. Assembly (also genome assembly): the process of taking a large number of short DNA sequencing reads and putting them back together to create contigs from which the DNA originated.

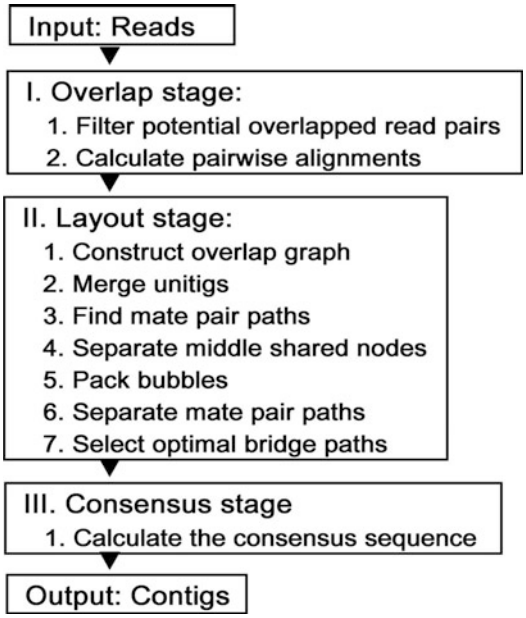
Introduction

MAP (metagenomic assembly program) is a de novo assembler designed to be applicable to shotgun DNA reads (recommended as >200 bp) for metagenome sequencing project (Lai et al. 2012). The program focuses on the metagenomic assembly problem of longer reads produced by, for example, Sanger (typically 700–1,000 bp) and 454 sequencing (typically 200–500 bp). Meanwhile, mate-pair information from both ends of a DNA fragment for a given size (e.g., an insert in a vector plasmid in Sanger sequencing or a mate-pair template in 454 sequencing) in sequencing is introduced, which is commonly available in Sanger sequencing and most of the new sequencing technologies including 454 sequencing.

Although processing of shotgun metagenomic sequence data usually does not have a fixed end point to recover one or more complete genomes as for isolated microbial genomes, the assembly tools, which aim to combine sequence reads into contigs, are still expected to play an important role in sequence processing, due to more valuable genomic content they can provide (Tyson et al. 2004; Venter et al. 2004). In the past decade, a good many assembly algorithms have been proposed to deal with the sequence assembly problem, among of which are the early algorithms targeted to the Sanger sequencing technology, such as Phrap (http://www.phrap.org), Celera (Myers et al. 2000; Miller et al. 2008), and PCAP (Huang et al. 2003), and the up-to-date algorithms targeted to the next-generation technology, such as Velvet (Zerbinor and Birney 2008) and SOAPdenovo (Li et al. 2010). However, these methods are not targeting the metagenome sequencing in spite of the situation that they are still usually employed to undertake assembling of the metagenomic sequencing reads.

Compared to isolated genome assembly problem, the metagenomic assembly problem is more complicated due to two challenges (Kunin et al. 2008): (1) the genomic repeats may originate from either the same genome or the different genomes; therefore, large numbers of mixed short DNA reads belong to many different species (we even know little about the population structure for some environmental samples); and (2) the inhomogeneous coverage distribution and the low abundance of organisms provide limited information to handle repeats. Due to the specific challenges of the metagenomic assembly problem, traditional assembly methods developed for single genome assembly problem usually generate poor quality draft assembly on metagenomic data (Mavromatis et al. 2007). Thus, it is in need to develop highly efficient assembly method specifically for metagenomic data.

Moreover, compared with Sanger and 454 sequencing, the current limitation of shorter reads (<200 bp, typically 25–100 bp) and higher errors by the new sequencing platforms does not allow a significant utility for metagenomic analyses for the difficulty in phylogenetic study or gene function inference. In fact, shorter reads technologies have not been widely used in metagenome sequencing, and meanwhile the sequencing technologies producing longer reads, such as Sanger (usually 700–1,000 bp)



A De Novo Metagenomic Assembly Program for Shotgun DNA Reads, Fig. 1 The flowchart of MAP algorithm

and 454 sequencing (usually 200–500 bp), are still the overwhelming recommendation and thus remain the major source of metagenomic sequence data. Therefore, it is never trivial to continue to emphasize the importance of longer reads to metagenomic analyses, clearly including the reads assembly tool designed specifically.

Algorithm of MAP

MAP designs an improved approach of the classical overlap/layout/consensus (OLC) strategy, in which several special algorithms are incorporated into its stages, to calculate correct contigs by connecting the fragments linked by mate pairs to prevent the false merge of unrelated reads. For the improved OLC strategy, MAP deploys a series of algorithms in three stages as shown in Fig. 1. In the overlap stage, the filter algorithm based on q-gram (Mullikin et al. 2003) is used to obtain the read pairs that are supposed to have the overlaps, and the seed and extend alignment approach, similar to that used by BLAST (Altschul et al. 1990), is employed in the pairwise alignment calculation. While in the consensus stage, a consistency-based consensus algorithm is used (Rausch et al. 2009), which is based on a multi-read alignment algorithm aligning the reads with a consistency-enhanced alignment graph of shared sequence segments identified in advance. The most important innovation of MAP is the layout stage which applies mate-paired information to deal with repeat problem, which is described below.

In the OLC approach of MAP, the overlap graph is used to facilitate the assembly process. Conceptually, reads and overlaps are represented in the graph G by nodes and bidirected edges, respectively. The arrows of both ends of the edge are determined by the way how two reads overlap. Herein, a dovetail path is defined as an acyclic path with each node has only one arrow outward it and one arrow inward it. Thus, a dovetail path can determine a certain contig by means of threading the reads corresponding to the nodes in this path. Thus, the goal of the layout stage is to separate the graph into disconnected dovetail paths. However, since there may be quite many misleading edges in the graph that represent the false overlaps mainly originated from two repetitive DNA regions or similar fragments of different genomes, this goal seems to be a formidable task. To this end, MAP is designed to determine the optimal dovetail paths with the aids of the clues given by mate pairs (Lai et al. 2012).

Compared with other assemblers, several distinct features of MAP algorithm should be pointed out. First, MAP does not refer to any other information such as genome length or sequencing coverage that is often used in the assemblers targeting the isolated genomes, because such information is clearly not applicable to the situation of metagenomic assembly. What is more important is that MAP employs mate-paired information different from other assemblers do. For example, the Celera Assembler (Myers et al. 2000) used mate-paired information in the scaffold construction. The Celera Assembler later developed a new pipeline CABOG, which finds the best overlap graph in the unitigger module (Miller et al. 2008). In this algorithm, mate pairs are used to correct the misassemblies by breaking the unitigs which are found violated with the mate-pair constrains. PCAP (Huang et al. 2003) used mate-paired information to correct contigs and to link contigs into scaffolds. Different from these assemblers, MAP uses mate pairs as a core measure to construct contigs when repeats hamper the assembly. Based on mate-paired information, MAP designs a series of procedures to implement the layout stage.

Performance of MAP

MAP is designed for metagenomic assembly on long reads data with mate pairs, such as Sanger reads (700–1,000 bp) and 454 reads (200–500 bp). MAP method was assessed on simulated data compared with widely used assemblers on long reads data. Specifically, the assessment test results on simulated dataset with 800 bp reads demonstrate that the total assembly performance of MAP can be superior to both Celera and Phrap for typical longer reads by Sanger sequencing, and the results on simulated dataset with 200 bp reads show that MAP has evident advantage over Celera, Newbler (Margulies et al. 2005), and Genovo (Laserson et al. 2011), for typical shorter reads by 454 sequencing (Lai et al. 2012).

Availability

MAP is written in C++ and the source code is freely available under GNU GPL license. The MAP is freely available at http://bioinfo.ctb.pku.edu.cn/MAP/.

References

Altschul SF, Gish W, et al. Basic local alignment search tool. J Mol Biol. 1990;215(3):403–10.
Huang X, Wang J, et al. PCAP: a whole-genome assembly program. Genome Res. 2003;13:2164–70.
Kunin V, Copeland A, et al. A bioinformatician's guide to metagenomics. Microbiol Mol Biol Rev. 2008;72:557–178.
Lai B, Ding R, et al. A de novo metagenomic assembly program for shotgun DNA reads. Bioinformatics. 2012;28(11):1455–62.
Laserson J, Jojic V, et al. Genovo: de novo assembly for metagenomes. J Comput Biol. 2011;18:429–43.
Li R, Zhu H, et al. De novo assembly of human genomes with massively parallel short read sequencing. Genome Res. 2010;20:265–72.
Margulies M, Egholm M, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 2005;437:376–80.
Mavromatis K, Ivanova N, et al. Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. Nat Methods. 2007;4:495–500.
Miller JR, Delcher AL, et al. Aggressive assembly of pyrosequencing reads with mates. Bioinformatics. 2008;24:2818–24.
Mullikin JC, Ning Z, et al. The phusion assembler. Genome Res. 2003;13:81–90.
Myers EW, Sutton GG, et al. A whole-genome assembly of Drosophila. Science. 2000;287:2896–204.
Rausch T, Koren S, et al. A consistency-based consensus algorithm for de novo and reference-guided sequence assembly of short reads. Bioinformatics. 2009;25:1118–24.
Tyson GW, Chapman J, et al. Genomic structure and metabolism through reconstruction of microbial genomes from the environment. Nature. 2004;428:37–43.
Venter JC, Remington K, et al. Environmental genome shotgun sequencing of Sargasso sea. Science. 2004;304:66–74.
Zerbinor DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008;18:821–9.

Ab Initio Gene Identification in Metagenomic Sequences

Shiyuyun Tang[1] and Mark Borodovsky[2]
[1]School of Biology, Biodiversity Research Center, Georgia Institute of Technology, Atlanta, GA, USA
[2]Joint Georgia Tech and Emory Wallace H Coulter Department of Biomedical Engineering, Center for Bioinformatics and Computational Genomics, Atlanta, GA, USA

Synonyms

Statistical or intrinsic methods of gene prediction

Definition

Computational inference of how a metagenomic sequence is divided into protein-coding and non-coding regions based on presence or absence of characteristic oligonucleotide frequency patterns.

Introduction

As of April 2013 sequences of 370 metagenomes were available in databases. On the other hand, Genomes Online Database (www.genomesonline.org) lists 186 complete archaeal and 3,956 complete bacterial genomes; also there are about 15,000 incomplete (draft) prokaryotic genomes. With the average size of a metagenome being 100 times larger than an average prokaryotic genome, the current volume of metagenomic sequences is twice as large as the total sequence in "genomic" data. Therefore, current metagenomes carry a larger wealth of genes than all the prokaryotic genomes, and this gap is growing.

Notably, gene prediction and annotation of gene and protein function is more challenging in metagenomes than in draft or complete genomes. To give a historic perspective, one can compare gene annotation of a metagenome with annotation of the first completely sequenced archaeal genome, *Methanococcus jannaschii* (Bult et al. 1996). All the *M. jannaschii* genes were predicted by the ab initio statistical method (Borodovsky and McIninch 1993) while function of 2/3 of them was a mystery since the translated protein sequences did not show sequence similarity to proteins in databases.

The history repeats itself in metagenomes, since majority of protein-coding regions in a new metagenome may code for proteins that do not show similarity to already known proteins. "Evidence-based" or "similarity-based" methods of gene finding (Kunin et al. 2008) provide gene prediction along with valuable information about function of encoded proteins. Similarity-based gene finders possess high specificity, close to 100 % (Altschul et al. 1997; Badger and Olsen 1999; Frishman et al. 1998; Gish and States 1993). Still, the drawback of similarity-based methods is low sensitivity; they cannot predict novel genes.

The similarity-based methods are less useful for gene prediction in metagenomes that carry many novel genes, while the ab initio gene prediction methods, not depending on presence of homologs in protein databases, are both effective and efficient for annotating genes in metagenomic sequences (Kunin et al. 2008).

Ab Initio Gene Finding

Ab initio gene prediction tools have high sensitivity (above 90 % for the best tools) and high specificity (above 90 % as well). Ab initio gene finders use statistical pattern recognition methods (Wooley et al. 2010). Statistical models such as Markov models, hidden Markov models (HMM), and hidden semi-Markov models (HSMM, also called hidden Markov model with duration) proved to be very useful to model statistical patterns of nucleotide ordering in protein-coding and noncoding regions. Accurate ab initio gene finding in isolated genomes requires ample sequence data for estimation of algorithm parameters (model training).

Contrary to isolated (complete and draft) genomes metagenomic sequences are derived from numerous genomes of heterogeneous microbial communities (microbiomes). A typical metagenomic sequence is short; its genomic context and the phylogenetic origin are rarely known. Gene identification is also affected by sequencing and assembly errors; for example, errors that lead to frameshifts (change of coding frame).

The major challenge for ab initio gene prediction in metagenomic sequences is that the metagenomic sequences are often too short for reliable estimation of parameters of species-specific models of coding and noncoding regions. Special training techniques have to be developed to address the challenging task of parameter estimation (see below). Similarly to gene prediction in isolated genomes, newly predicted genes are immediately translated into proteins and the similarity search is used in an attempt of function annotation.

Gene Finders Currently Available for Metagenomes

Current metagenomic gene-finding tools include FragGeneScan (Rho et al. 2010), Glimmer-MG (Kelley et al. 2012), MetaGene Annotator (Noguchi et al. 2008), MetaGeneMark (Zhu et al. 2010), and Orphelia (Hoff et al. 2009, 2008). Glimmer-MG and MetaGeneMark are extensions of gene finders for complete or draft genomes Glimmer3 (Delcher et al. 2007) and GeneMarkS (Besemer et al. 2001), respectively.

The MetaGeneMark algorithm uses HSMM architecture, originally developed in GeneMarkS (Besemer et al. 2001). The HSMM parameter derivation approach used in MetaGeneMark is to arrive to a large set of parameters (thousands of parameters related to oligonucleotide frequencies) from a small set (nucleotide frequencies determined in a short fragment) using the dependencies between oligonucleotide and nucleotide frequencies that have been formed in evolution. The original idea of this approach (Besemer and Borodovsky 1999) has been developed for small viral genomes before the start of "metagenomic era" (see below for more details).

Glimmer-MG is based on interpolated Markov models or IMMs (Salzberg et al. 1998). Glimmer-MG scores metagenomic sequences and assigns them into clusters; then, the algorithm iteratively estimates the IMM parameters and reassigns sequences to clusters.

FragGeneScan (Rho et al. 2010), an HMM-based gene finder, has an additional ability to predict frameshifts caused by sequencing errors. Transition probabilities between coding frames are determined with respect to the error models of sequencing technologies used to derive the input sequence.

MetaGene Annotator (Noguchi et al. 2008) works in two steps: in the first step the program scores open reading frames (ORFs) with respect to base composition and lengths; in the second step, it connects high-scoring ORFs using dynamic programming.

Machine learning classification algorithms such as support vector machines and neural networks are also used for ab initio gene finding. In order to classify coding or noncoding ORFs, Orphelia (Hoff et al. 2009, 2008) uses an artificial neural network combining multiple features to get ORF's scores.

Parameter Estimation for Metagenomic Gene-Finding Algorithms

Patterns of oligonucleotide frequencies differ in coding and noncoding regions; these patterns are more pronounced when frequencies of longer oligomers are considered. Sequences with specific oligomer frequencies can be modeled by Markov chain models and in the important case of protein-coding sequences by three-periodic Markov chain models (Borodovsky et al. 1986). The number of parameters of a three-periodic Markov chain model increases exponentially with the model order; estimation of parameters of the practically useful fifth order model requires at least several hundred thousand nucleotide long sequence. Use of a shorter training sequence leads to over-fitting and will corrupt gene prediction. If the origin of the metagenomic sequence is known, sequences from the whole parent genome could be used for training. Alternatively, if novel metagenomic sequences from a single species are assembled in sufficiently long contig the model parameters can be estimated by self-training on the contig sequence (Besemer et al. 2001; Kelley et al. 2012). Most frequently, however, metagenomic sequences are short and novel (of the order of a few hundred nucleotides). Therefore, new approach to the model parameter derivation is needed.

A novel approach for constructing parameters and making efficient models for gene prediction in short genomic sequences was proposed back in 1999 (Besemer and Borodovsky 1999). The idea was to use observed trends in the nucleotide frequencies in the three codon positions in genomes with various GC content. Use of these dependencies allows for reconstructing the species-specific codon usage pattern in the whole genome starting from a short fragment of this genome whose length is sufficient to estimate the genome GC content. This approach is based on the assumption of genome compositional uniformity that is largely valid for prokaryotic genomes. It was shown that parameters provided by this approach allow sufficiently accurate gene prediction in short metagenomic sequences. Later on, with more genomes becoming available, this idea was extended (Zhu et al. 2010) to longer oligonucleotides (e.g., hexamers). With GC content of a genome being an independent variable X, it could be shown that frequency of phased K-mers in any of three frames, variable Y, can be approximated by a polynomial of order K. Particularly, the mononucleotide frequencies in three codon positions can be approximated by linear functions. These dependencies indicate that GC content is a major driving factor that determines evolution of genome-wide codon usage pattern (Chen et al. 2004). In MetaGeneMark, the value of GC content determined for a short metagenomic sequence is used as an estimate of GC content of the whole genome the sequence originated from. This value allows immediate reconstruction of frequencies of phased oligonucleotides and, at the next step, parameters of three-periodic Markov chain models of the heuristic model (Zhu et al. 2010).

Interestingly, the heuristic models can also be used for gene prediction in complete genomes or draft genomes. In comparison with the "native" models (models trained on a genome of interest), heuristic models are more sensitive to so-called "atypical" genes. Many atypical genes appear to be horizontally transferred genes with codon frequencies deviating from dominant codon usage pattern of the "host" genome.

Another approach to model parameter estimation is attempting to make a sufficiently large set of training sequences by linking anonymous sequences that appear to be taxonomically close. For example, Glimmer-MG assigns a taxon for a metagenomic sequence by a classification method called Phymm (Brady and Salzberg 2009) and then searches databases for genomes that belong to this taxon. Since such type of training is executed in real time, the running time of gene-finding algorithm may increase significantly in comparison with the algorithm selecting a heuristic model from a set of models precomputed for possible values of GC contents.

Additional Sequence Features Used by Metagenomic Gene Finders

Besides function-specific patterns in oligonucleotide composition, gene identification algorithms can use additional features that help discriminate protein-coding and noncoding regions. Such features include empirical length distributions of coding and noncoding regions, mutual orientation of neighboring coding regions, and sequence patterns related to functional sites such as ribosomal binding sites (RBS). The two-component model of RBS, containing positional frequency matrix as a model of the RBS motif and the length distribution of a "spacer," the sequence between RBS and gene start, carries important additional information for improving accuracy of gene start prediction. In prokaryotic genomes an average spacer length is 5–7 nt. The RBS positional

Ab Initio Gene Identification in Metagenomic Sequences, Table 1 Gene prediction accuracy for five ab initio gene finders. Sn stands for sensitivity and Sp stands for specificity

Programs	Test set	Sequence length (bp)	Sn (%)	Sp (%)	(Sn + Sp)/2 (%)	Publication
Orphelia	Fragments from 12 test species	300	82.1	91.7	86.9	Hoff et al. (2009)
FragGeneScan	Simulated short reads of 9 genomes	400	91.3	86.1	88.7	Rho et al. (2010)
MetaGeneMark	Fragments from 50 microbial chromosomes	400	97.0	94.6	95.8	Zhu et al. (2010)
Glimmer-MG	Simulated 454 sequences	535	98.4	71.8	85.1	Kelley et al. (2012)
MetaGeneAnnotator	Subsequences of 13 genomes	700	95.1	91.0	93.1	Noguchi et al. (2008)
FragGeneScan	Simulated reads with 1 % sequencing error rate	400	85.4	79.5	82.5	Rho et al. (2010)
Glimmer-MG	Simulated 454 reads with 1 % sequencing error rate	535	83.6	62.5	73.1	Kelley et al. (2012)

frequency matrix can be derived by algorithms such as MCMC (Markov chain Monte Carlo)-based Gibbs sampler (Lawrence et al. 1993) or EM (Expectation Maximization)-based MEME (Bailey and Elkan 1994); detection of the RBS motif is done by finding the most conserved set of ungapped sequence fragments within the multiple alignment window. The structure of two-component RBS model is convenient for incorporation into HMM-based framework of several algorithms such as MetaGeneMark and FragGeneScan

Another feature, the prokaryotic gene length distribution, is approximated for complete or draft genomes by the gamma distribution with mean value about 900 nt; yet another one, the distribution of length of noncoding region is approximated by exponential distribution. These two distributions, as well as the RBS spacer length distribution, are used as in the HSMM-based algorithms (Besemer et al. 2001). Since short metagenomic sequences are more likely to contain partial genes than complete genes, length distributions of partial genes are used in HSMM-based metagenomic gene finders (Rho et al. 2010; Zhu et al. 2010).

About 70 % of neighboring genes in prokaryotic genomes have the same orientation (Noguchi et al. 2006), and many of them make co-transcribed "chains" or operons. Genes in an operon are located on a close distance or even overlap. Four base-pair overlap ATGA is very common in adjacent genes as an overlap of stop and start codons ATG and TGA. Average distance between adjacent genes having the same orientation is shorter than that between neighbor genes residing in complementary strands, especially in gene start-to-gene start configuration where additional space has to be available for promoters.

All these features are incorporated in metagenomic gene finders, e.g., MetaGeneMark. Tests of ab initio gene finders on simulated metagenomic sequences have shown that these algorithms are quite accurate, with average values of sensitivity and specificity above 90 %; see Table 1. However, the sensitivity drops if the sequence length goes below 200 nt (Yok and Rosen 2011; Zhu et al. 2010).

An Initio Gene Finding in Metagenomic Sequences with Errors

Real metagenomic sequences contain errors: substitutions, insertion, and deletions (indels), as well

Ab Initio Gene Identification in Metagenomic Sequences, Table 2 Frameshift prediction accuracy

Programs	Sequence length (bp)	Sn (%)	Sp (%)	(Sn + Sp)/2	Test set	Publication
FragGeneScan	400	81.0	43.2	62.1	Fragments from 18 prokaryotic genomes with 20 % containing frameshifts	Tang et al. 2013
	600	81.9	35.1	58.5		
	800	82.8	29.4	56.1		
MetaGeneTrack	400	75.8	70.2	73.0		
	600	80.1	61.7	70.9		
	800	81.5	51.9	66.7		

as chimerisms, when two reads from different species are joined due to assembly error. Indels can cause frameshifts in coding regions; thus gene prediction accuracy is affected by sequencing errors. The overall effect on accuracy depends on error rates specific to sequencing and finishing technologies; for example, the error rates reported for Sanger sequencing may be as low as 0.001 % while sequencing errors in NGS technologies can go above 1 %. In both simulated Sanger reads and simulated 454 reads significant decrease of gene prediction sensitivity is observed when error rate exceeds 1 % (Hoff 2009). Still, in assembled sequences, the per-nucleotide error rate of 0.5 % in raw reads can be reduced to as low as 0.005 %. This error rate is still large enough to affect ~3–4.5 % of genes in assembled sequences (Luo et al. 2012).

To identify frameshift errors in metagenomic sequences, gene-finding algorithms have to model frame transitions that occur due to sequencing errors. In HSMM-based gene finders, e.g., FragGeneScan, new hidden states designating transitions between coding frames in the same strand were incorporated into the HSMM architecture. Another recent tool able to detect frameshift in metagenomic coding regions is MetaGeneTack (Tang et al. 2013). It combines the original HSMM-based MetaGeneMark with an ab initio frameshift finding program GeneTack (Antonov and Borodovsky 2010). Several filters of false-positive predictions were employed in MetaGeneTack to achieve higher accuracy. MetaGeneTack is reported to have higher frameshift prediction specificity than FragGeneScan (Table 2) in reads with error rate typical for metagenomic projects (Tang et al. 2013).

Yet another approach was used in Glimmer-MG, which, to trace possible indel errors, splits an ORF into three branches (frames), starting from the position of a nucleotide called with low confidence (Kelley et al. 2012). This approach was reported to have higher gene prediction accuracy on error-contained reads than FragGeneScan. Methods that account for sequencing errors generally perform better in real error-prone metagenomic sequences than "idealistic" approaches. The accuracy of sequencing error detection, however, depends on how accurate is the modeling of sequencing errors is.

Summary

Accurate ab initio gene prediction in metagenomic sequences is necessary for reliable functional annotation. Ab initio algorithms identify genes in metagenomic sequences by detecting intrinsic statistical patterns of coding and noncoding regions. Being independent of data stored in databases, these methods are especially useful for discovering novel genes. Special techniques have been developed for derivation of parameters of the ab initio algorithms working with short anonymous metagenomic sequences. We have reviewed several ab initio gene finders developed for metagenomic sequences including the latest tools that take into account possible sequencing errors (frameshifts).

Cross-References

- ▶ Computational Approaches for Metagenomic Datasets
- ▶ FragGeneScan: Predicting Genes in Short and Error-Prone Reads
- ▶ Metagenomics, Metadata, and Meta-analysis
- ▶ Protein-Coding Genes as Alternative Markers in Microbial Diversity Studies
- ▶ Proteomics and Metaproteomics
- ▶ RITA: Rapid Identification of High-Confidence Taxonomic Assignments for Metagenomic Data

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25(17):3389–402.

Antonov I, Borodovsky M. Genetack: frameshift identification in protein-coding sequences by the viterbi algorithm. J Bioinforma Comput Biol. 2010;8(3):535–51. PubMed PMID: 20556861.

Badger JH, Olsen GJ. CRITICA: coding region identification tool invoking comparative analysis. Mol Biol Evol. 1999;16(4):512–24.

Bailey TL, Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proceedings/International Conference on Intelligent Systems for Molecular Biology; ISMB International Conference on Intelligent Systems for Molecular Biology, Vol. 2; 1994; p. 28–36. PubMed PMID: 7584402.

Besemer J, Borodovsky M. Heuristic approach to deriving models for gene finding. Nucleic Acids Res. 1999;27(19):3911–20. PubMed PMID: 10481031. Pubmed Central PMCID: 148655.

Besemer J, Lomsadze A, Borodovsky M. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. Nucleic Acids Res. 2001;29(12):2607–18. PubMed PMID: 11410670. Pubmed Central PMCID: 55746.

Borodovsky M, McIninch J. GENMARK: parallel gene recognition for both DNA strands. Comp Chem. 1993;17(2):123–33.

Borodovsky MY, Sprizhitskii Y, Golovanov E, Aleksandrov A. Statistical patterns in primary structures of functional regions in the *E. coli* genome. III. Computer recognition of coding regions. Mol Biol. 1986;20:1145–50.

Brady A, Salzberg SL. Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. Nat Methods. 2009;6(9):673–6. PubMed PMID: 19648916. Pubmed Central PMCID: 2762791.

Bult CJ, White O, Olsen GJ, Zhou L, Fleischmann RD, Sutton GG, et al. Complete genome sequence of the methanogenic archaeon. *Methanococcus jannaschii*. Science. 1996;273(5278):1058–73. PubMed PMID: 8688087.

Chen SL, Lee W, Hottes AK, Shapiro L, McAdams HH. Codon usage between genomes is constrained by genome-wide mutational processes. Proc Natl Acad Sci U S A. 2004;101(10):3480–5. PubMed PMID: 14990797. Pubmed Central PMCID: 373487.

Delcher AL, Bratke KA, Powers EC, Salzberg SL. Identifying bacterial genes and endosymbiont DNA with Glimmer. Bioinformatics. 2007;23(6):673–9. PubMed PMID: 17237039. Pubmed Central PMCID: 2387122.

Frishman D, Mironov A, Mewes H-W, Gelfand M. Combining diverse evidence for gene recognition in completely sequenced bacterial genomes. Nucleic Acids Res. 1998;26(12):2941–7.

Gish W, States DJ. Identification of protein coding regions by database similarity search. Nat Genet. 1993;3(3):266–72.

Hoff KJ. The effect of sequencing errors on metagenomic gene prediction. BMC Genomics. 2009;10:520. PubMed PMID: 19909532. Pubmed Central PMCID: 2781827.

Hoff KJ, Tech M, Lingner T, Daniel R, Morgenstern B, Meinicke P. Gene prediction in metagenomic fragments: a large scale machine learning approach. BMC Bioinforma. 2008;9:217. PubMed PMID: 18442389. Pubmed Central PMCID: 2409338.

Hoff KJ, Lingner T, Meinicke P, Tech M. Orphelia: predicting genes in metagenomic sequencing reads. Nucleic Acids Res. 2009 Jul 37(Web Server issue): W101-5. PubMed PMID: 19429689. Pubmed Central PMCID: 2703946.

Kelley DR, Liu B, Delcher AL, Pop M, Salzberg SL. Gene prediction with Glimmer for metagenomic sequences augmented by classification and clustering. Nucleic Acids Res. 2012;40(1):e9. PubMed PMID: 22102569. Pubmed Central PMCID: 3245904.

Kunin V, Copeland A, Lapidus A, Mavromatis K, Hugenholtz P. A bioinformatician's guide to metagenomics. Microbiol Mol Biol Rev. 2008;72(4):557–78. Table of Contents. PubMed PMID: 19052320. Pubmed Central PMCID: 2593568.

Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. Science. 1993;262(5131):208–14. PubMed PMID: 8211139.

Luo C, Tsementzi D, Kyrpides N, Read T, Konstantinidis KT. Direct comparisons of Illumina vs. Roche 454 sequencing technologies on the same microbial community DNA sample. PloS ONE. 2012;7(2): e30087.

Noguchi H, Park J, Takagi T. MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. Nucleic Acids Res. 2006;34(19):5623–30. PubMed PMID: 17028096. Pubmed Central PMCID: 1636498.

Noguchi H, Taniguchi T, Itoh T. MetaGeneAnnotator: detecting species-specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes. DNA Res Int J Rapid Publ Rep Genes Genomes. 2008;15(6):387–96. PubMed PMID: 18940874. Pubmed Central PMCID: 2608843.

Rho M, Tang H, Ye Y. FragGeneScan: predicting genes in short and error-prone reads. Nucleic Acids Res. 2010;38(20):e191. PubMed PMID: 20805240. Pubmed Central PMCID: 2978382.

Salzberg SL, Delcher AL, Kasif S, White O. Microbial gene identification using interpolated Markov models. Nucleic Acids Res. 1998;26(2):544–8. PubMed PMID: 9421513. Pubmed Central PMCID: 147303.

Tang S, Antonov I, Borodovsky M. MetaGeneTack: ab initio detection of frameshifts in metagenomic sequences. Bioinformatics. 2013;29(1):114–6. PubMed PMID: 23129300. Pubmed Central PMCID: 3530910.

Wooley JC, Godzik A, Friedberg I. A primer on metagenomics. PLoS Comput Biol. 2010;6(2): e1000667. PubMed PMID: 20195499. Pubmed Central PMCID: 2829047.

Yok NG, Rosen GL. Combining gene prediction methods to improve metagenomic gene annotation. BMC Bioinforma. 2011;12:20. PubMed PMID: 21232129. Pubmed Central PMCID: 3042383.

Zhu W, Lomsadze A, Borodovsky M. Ab initio gene identification in metagenomic sequences. Nucleic Acids Res. 2010;38(12):e132. PubMed PMID: 20403810. Pubmed Central PMCID: 2896542.

AbundanceBin, Metagenomic Sequencing

Yuzhen Ye
Indiana University, School of Informatics and Computing, Bloomington, IN, USA

Definition

Binning is unsupervised clustering of metagenomic sequences into an unknown set of species.

AbundanceBin is a binning tool utilizing the different abundances of the species in a community.

Introduction

Binning is one of the challenging problems in the metagenomics field. It has two main applications. One application is for studying the structure of microbial communities. The other application is for improving the downstream analysis of metagenomic sequences, including metagenome assembly (which has shown to be extremely difficult), considering that assembling reads one bin at a time significantly reduces the complexity of the metagenome assembly problem.

Composition-based methods have been the main approaches to unsupervised classification of reads. The basis of these approaches is that the genome composition (G + C content, dinucleotide frequencies, and synonymous codon usage) vary among organisms and are generally characteristic of evolutionary lineages. Tools in this category include TETRA (Teeling et al. 2004), TACOA (Diaz et al. 2009), and MetaCluster (Leung et al. 2011). Due to the substantial variance in sequence properties along a genome, the main limitation of composition-based approaches is that they require relatively long reads (at least 800 bp), although it is shown that MetaCluster (Leung et al. 2011) can bin reads of 300 bp by employing a different distance metric (Spearman Footrule Distance) to reduce the local variations for 4-mers.

Note a large collection of methods have been developed to classify sequencing reads in a supervised manner. MEGAN (Huson and Mitra 2012) is a representative approach of this kind. These methods either use composition information (as in NCB, a naïve Bayes classifier to metagenomic sequence classification (Rosen et al. 2011)) or employ similarity searches of metagenomic sequences against a database of known genes/proteins (as in MEGAN) and assign metagenomic sequences to taxa accordingly, with or without using phylogeny. They also differ in the algorithms used for classification: MEGAN pioneers the lowest common ancestor (LCA) algorithm (Huson et al. 2007), MTR (Gori et al. 2011) improves on LCA algorithm considering multiple taxonomic ranks, and MetaPhyler (Liu et al. 2011) achieves better classification results by tuning the taxonomic classifier to each matching length, reference gene, and taxonomic level. Note that some tools in this category can only classify a subset of the metagenomic sequences instead of all. MLTreeMap (Stark et al. 2010) uses phylogenetic analysis of 31 marker genes for taxonomic distribution estimation. CARMA (Krause et al. 2008) searches for conserved Pfam domains and protein families in raw metagenomic sequences and classifies them into a higher-order taxonomy. RDP classifier is designed for classification of 16S rRNA genes, and later extended to classification of 18S rRNA genes using a naïve Bayes classifier (Cole et al. 2009).

AbundanceBin

AbundanceBin (Wu and Ye 2011) is the first unsupervised clustering algorithm that utilizes abundance information of the species in the same microbial community to group reads into bins. The fundamental assumption of the AbundanceBin algorithm is that reads are sampled from genomes following a Poisson procedure, such that the sequencing reads can be modeled as a mixture of Poisson distribution.

An expectation–maximization (EM) algorithm is used in AbundanceBin to find parameters for the Poisson distributions (i.e., the means), which reflect the relative abundance levels of the source species. AbundanceBin then assigns reads to bins based on the fitted Poisson distributions. AbundanceBin gives an estimation of the genome size (or the concatenated genome size of species of the same or very similar abundances) and the coverage (which reflects the abundances of species) of each bin in an unsupervised manner without requiring prior knowledge of the structure of the microbial communities. The EM algorithm needs an important parameter, the number of bins, which is typically unknown, as for most metagenomic projects. AbundanceBin solves this problem by using a recursive binning approach to determine the total number of bins automatically. The recursive binning approach works by separating a dataset into two bins and proceeds by further splitting bins. The recursive procedure continues if (1) the predicted abundance values of two bins differ significantly; (2) the predicted genome sizes are larger than a certain threshold; and (3) the number of reads associated with each bin is larger than a certain threshold proportion of the total number of reads classified in the parent bin.

AbundanceBin achieves accurate classification of even very short sequences sampled from species with different abundance levels, as tested on simulated and real metagenomic datasets. The software is available for download at http://omics.informatics.indiana.edu/AbundanceBin.

Integrated Binning Methods

MetaCluster 3.0 is an integrated binning method based on the unsupervised top–down separation and bottom–up merging strategy, which can bin metagenomic fragments of species with very balanced abundance ratios to very different abundance ratios (Leung et al. 2011). MetaCluster 4.0 further improves the binning algorithm and is able to handle datasets with large number of species (e.g., 100 species) (Wang et al. 2012). MetaCluster is available for download at http://i.cs.hku.hk/~alse/MetaCluster/.

Joint Analysis of Multiple Metagenomic Samples

Baran and Halperin proposed an abundance-based (also termed as coverage-based) binning algorithm (MultBin) that operates on multiple samples of the same environment simultaneously, assuming that the different samples contain the same microbial species, possibly in different proportions (Baran and Halperin 2012). MultBin employs a *k*-medoids clustering algorithm to cluster reads according to their coverage across the samples. Testing of MultBin on simulated metagenomic datasets shows that integrating information across multiple samples yields more precise binning on each of the samples.

Summary

Abundance-based (or coverage-based) binning approaches achieve an accurate performance even for extremely short reads – when there exist species abundance differences, an ability that cannot be achieved by composition-based approaches which suffer from the variances of the compositions of short reads. Approaches that integrate abundance and composition information and approaches that utilize multiple samples have shown promising binning results.

References

Baran Y, Halperin E. Joint analysis of multiple metagenomic samples. PLoS Comput Biol. 2012;8(2):e1002373.

Cole JR, Wang Q, Cardenas E, et al. The ribosomal database project: improved alignments and new tools for rRNA analysis. Nucleic Acids Res. 2009;37(Database issue):D141–5.

Diaz NN, Krause L, Goesmann A, et al. TACOA: taxonomic classification of environmental genomic fragments using a kernelized nearest neighbor approach. BMC Bioinformatics. 2009;10:56.

Gori F, Folino G, Jetten MS, et al. MTR: taxonomic annotation of short metagenomic reads using clustering at multiple taxonomic ranks. Bioinformatics. 2011;27(2):196–203.

Huson DH, Mitra S. Introduction to the analysis of environmental sequences: metagenomics with MEGAN. Methods Mol Biol. 2012;856:415–29.

Huson DH, Auch AF, Qi J, et al. MEGAN analysis of metagenomic data. Genome Res. 2007;17(3):377–86.

Krause L, Diaz NN, Goesmann A, et al. Phylogenetic classification of short environmental DNA fragments. Nucleic Acids Res. 2008;36(7):2230–9.

Leung HC, Yiu SM, Yang B, et al. A robust and accurate binning algorithm for metagenomic sequences with arbitrary species abundance ratio. Bioinformatics. 2011;27(11):1489–95.

Liu B, Gibbons T, Ghodsi M, et al. Accurate and fast estimation of taxonomic profiles from metagenomic shotgun sequences. BMC Genomics. 2011;12 Suppl 2:S4.

Rosen GL, Reichenberger ER, Rosenfeld AM. NBC: the naive bayes classification tool webserver for taxonomic classification of metagenomic reads. Bioinformatics. 2011;27(1):127–9.

Stark M, Berger SA, Stamatakis A, et al. MLTreeMap–accurate maximum likelihood placement of environmental DNA sequences into taxonomic and functional reference phylogenies. BMC Genomics. 2010;11:461.

Teeling H, Waldmann J, Lombardot T, et al. TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. BMC Bioinformatics. 2004;5:163.

Wang Y, Leung HC, Yiu SM, et al. MetaCluster 4.0: a novel binning algorithm for NGS reads and huge number of species. J Comput Biol. 2012;19(2):241–9.

Wu YW, Ye Y. A novel abundance-based algorithm for binning metagenomic sequences using l-tuples. J Comput Biol. 2011;18(3):523–34.

Accurate Genome Relative Abundance Estimation Based on Shotgun Metagenomic Reads

Fengzhu Sun and Li Charlie Xia
Molecular and Computational Biology Program, Department of Biological Sciences, University of Southern California, Dana and David Dornsife College of Letters, Arts and Sciences, Los Angeles, CA, USA

Synonyms

Genome Relative Abundance estimation using Mixture Model theory (GRAMMy)

Introduction

Accurate estimation of microbial community composition based on metagenomic sequencing data is fundamental for subsequent metagenomic analysis. However, it is also a challenging computational problem because of the mixed nature of metagenomes and the fact that only a small fraction of them get sequenced.

With the advents of next-generation sequencing (NGS) technologies, there has been significant increase in sequencing capacity yet reduction in single read length. This paradigm shift in sequencing technologies has impacted downstream analyses. Specifically, the identification of the origin of a read becomes more difficult for several reasons. First, a large number of short reads cannot be uniquely mapped to a specific location of one genome. Instead, they map to multiple locations of one or multiple genomes.

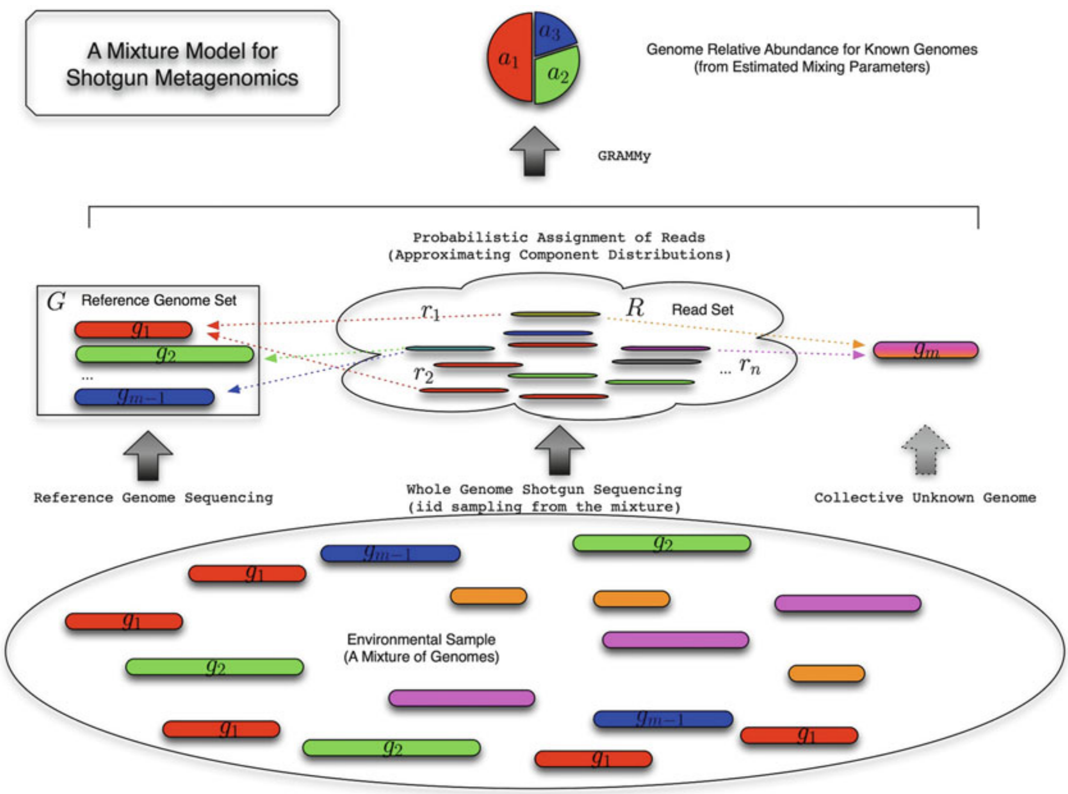
These ambiguities are directly associated with the read length reduction in NGS technologies. Second, communities usually consist of many microbes with similar genomes, different only in some parts, making it indeed impossible to determine the origin of a particular short read based solely on its sequence.

Despite these difficulties, NGS read sets have brought in richer abundance information of microbial communities than traditional datasets because of the significant increase in the number of reads. Along with the increase of read set size, efforts to assemble more reference genomes are ongoing. In addition, new experimental techniques, such as single-cell sequencing approaches, are being developed to sequence reference genomes directly from environmental samples. In face of the challenges from short reads and the opportunities from fast-expanding reference genome databases, GRAMMy is a statistical framework developed to accurately and efficiently estimate the relative abundance of microbial organisms within the community (Xia et al. 2011).

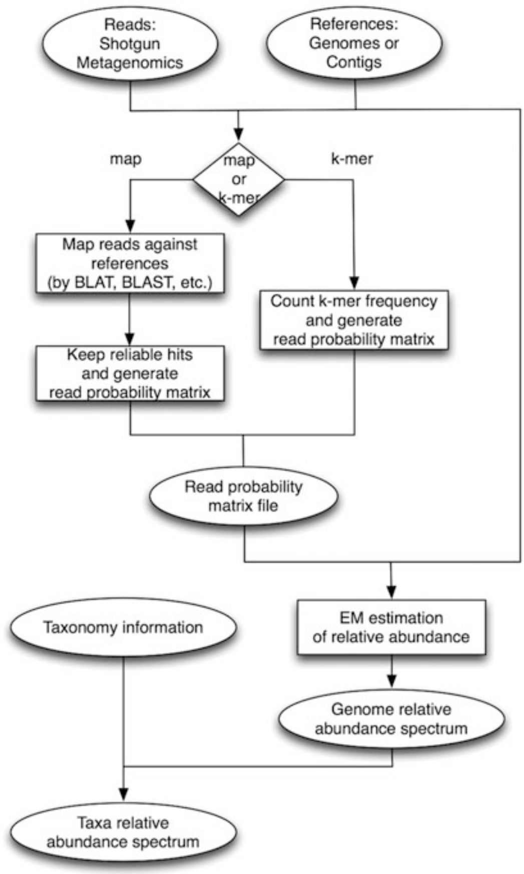
Description

The GRAMMy Framework

The GRAMMy framework is based on a mixture model for the short metagenomic sequencing and an expectation-maximization (EM) algorithm, as outlined in the model schema and the analysis flowchart in Figs. 1 and 2. GRAMMy accepts a set of shotgun reads as well as external references (e.g., genomes, scaffolds, or contigs) as inputs and subsequently performs the maximum-likelihood estimation (MLE) of the genome relative abundance (GRA) levels.



Accurate Genome Relative Abundance Estimation Based on Shotgun Metagenomic Reads, Fig. 1 The GRAMMy model. A schematic diagram of the finite mixture model underlies the GRAMMy framework for shotgun metagenomics. In the figure, "iid" stands for "independent identically distributed"



Accurate Genome Relative Abundance Estimation Based on Shotgun Metagenomic Reads, Fig. 2 The GRAMMy flowchart. A typical flowchart of GRAMMy analysis pipeline employs "map" and "k-mer" assignment

In the typical GRAMMy workflow, which is shown in Fig. 2, the end user starts with the metagenomic read set and reference genome set and then chooses between mapping-based ("map") and k-mer composition-based ("k-mer") assignment options (He and Xia 2007). In either option, after the assignment procedure, an intermediate matrix describing the probability that each read is assigned to one of the reference genomes is produced. This matrix, along with the read set and reference genome set, is fed forward to the EM algorithm module for estimation of the GRA levels. After the calculation, GRAMMy outputs the GRA estimates as a numerical vector, as well as the log-likelihood and standard errors for the estimates. If the taxonomy information for the input reference genomes is available, strain (genome) level GRA estimates can be combined to calculate high taxonomic level abundance, such as species- and genus-level estimates.

Accurate GRAMMy Estimates with EM Algorithm

Formally, GRA is defined as the normalized abundance for m unique genomes, where the relative abundance for the jth known genome is

$$a_j = \frac{\#\text{j-th genome}}{\#\text{known genomes}}$$

Note that g_m is a collective surrogate for unknown genomes and cannot be estimated in the model. Knowing length l_j, a_j is one-to-one related to the corresponding mixing parameter π_j by

$$a_j = \frac{\pi_j}{l_j \sum_{k=1}^{m-1} \frac{\pi_k}{l_k}}$$

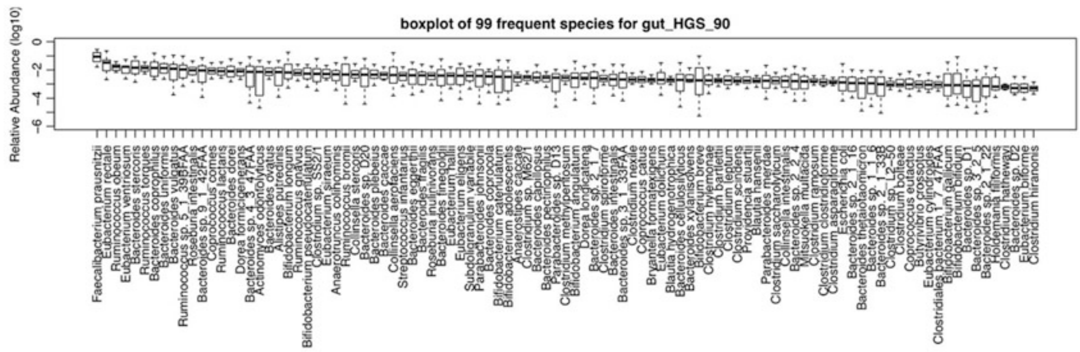
Mixing component distributions are needed to solve for mixing parameter π, which are $p(r_i | z_{ij} = 1; \mathbf{g})$'s – i.e., the probabilities of generating a read r_i from g_j. They are approximated empirically. The first approach is to use the number of high-quality hits s_{ij} from BLAST, BLAT, or other mapping tools and approximate by $\frac{s_{ij}}{l_j}$; the second approach is to use k-mer composition as detailed in the original study (Xia et al. 2011). The EM algorithm to calculate π iterates between E-step

$$z_{ij}^{(t)} = \frac{p\left(r_i | z_{ij} = 1; g\right)\pi_j^{(t)}}{\sum_{k=1}^{m} p(r_i | z_{ik} = 1; g)\pi_k^{(t)}}$$

and M-step

$$\pi_j^{(t+1)} = \frac{\sum_{i=1}^{n} z_{ij}^{(t)}}{n}$$

until convergence, where n is the total number of reads and z_{ij}'s are entries in the missing read



Accurate Genome Relative Abundance Estimation Based on Shotgun Metagenomic Reads, Fig. 3 Frequent species of human gut microbiome. The 99 species occurring in at least 50 % of the 33 human gut samples with a minimum relative abundance of 0.05 % were selected

origin matrix $\boldsymbol{Z}$. The estimated mixing parameters π are then converted back to GRA estimates $\mathbf{a}$.

GRAMMy Estimates for Human Gut Metagenomes

The human gastrointestinal tract harbors the largest group of human symbiotic microbes. Figure 3 shows the 99 most frequent species of human gut based on the GRAMMy analysis of the 33 metagenomic samples collected from three human gut metagenome projects (Gill et al. 2006; Kurokawa et al. 2007; Turnbaugh et al. 2009). The medians of estimated average genome lengths for these metagenomes range from 2.8 to 3.7 Mbp. Among the top ten most frequent species, there are eight from the *Firmicutes* phylum including members of *Faecalibacterium*, *Eubacterium*, and *Ruminococcus* genera, and two from the *Bacteroides* genus of the *Bacteroidetes* phylum. *Firmicutes* and *Bacteroidetes* dominate the human gastrointestinal tract. Species' relative abundance displays a long-tail distribution, suggesting that many are detected across samples, though most of them are not highly abundant. The abundance levels of some species are highly variable (with larger box size), while most others are relatively constant.

Conclusions

GRAMMy is a rigorous probabilistic framework for accurately and efficiently estimating genome relative abundance (GRA) based on shotgun metagenomic reads. Users have a wide choice of mapping and alignment tools to assign reads to references. The method is particularly suitable for NGS short read datasets due to its better handling of read assignment ambiguities. GRAMMy tools are packaged as a C++ extension to Python, which can be downloaded freely from GRAMMy's homepage: http://meta.usc.edu/softs/grammy.

Cross-References

- ▶ Approaches in Metagenome Research: Progress and Challenges
- ▶ Computational Approaches for Metagenomic Datasets
- ▶ Extended Local Similarity Analysis (eLSA) of Biological Data
- ▶ Metagenomic Research: Methods and Ecological Applications
- ▶ Metagenomics, Metadata, and Meta-analysis
- ▶ Molecular Ecological Network of Microbial Communities

References

Gill SR, Pop M, Deboy RT, et al. Metagenomic analysis of the human distal gut microbiome. Science. 2006;312(5778):1355–9.
He PA, Xia L. Oligonucleotide profiling for discriminating bacteria in bacterial communities. Comb Chem High Throughput Screen. 2007;10(4):247–55.
Kurokawa K, Itoh T, Kuwahara T, et al. Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. DNA Res. 2007;14(4):169–81.
Turnbaugh PJ, Hamady M, Yatsunenko T, et al. A core gut microbiome in obese and lean twins. Nature. 2009;457(7228):480–4.
Xia LC, Cram JA, Chen T, et al. Accurate genome relative abundance estimation based on shotgun metagenomic reads. PLoS One. 2011;6(12):e27992.

All-Species Living Tree Project

Pablo Yarza[1], Raul Munoz[2], Jean Euzéby[3], Wolfgang Ludwig[4], Karl-Heinz Schleifer[4], Rudolf Amann[5], Frank Oliver Glöckner[6,7] and Ramon Rosselló-Móra[2]
[1]Ribocon GmbH., Bremen, Germany
[2]Marine Microbiology Group, Department of Ecology and Marine Resources, Institut Mediterrani d'Estudis Avançats (CSIC-UIB), Illes Balears, Spain
[3]Society of Systematic Bacteriology and Veterinary (SBSV) & National Veterinary School de Toulouse (ENVT), Toulouse, France
[4]Lehrstuhl Für Mikrobiologie, Technische Universität München, Freising, Germany
[5]Molecular Ecology Group, Max Planck Institute for Marine Microbiology, Bremen, Germany
[6]Microbial Genomics and Bioinformatics Group, Max Planck Institute for Marine Microbiology, Bremen, Germany
[7]Jacobs University Bremen gGmbH, Bremen, Germany

Synonyms

16SrRNA(SSU) and 23SrRNA(LSU) gene sequence databases; Alignments; LTP project; Manual curation; "Orphan" species; Taxa boundaries; Taxonomy/classification/phylogeny of Bacteria and Archaea; Type strains

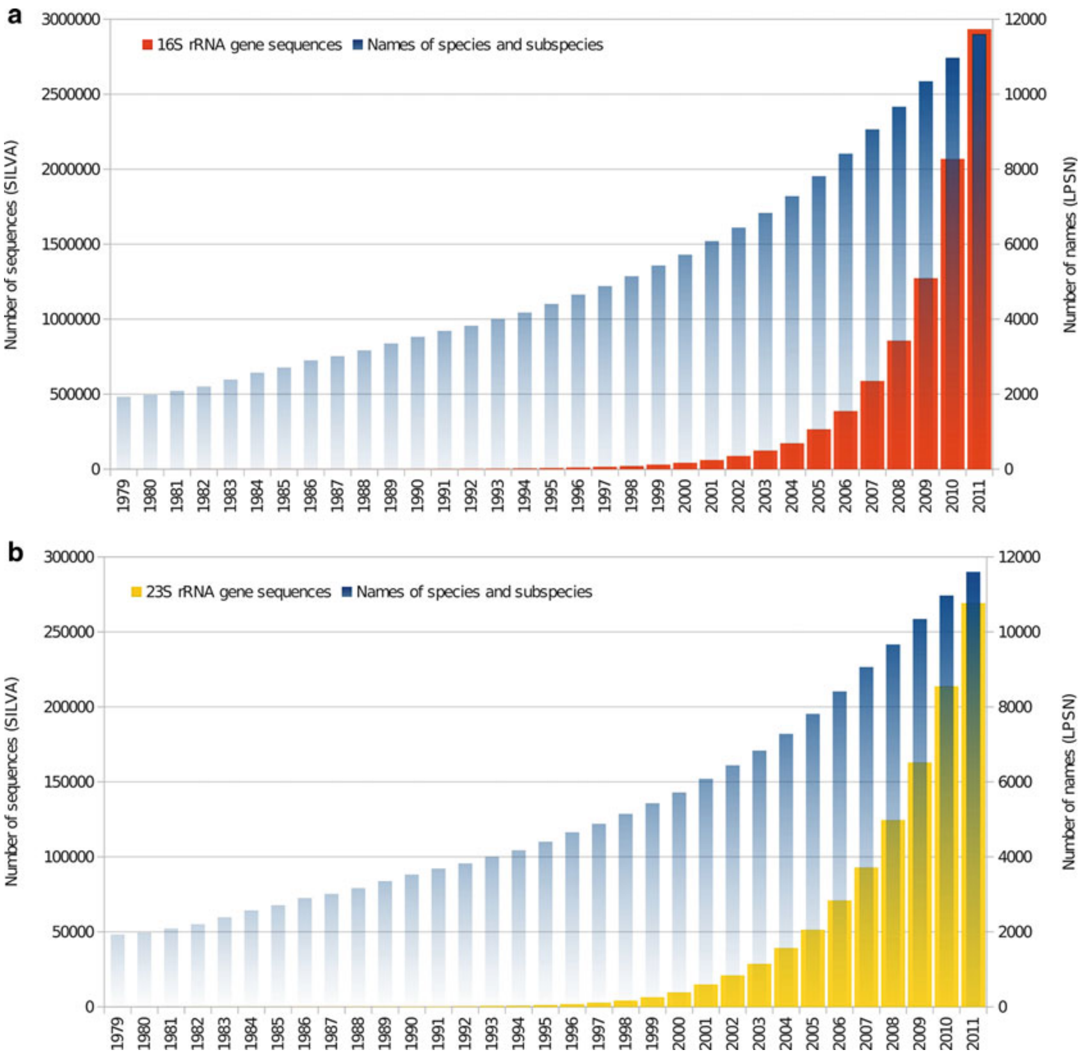
Definition

The All-Species Living Tree Project (LTP) is an international initiative for the creation and maintenance of highly curated 16SrRNA and 23SrRNA gene sequence databases, alignments, and phylogenetic trees for all the type strains of *Bacteria* and *Archaea*.

Introduction

Classification and identification of *Bacteria* and *Archaea* came across to a turning point around 35 years ago. It was the time when Carl Woese and co-workers demonstrated that ribosomal markers were appropriate to infer genealogical relationships by means of phylogenetic reconstructions (Fox et al. 1977). Rapidly, comparative analysis of rRNA gene sequences became a standard procedure with mature implications in microbial ecology and taxonomy: culture-independent exploration of ecosystems' diversity (Amann et al. 1995) and settlement of the phylogenetic backbone (i.e., our current accepted classification of *Bacteria* and *Archaea*; Garrity 2001). As a result, the total amount of ribosomal RNA entries in the public DNA databases has grown exponentially since early 1990s, currently comprising at least 3,500,000 small (SSU) and 300,000 large (LSU) ribosomal subunit gene sequence entries. On the other hand, the number of bacterial and archaeal species with validly published names has followed arithmetic trends with a ratio of around 500–700 annual descriptions during the last 7 years (Fig. 1), currently (December 2012) exceeding the total number of 10,300 species and subspecies. A comparative overview of these trends until December 2011 is shown in Fig. 1.

As from early 1990s, the 16S rRNA has been, by orders of magnitude, the most often sequenced gene, there is no alternative phylogenetic marker with such a high coverage in public repositories. However, abundance is not the single requisite for a proper phylogenetic inference and other single molecules (e.g., 23S rRNA) or combinations of them might perform better at reflecting



All-Species Living Tree Project, Fig. 1 Annual growth of ribosomal 16S rRNA (**a**) and 23S rRNA (**b**) gene sequence databases and species and subspecies names with standing in nomenclature until December 2011. SILVA SSU-Parc111 and LSU-Parc111 databases (http://www.arb-silva.de/documentation/release-111/) were filtered by submission date until December 2011 and its cumulative annual growth was plotted in red (SSU, 1A) and yellow bars (LSU, 1B). The cumulative growth of published species and subspecies names (according to LPSN; http://www.bacterio.cict.fr/number.html) since 1980 until December 2011 is plotted in blue. Note that the total number of names is around 2,000 above the total number of distinct type strains due to homotypic synonyms, new combinations, nomina nova, later heterotypic synonyms, or illegitimate names

genealogies of certain groups given the higher information content (Ludwig and Klenk 2001). Although far from reaching 16S rRNA levels, submission of alternative markers is growing fast, mostly because (i) the number of metagenomes and complete genomes is growing exponentially due to the reduction on sequencing and analysis costs and (ii) the recent initiative to complete the genome sequence of all type strains (GEBA initiative). Undoubtedly, comparative genomics will involve a new breakthrough for microbial taxonomy and the current phylogenetic backbone based on ribosomal sequences will be carefully reviewed (Coenye et al. 2005). Nevertheless, at this point, the number of sequenced genomes of type strains is still low and therefore the current possibilities for an in-depth taxonomic study are sparse.

The responsible teams of the ARB, SILVA, and LPSN projects (www.arb-home.de, www.arb-silva.de, and www.bacterio.net) together with the journal Systematic and Applied Microbiology (SAM) started the "All-Species Living Tree Project" (LTP; http://www.arb-silva.de/projects/living-tree), a project conceived to provide a tool especially designed for the microbial taxonomist scientific community (Yarza et al. 2008). The main objectives considered so far are (1) provide a curated 16S and 23S rRNA gene database for the type strains of all species with validly published names; (2) set up an optimized and universally usable alignment; (3) reconstruct reliable phylogenetic trees with all the type strains; (4) maintain the database, alignments, and trees through regular updates including the new validly published taxa and their respective 16S and 23S rRNA gene sequences; and (5) investigate, with the use of the database, fundamental aspects about taxonomy of *Bacteria* and *Archaea* such as phylogenetic thresholds in new taxa circumscriptions, coherence of current taxonomy by means of phylogenetic schemes, and relevance of the ribosomal RNA genes in taxonomic studies.

Creation and Maintenance of LTP Releases

LTP Datasets

First LTP datasets (release LTPs93 for SSU (Yarza et al. 2008), release LTPs102 for LSU (Yarza et al. 2010)) were prepared following six main steps:

1. Set up a list of candidate sequences. An initial sequence dataset consisted on a subsample of the SILVA database, filtering by "type" (T) or "cultured" (C) strains; this information mainly came from StrainInfo.
2. Set up a list of species names. In parallel we built a comprehensive, updated, and nonredundant (i.e., free of synonyms and according to latest valid nomenclature) list of validly published species and subspecies names from LPSN. When a species is divided into subspecies, we substituted the original species name by that of the subspecies (e.g., *Staphylococcus sciuri* subsp. *sciuri* instead of *Staphylococcus sciuri*). We avoided the "Candidatus" names (e.g., "*Candidatus* Aciduliprofundum boonei"), *Cyanobacteria* not validly published under the Bacteriological Code (e.g., *Anabaena oscillatorioides*), and later heterotypic synonyms (e.g., *Pseudomonas chloritidismutans*).
3. Manual cross-check. Then, each entry from our initial list of sequences was assigned to a species name by manually examining the companion contextual metadata. This process had to be done manually given the often outdated, mistaken, or absent taxonomic information such as the organism name or the strain numbers.
4. Quest for missing type strains. We realized that not all species names were represented in the list of sequences. Then, we inverted the process by searching in resources like EMBL, Bergey's Outlines, issues of the International Journal of Systematic and Evolutionary Microbiology (IJSEM), etc. with the aim to find a good-quality sequence entry for each missing type strain.
5. "Orphan" species recognition. Finally, we got a group of type strains whose 16S/23S rRNA genes had never been sequenced or that the existing sequences were of too low quality to be considered for the project (i.e., in terms of sequence length, number of ambiguities, etc.). We called them "orphan" species. The LTP project together with eleven international culture collections has driven the sequencing of these "orphan" species through the SOS initiative (Yarza et al. 2013).
6. Keep one sequence per species. On the other hand, the list of type-strain sequences was redundant in the sense that one single type strain could be represented by multiple sequence entries. This is the case of multiple independent sequencings and submissions, or the existence of several sequences due to multiple copies of the ribosomal operon. The aim of the LTP is, whenever possible, to keep one

All-Species Living Tree Project, Table 1 Summary of LTP releases. "Sync" fields correspond to IJSEM and EMBL release dates. "Net increase" of a release is the number of new entries minus the number of deleted entries. "% incorrect" refers to the percentage of new entries whose INSDC records carried incorrect information in the organism name field. Averages include standard deviation

Release	Type	IJSEM sync	EMBL sync	Total entries	New entries	Deleted entries	Net increase	% incorrect[a]	Average length[a]	Average ambig.[b]
LTPs93	SSU	Dec. 2007	Dec. 2007	6,728	6,728	0	6,728	22	1,465.0 ± 51.2	0.10 ± 0.26
LTPs95	SSU	Jun. 2008	Jun. 2008	7,006	299	21	278	45	1,446.0 ± 46.3	0.04 ± 0.11
LTPs100	SSU	Aug. 2009	Jun. 2009	7,710	750	46	704	50	1,448.0 ± 54.2	0.03 ± 0.11
LTPs102	SSU	Feb. 2010	Nov. 2009	8,029	363	44	319	58	1,453.6 ± 52	0.33 ± 0.12
LTPs102	LSU	Feb. 2010	Nov. 2009	792	792	0	792	6	2,866.1 ± 177.6	0.02 ± 0.11
LTPs104	SSU	Dec. 2010	May 2010	8,545	545	29	516	74	1,444.6 ± 62	0.27 ± 0.11
LTPs106	SSU	May 2011	Dec. 2010	8,815	279	9	270	77	1,445.9 ± 51.1	0.03 ± 0.12
LTPs108	SSU	Dec. 2011	Jun. 2011	9,279	490	26	464	60	1,455.4 ± 51.9	0.02 ± 0.09

[a] Average length for the "new entries"
[b] Average percentage of ambiguities for the "total entries"

sequence per type strain in order to maintain simplicity, avoid confusion, and improve tree navigation and database usability. In general, the best quality available (including manual inspection of the alignment) was selected for the project and, in case of doubt, the earliest submission to an INSDC partner (www.insdc.org). From release LTPs102 (Yarza et al. 2010), when multiple paralogues exist due to rRNA operon copy number, several copies are kept if they show less than 98 % sequence identity (see below for further details).

LTP is maintained by a scrutiny of the new described species, nomenclatural changes, taxonomic notes, and opinions that are monthly published in the IJSEM journal. Their respective 16S and 23S rRNA gene sequence entries are acquired from the latest SILVA release and appended to the existing LTP database. Therefore, SILVA's Reference (Ref) ARB databases (http://www.springerreference.com/docs/html/chapterdbid/304116.html) serve as template for the new LTP-ARB databases. Until now (December 2012) one LSU-based and seven SSU-based LTP releases have been produced (Table 1). New species are incorporated into the database only if they account a good-quality sequence existing in the respective SILVA release. Certain entries can be deleted if their corresponding species names are seen to be later heterotypic synonyms, if they become rejected, or as a matter of taxonomic opinions. Sequence entries existing in an LTP database can also change by means of their metadata. Thus, for example, new combinations (i.e., a type strain which changes its name due to reclassification) or subdivision of a species into subspecies leads to an entry modification at its taxonomic information fields.

Inaccurate or Mistaken Metadata

Inaccurate sequence-associated metadata tend to happen in more than 50 % of the new added 16S rRNA entries (Table 1). Often, these "mistakes" consist on a lack of entries' updating tasks at the time of their first appearance in a scientific publication. It mainly occurs in taxonomy-associated information fields. To prove the uniqueness of a new species and to name it take time and, in the meanwhile, sequences are quickly produced and easily submitted to nucleotide databases. Most often, these submissions only show genus specifications, for example, sequence entry GU808562 appears as "Hymenobacter sp. HMD1010" but its real name is *Hymenobacter yonginensis*. Indeed, a Bacteriological Code-compliant (Lapage et al. 1992) nomenclature may be somewhat tricky and is frequent to consider several Latin terms and derivations until one species name is finally accepted by authors and reviewers. Unavoidably, this bad-quality information is propagated from INSDC databases (primary sources) to other technological services like dedicated ribosomal databases (e.g., SILVA). Although extensive data curation is not a task of primary sources of information, it would be very beneficial that authors enhance their commitment with the correctness of the metadata provided (e.g., like the species name) or that authors are forced to update their INSDC entries prior to manuscript acceptation (recommended action for scientific journals). Successively, this rough data arrives finally to resources like LTP, which have no choice but checking it carefully to provide new informational fields with corrected information; curated information can return back to other resources of information.

Multiple Copies of the Ribosomal Operon

In 2010, a comprehensive study was conducted to evaluate the intra-genomic variability of the 16S rRNA gene on complete type-strain genomes (Yarza et al. 2010). We observed that in very unusual exceptions, the intra-genus (94.5 %; Yarza et al. 2008) or intraspecies (98.7 %; Stackebrandt and Ebers 2006) boundaries could be exceeded within a single genome. In such cases, the selection of one or another sequence might seriously affect the interpretation of a phylogenetic inference. However, despite the fact that the vast majority of strains contain multiple copies of the *rrn* operon, only 2 % of them reveal divergences beyond 2 % (30 nucleotides) sequence identity. Thus, most likely, the selection of one or another copy should not affect the phylogenetic reconstructions. Consequently, starting from release s104 (Munoz et al. 2011), the LTP database includes all paralogues with higher divergences than 2 %. By now, it is the case of three species: *Haloarcula marismortui* ATCC 43049^{T}, accession number AY596297, with 5.7 % of maximum inter-operonic divergence; *Thermoanaerobacter pseudethanolicus* ATCC 33223^{T}, accession number CP000924, with 3.66 % of maximum inter-operonic divergence; and *Desulfitobacterium hafniense* DCB-2^{T}, accession number CP001336, with 4.34 % of maximum inter-operonic divergence.

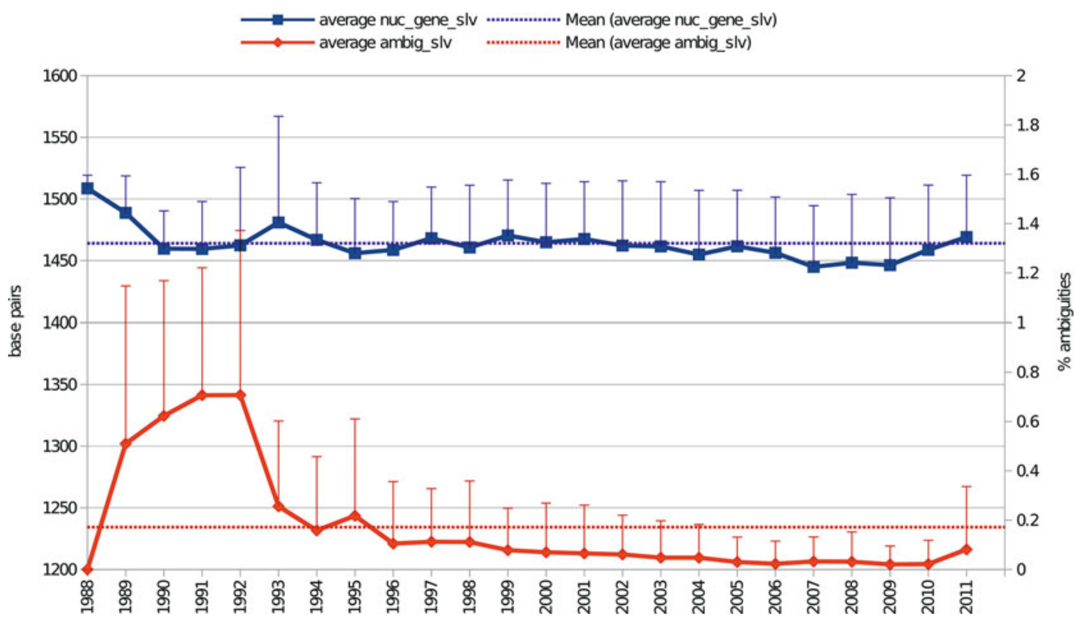
Sequence Quality in LTP Datasets

It has been suggested that sequences produced for taxonomic purposes should be equal or larger than 1,450 bases with less than 0.5 % ambiguities (Stackebrandt et al. 2002). Reason is that informative content of a molecular clock is linked to the total number of its variable positions (Ludwig and Klenk 2001). Statistics derived from LTP datasets indicate that in general, sequence quality is acceptable for in-depth phylogenetic studies (~1,455 bases and 0.02 % ambiguities for LTPs108; Table 1). Figure 2 shows annual variation of gene sequence length and percentage of ambiguities. Quality increase is mainly observed in terms of ambiguities reduction, probably related to amelioration of sequencing techniques. In any case, the completion of more full genome sequences of type strains will substantially increase the sequence quality (indicated by these two parameters) in the LTP database. Researchers should be encouraged to complete 5′ ends of 16S rRNA gene sequences, as first 250 bases contain hypervariable regions V1 and V2 which play an important role in comparisons between highly related organisms (Chakravorty et al. 2007).

Curated Metadata Introduced by the LTP

In addition to regular fields provided by the ARB-SILVA databases, sequence entries include now the following LTP-specific metadata fields:

1. *fullname_ltp*: corrected species name according to LPSN (http://www.bacterio.net).
2. *rel_ltp*: name of the LTP release where a sequence entry appeared for the first time.
3. *hi_tax_ltp*: name of the family where the taxon is classified. For unclassified genera, the name of the next available higher taxon above genus (e.g., "*Acidobacteria*" for *Bryobacter aggregatus*).
4. *type_ltp*: type species receive the label "type sp." in this field.
5. *riskgroup_ltp*: risk-group classification of microorganisms risk-group classification of microorganisms obtained from the DSMZ



All-Species Living Tree Project, Fig. 2 Annual distribution of the 16S rRNA gene sequence length and % of ambiguities in the 9,279 type-strain sequences corresponding to LTP release s108. Gene sequence length is given by the SILVA parameter "nuc_gene_slv" which cuts off the bases at the extremes when beyond the *E.coli'* s16S rRNA gene limits. Percentage of ambiguities is given by the SILVA descriptor "ambig_slv"

(Deutsche Sammlung von Mikroorganismen und Zellkulturen), according to the Federal Institute for Occupational Safety and Health (BAuA) in Germany.

6. *tax_ltp*: taxonomic classification into higher taxonomic ranks according to LPSN (http://www.bacterio.cict.fr/classifphyla.html).
7. *url_lpsn_ltp*: it contains the variable part of the URL leading to the LPSN's species file (e.g., http://www.bacterio.net/bryobacter.html).

Alignments and Phylogenetic Trees

Setting up universal alignments is a key step in order to achieve optimal and comparable phylogenetic reconstructions. It has been one of the constant motivations of Wolfgang Ludwig and co-workers who dealt with the huge task of preparing common and reliable alignment of ribosomal SSU and LSU sequences of *Bacteria*, *Archaea*, and *Eukarya* (Ludwig and Schleifer 1994). They found out that secondary structure formations such as loops and helices occurred at the same relative positions along the molecule. This helped to refine the alignments because variable stretches, with low sequence similarities, could be optimally positioned by recognizing functional homology (due to evolutionary pressure) and functional stability of helices (due to chemical stability of base pairs' bounds). A core dataset of sequences with highly curated alignments was incorporated into the SILVA system so new added sequences can be automatically aligned using this "seed alignment" as a reference (Ludwig et al. 2004; Pruesse et al. 2007). Periodically more and more manually curated sequences are added into the seed which improves its quality over time.

Although all new sequences incorporated into the LTP come from an ARB-SILVA database, they are again manually revised to further correct misplaced bases and to check highly variable regions. Before tree calculation, the complete alignment is shifted using maximum frequency filters (Table 2) that remove dubious orthologous positions caused by sequencing errors and hypervariability. Typically, LTP phylogenetic trees are calculated using the 40 % maximum frequency filter.

All-Species Living Tree Project, Table 2 Maximum frequency filters implemented into the LTPs 108ARB database

Filter name	Start position	Stop position	%min[a]	%max[a]	No. of positions[b]
LTPs108_ssu_10	0	50,000	10	100	1,433
LTPs108_ssu_20	0	50,000	20	100	1,433
LTPs108_ssu_30	0	50,000	30	100	1,432
LTPs108_ssu_40	0	50,000	40	100	1,390
LTPs108_ssu_50	0	50,000	50	100	1,288

[a]Minimum and maximum sequence identity. For tree reconstructions, only columns are taken into account if they have a positional conservation above the respective minimum values
[b]Number of homologous positions (columns) taken into account for tree reconstructions

The first 16S rRNA-based phylogenetic tree was calculated for the release LTPs93 (Yarza et al. 2008). The sequence dataset consisted of 6,728 type-strain sequences plus 3,247 supporting sequences belonging to non-type strains used to reinforce underrepresented groups and to stabilize the topology. The multiple alignment of 9,975 16S rRNA gene sequences was submitted to different treeing methodologies including neighbor-joining, maximum likelihood, and maximum parsimony, all tested with several filters (30 %, 40 %, and 50 % maximum frequency filters) and all implemented in the ARB software package (Ludwig et al. 2004). A high degree of congruence was observed among them. The tree considered as optimal was a 40 %-filtered maximum likelihood reconstruction calculated using the RAxML algorithm (Stamatakis 2006), with the GTRGAMMA correction, with 100 bootstrap replicates, in a 5-node and 20-processor parallel environment. The last de novo phylogenetic reconstruction appears in the release LTPs108 and was similarly calculated; tree calculation was run with a dataset of 12,166 16S rRNA gene sequences.

The phylogenetic tree calculated using the 23S rRNA gene was particularly challenging due to data shortage in many groups. In order to set up a reliable phylogeny based on 23S rRNA data, we defined a core dataset made of high-quality sequences (type and non-type strains). The stringent quality filtering approach ended with around 2,000 high-quality and nonredundant LSU sequences. This dataset was submitted to a maximum likelihood reconstruction in combination with a 50 % maximum frequency filter allowing 2,463 positions of the entire alignment. The missing partial or lower-quality type-strain sequences were added to the tree using the ARB parsimony tool with the option for keeping the initial topology while inserting additional data.

The groups shown in the trees are defined by recognizing the type members and according to the taxonomic classification. The trees are carefully compared against previously reported topologies and current taxonomic classifications (Yarza et al. 2010). All the additional supporting sequences used to reconstruct the phylogeny are removed from the final tree by keeping its topology intact. Within the ARB database, the type species are labeled with a distinct color for easy recognition and tree handling.

Files Provided by the LTP

As a taxonomic tool, the LTP must be understood as a collection of reference materials, all publicly available at the project's Web page (http://www.arb-silva.de/projects/living-tree), including:

1. Release documentation: (I) readme file with a release description and (II) PDF document describing the metadata fields introduced by the LTP
2. Tables: (I) new entries with outdated submission names and (II) list of changes in the dataset: added/deleted/modified entries
3. Export filter: ARB-export filter (*.eft* format) to extract data from LTP-ARB databases
4. Databases: (I) complete ARB databases including sequences, alignments, metadata, filters, and trees and (II) datasets in CSV format including LTP metadata

5. Alignments: (I) gapped exports in multi-FASTA format and (II) compressed exports in multi-FASTA format
6. Phylogenetic trees: (I) collapsed overviews in PDF format showing the distinct phyla, (II) full SSU (more than 80 pages long) and LSU trees in PDF format, and (III) full trees in NEWICK format, including group names and branch lengths

Side Research

Sequencing the Orphan Species Initiative (SOS)

The understanding that around 6 % of all classified species were missing from the ribosomal SSU sequence catalogues motivated us to start the "Sequencing the Orphan Species" (SOS) initiative (Yarza et al. 2013). During 3 years of work, the LTP team coordinated a network of 12 partner researchers and culture collections (ATCC, BZF, CECT, CIP, CCUG, DSMZ, JCM, ICMP, BCCM/LMG, MMG, NBRC, NCCB) in order to improve this situation by (re)sequencing the 16S rRNA gene of the "orphan" species. As a result, 351 type strains appear represented now by a good-quality SSU gene sequence in the databases. They comprise representatives of 14 bacterial and archaeal phyla, 76 type species, and 78 pathogenic species. However, 201 type strains could not be accessed as cultivable strains were not available at recognized culture collections. They represent 10 phyla and 17 type species.

Taxonomic Boundaries

In order to understand how the higher taxonomic categories could be circumscribed by means of a sequence identity threshold, we performed a statistical procedure to get the lowest similarity found within the members of a certain taxon (Yarza et al. 2008, 2010). By taking into account all the taxa at a particular taxonomic rank, we obtained general lower cutoff values of sequence identity for genus, family, and phylum based on 16S rRNA and 23S rRNA. In general, minimum 16S rRNA gene sequence identities of 94.9 % $\pm$ 0.4, 87.5 % $\pm$ 1.3, and 78.4 % $\pm$ 2.0 lead to the circumscription of a new genus, family, and phylum, respectively. For 23S rRNA genes, these values are slightly different: 93.2 % $\pm$ 1.3 (genus), 87.7 % $\pm$ 2.5 (family), and 75.3 % (phylum). As shown by the low errors, historically used criteria for genera, families, and phyla are quite homogeneous and do not lead to unambiguous circumscriptions. These cutoffs should be used with caution and always as a complementary approach. They are especially recommended for prospective studies in clone libraries and as additional support for the circumscription of new taxa or emendation of existing ones.

Summary

SSU and LSU databases made by the All-Species Living Tree Project (LTP; http://www.arb-silva.de/projects/living-tree) provide high-quality nearly full-length sequences of the type strains of all *Archaea* and *Bacteria* with validly published names. Setting up a type-strain sequences database included the sieving of the public DNA databases whose sequence entries often appeared outdated or mistaken at their taxonomic metadata. It involved the initial manual cross-check of nearly 14,000 SSU and 6,000 LSU sequence entries against the catalogue of distinct species with validly published names retrieved from LPSN. Databases are complemented with manually curated metadata, manually curated alignments, and state-of-the-art phylogenetic reconstructions (in contrast to other similar resources like the EzTaxon (Santamaria et al. 2012)). The LTP team wants to remark that the aim of the project is not to reconstruct the currently described species genealogy with total fidelity but to provide a curated taxonomic tool for the scientific community. Our small but very representative SSU and LSU datasets may be used as a reference for identification and classification purposes in several fields of application, for example, facilitating the collection of sequences for the reconstruction of taxa genealogies (Cousin et al. 2012), enabling fast and reliable taxonomic affiliations in rRNA surveys (Santamaria et al. 2012), or serving as reference datasets for testing bioinformatic procedures (Mizrahi-Man et al. 2013).

Cross-References

▶ Culture Collections in the Study of Microbial Diversity, Importance
▶ Phylogenetics, Overview
▶ SILVA Databases

References

Amann R, Ludwig W, Schleifer KH. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. Microbiol Rev. 1995;59:143–69.

Chakravorty S, Helb D, Burday M, et al. A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. J Microbiol Methods. 2007;69:330–9.

Coenye T, Gevers D, Van de Peer Y, et al. Towards a prokaryotic genomic taxonomy. FEMS Microbiol Rev. 2005;29:147–67.

Cousin S, Gulat-Okalla ML, Motreff L, et al. Lactobacillus gigeriorum sp. nov., isolated from chicken crop. Int J Syst Evol Microbiol. 2012;62:330–4.

Fox GE, Pechman KR, Woese CR. Comparative cataloguing of 16S ribosomal ribonucleic acid: molecular approach to prokaryotic systematics. Int J Bacteriol. 1977;27:44–57.

Garrity GM. Bergey's manual of systematic bacteriology. 2nd ed. New York: Springer; 2001.

Lapage SP, Sneath PHA, Lessel EF, et al. International code of nomenclature of bacteria (1990 revision). Washington, DC: American Society for Microbiology; 1992. p. 295.

Ludwig W, Klenk HP. Overview: a phylogenetic backbone and taxonomic framework for prokaryotic systematics. In: Boone DR, Castenholz RW, Garrity GM, editors. Bergey's manual of systematic bacteriology. 2nd ed. New York: Springer; 2001. p. 49–65.

Ludwig W, Schleifer KH. Bacterial phylogeny based on 16S and 23S rRNA sequence analysis. FEMS Microbiol Rev. 1994;15:155–73.

Ludwig W, Strunk O, Westram R, et al. ARB: a software environment for sequence data. Nucleic Acids Res. 2004;32:1363–71.

Mizrahi-Man O, Davenport ER, Gilad Y. Taxonomic classification of bacterial 16S rRNA genes using short sequencing reads: evaluation of effective study designs. PLoS One. 2013;8:e53608.

Munoz R, Yarza P, Ludwig W, et al. Release LTPs104 of the all-species living tree. Syst Appl Microbiol. 2011;34:169–70.

Pruesse E, Quast C, Knittel K, et al. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Res. 2007;35:7188–96.

Santamaria M, Fosso B, Consiglio A, et al. Reference databases for taxonomic assignment in metagenomics. Brief Bioinform. 2012;13:682–95.

Stackebrandt E, Ebers J. Taxonomic parameters revisited: tarnished gold standards. Microbiol Today. 2006;33:152–5.

Stackebrandt E, Frederiksen W, Garrity GM, et al. Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. Int J Syst Evol Microbiol. 2002;52:1043–7.

Stamatakis A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics. 2006;22:2688–90.

Yarza P, Richter M, Peplies J, et al. The all-species living tree project: a 16S rRNA-based phylogenetic tree of all sequenced type strains. Syst Appl Microbiol. 2008;31:241–50.

Yarza P, Ludwig W, Euzéby J, et al. Update of the all-species living tree project based on 16S and 23S rRNA sequence analyses. Syst Appl Microbiol. 2010;33:291–9.

Yarza P, Spröer C, Swiderski J, et al. Sequencing Orphan Species initiative (SOS): filling the gaps in the 16S rRNA gene sequence database for all species with validly published names. Syst Appl Microbiol. 2013;36:69–73.

antiSMASH

Eriko Takano[1], Rainer Breitling[1] and
Marnix H. Medema[2]
[1]Manchester Institute of Biotechnology, University of Manchester, Manchester, UK
[2]Microbial Genomics and Bioinformatics Research Group, Max Planck Institute for Marine Microbiology, Bremen, Germany

Definition

antiSMASH (Medema et al. 2011) is a web server and a stand-alone software to identify, annotate, and compare gene clusters that encode the biosynthesis of secondary metabolites in bacterial and fungal genomes. antiSMASH offers a wide range of options to identify and analyze biosynthetic gene clusters, including protein domain analysis of the large multi-domain enzymatic assembly lines involved, prediction of core chemical structures of their end compounds, and multiple cluster alignments to a database of all currently sequenced gene clusters.

The antiSMASH web server can be found at http://antismash.secondarymetabolites.org.

Introduction

Microbial secondary metabolites are of great interest to society because of their diverse biological activities that are interesting starting points for drug development. Many of them are already used as antibiotics, antitumor agents, or cholesterol-lowering drugs (Hutchinson and McDaniel 2001; Fischbach and Walsh 2009). Automated computational identification of gene clusters in newly sequenced genomes is becoming a cornerstone of genome-based drug discovery, due to the affordability of sequencing large numbers of genomes from microorganisms that potentially produce novel secondary metabolites (Walsh and Fischbach 2010).

Functionalities

Gene Cluster Detection

antiSMASH detects a wide range of different types of biosynthetic gene clusters, including those encoding the pathways toward polyketides (PKs), nonribosomal peptides (NRPs), terpenoids, ribosomal peptides, aminoglycosides, and non-NRP siderophores. The detection is performed by screening the gene sequences from the input against a library of profile Hidden Markov Models (pHMMs) (Eddy 2011), each of which is specific for genes characteristic for a certain gene cluster type, and passing the results through a hierarchical logic filter. A second detection algorithm is also run, which detects genomic regions that are enriched in Pfam domains (Finn et al. 2010) linked to secondary metabolism.

Protein Domain Analysis of Polyketide Synthases and Nonribosomal Peptide Synthetases

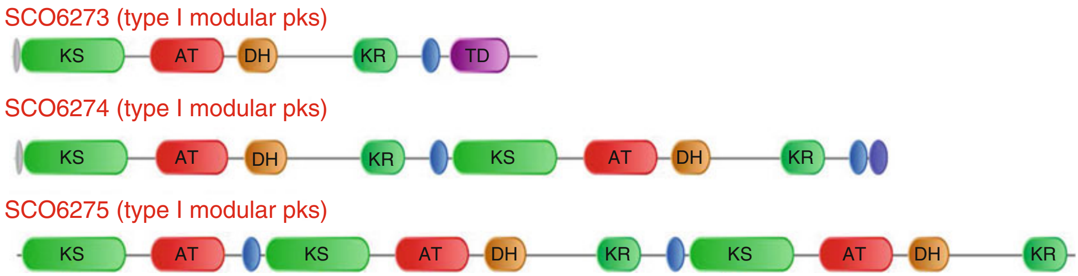
PKs and NRPs are synthesized by large megasynthase enzymes containing a multitude of protein domains, such as condensation (C) and adenylation (A) and PCP-binding domains in nonribosomal peptide synthetases (NRPSs), ketosynthase (KS), and acyltransferase (AT) and ACP-binding domains in polyketide synthases (PKSs) (Fischbach and Walsh 2006). antiSMASH contains a library of pHMMs that can recognize all these protein domains as well as distinguish between various subtypes of these domains. In the antiSMASH output, the domain structures of any NRPSs or PKSs encoded in a gene cluster are visualized, and several downstream analysis options are provided for each domain (Fig. 1).

Core Chemical Structure Prediction

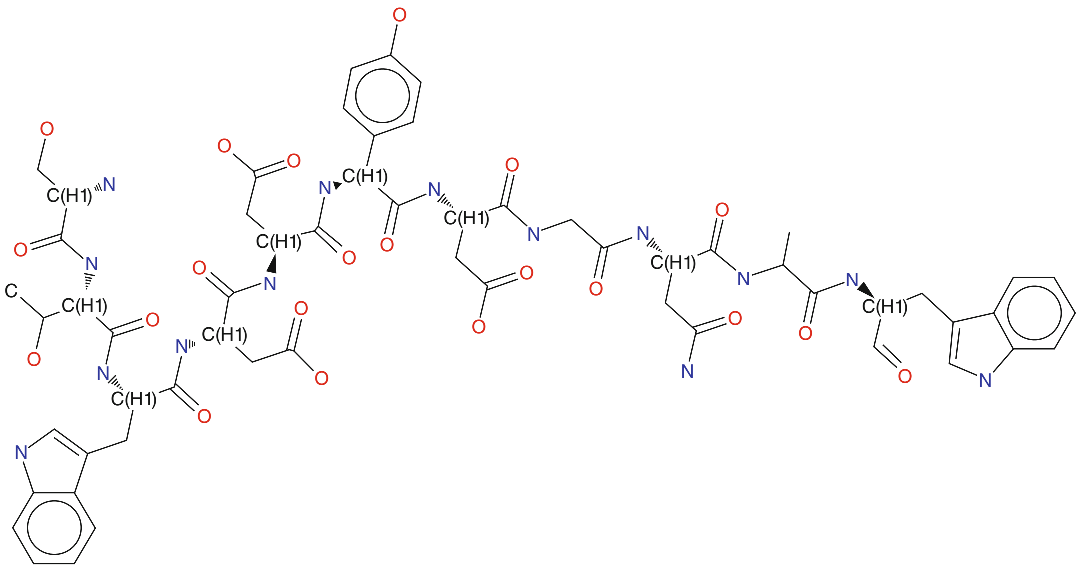
When a secondary metabolite biosynthesis gene cluster is detected, one of the key questions of course is what kind of chemical structure it produces. For NRPs and PKs, antiSMASH is able to already give a first approximation of the core chemical structure of the end compound (Fig. 2). To do so, it uses several substrate specificity prediction methods (Yadav et al. 2003; Minowa et al. 2007; Röttig et al. 2011) that are based on the amino acid sequences of the A domains of NRPSs and the AT domains of PKSs. To infer the sequential arrangement of the predicted substrates of the A/AT domains in the resulting polyketide or peptide, the order of the PKS enzymes in a multimodular assembly line is predicted using their estimated docking domain binding affinities (Yadav et al. 2009) or, alternatively, colinearity of the PKS or NRPS genes with their enzymes is assumed.

Comparative Analysis of Gene Clusters

In order to understand the architecture and function of a secondary metabolite biosynthesis gene cluster, much is gained by examining it within its evolutionary context through the



antiSMASH, Fig. 1 Domain structure of multi-domain enzymes such as PKSs and NRPSs as visualized by antiSMASH, offering several options for analysis when the mouse is positioned over a domain: one can, for example, run a BlastP search specifically with the sequence of this domain

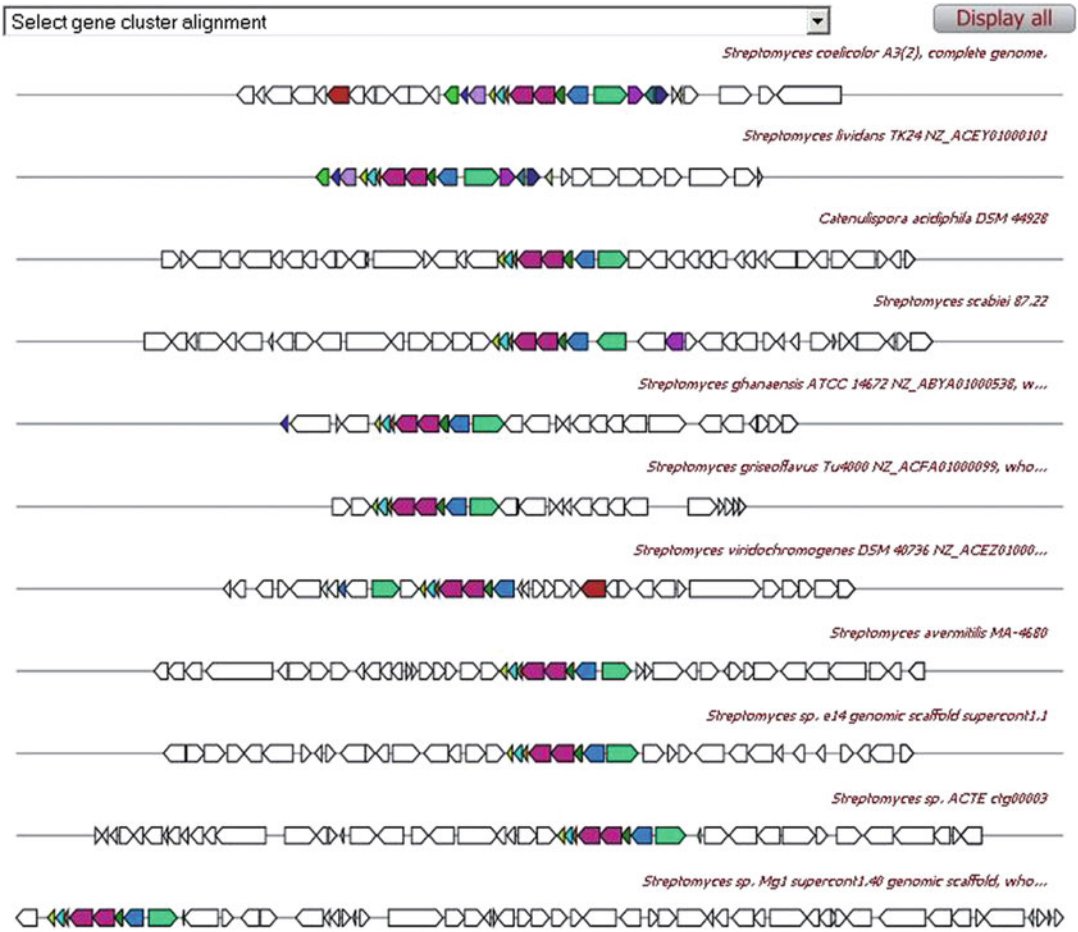


antiSMASH, Fig. 2 Prediction of the core chemical structure of an NRP by antiSMASH. The residues are based on a consensus between three prediction methods for the substrate specificities of the NRPS adenylation domains in the gene cluster

comparison with related gene clusters from species across the tree of life. To facilitate this, antiSMASH hosts a regularly updated database of gene clusters it has detected in all nucleotide sequences present in GenBank. antiSMASH then combines multiple BlastP runs into a comparative search of every identified gene cluster against all other known gene clusters. This is used to generate a multiple gene cluster alignment (Fig. 3), which can aid the biologist in assessment of the novelty of the gene cluster, detecting the borders of the gene cluster and identifying the conserved multigene modules that constitute its building blocks.

Secondary Metabolism-Specific Gene Family Analysis

Most genes involved in the biosynthesis of secondary metabolite have (close) homologues with similar functions in other secondary metabolite biosynthesis gene clusters. This can be used to infer the functions of the genes



antiSMASH, Fig. 3 Example of a multiple gene cluster alignment by antiSMASH, showing identified homologue clusters of the query gene cluster

residing in the biosynthetic gene cluster based on sequence homology. antiSMASH simplifies this process by categorizing the genes of every identified gene cluster into secondary metabolism-specific gene families and automatically generating approximate phylogenetic trees of each gene in the context of its gene family.

Genome-Wide Pfam and Blast Analysis

Finally, antiSMASH also offers the possibility (transferred from CLUSEAN; Weber et al. 2009) to do a comprehensive analysis of all genes within a submitted genome, identifying Pfam matches and running Blast for each gene against a database of all bacterial and fungal protein sequences.

Stand-Alone Version

Stand-alone versions of antiSMASH are available for download for Windows, Mac OS X, and Ubuntu Linux. Additionally, several related scripts are available from the antiSMASH website. An EMBL formatting script can be downloaded to format raw FASTA sequences together with a text file containing gene annotations into an EMBL file that can be submitted to antiSMASH. Also, a script is available which allows running antiSMASH on multiple files, in batch mode.

Development

antiSMASH is still under active development. Some features projected for the next release are batch input on the web server, protein sequence input, and subclass prediction for enzyme classes like terpene synthases and trans-AT PKSs. Feature requests, bug reports, or other questions/suggestions can be sent to the development team via the online contact form on the antiSMASH website.

Related Tools

Several other software tools for the study of secondary metabolism have been published. For example, ClustScan (Starcevic et al. 2008) and NP.searcher (Li et al. 2009) can both be used to detect bacterial polyketide and NRP biosynthesis gene clusters. The same is the case for CLUSEAN (Weber et al. 2009), the pipeline which has now been integrated entirely into antiSMASH. For the analysis of fungal sequences, SMURF (Khaldi et al. 2010) offers a gene cluster detection potential similar to that of antiSMASH. Structural analysis of polyketide synthases can be performed with the SBSPKS suite (Anand et al. 2010). Finally, draft genomes with many small contigs and metagenomes with fragments too small for gene cluster detection can be scrutinized with NaPDoS (Ziemert et al. 2012) in order to find protein domains related to secondary metabolite biosynthesis and analyze these phylogenetically.

Summary

antiSMASH is an easy-to-use web server for the detection of secondary metabolite biosynthesis gene clusters. Various functionalities – comparative, phylogenomic, enzymatic, etc. – are integrated in one single pipeline, making it straightforward for genomicists and natural product researchers to study the biosynthetic potential of any organism.

Cross-References

▶ Bacteriocin Mining in Metagenomes
▶ CLUSEAN, Overview
▶ Mining Metagenomic Datasets for Antibiotic Resistance Genes
▶ Phylogenetics, Overview

References

Anand S, Prasad MV, Yadav G, Kumar N, Shehara J, Ansari MZ, Mohanty D. SBSPKS: structure based sequence analysis of polyketide synthases. Nucleic Acids Res. 2010;38:W487–96.

Eddy SR. Accelerated profile HMM searches. PLoS Comput Biol. 2011;7:e1002195.

Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, et al. The Pfam protein families database. Nucleic Acids Res. 2010;38: D211–22.

Fischbach MA, Walsh CT. Assembly-line enzymology for polyketide and nonribosomal peptide antibiotics: logic, machinery, and mechanisms. Chem Rev. 2006;106:3468–96.

Fischbach MA, Walsh CT. Antibiotics for emerging pathogens. Science. 2009;325:1089–93.

Hutchinson CR, McDaniel R. Combinatorial biosynthesis in microorganisms as a route to new antimicrobial, antitumor and neuroregenerative drugs. Curr Opin Investig Drugs. 2001;2:1681–90.

Khaldi N, Seifuddin FT, Turner G, Haft D, Nierman WC, Wolfe KH, Fedorova ND. SMURF: genomic mapping of fungal secondary metabolite clusters. Fungal Genet Biol. 2010;47:736–41.

Li MH, Ung PM, Zajkowski J, Garneau-Tsodikova S, Sherman DH. Automated genome mining for natural products. BMC Bioinformatics. 2009;10:185.

Medema MH, Blin K, Cimermancic P, de Jager V, Zakrzewski P, Fischbach MA, Weber T, Takano E, Breitling R. antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. Nucleic Acids Res. 2011;39:W339–46.

Minowa Y, Araki M, Kanehisa M. Comprehensive analysis of distinctive polyketide and nonribosomal peptide structural motifs encoded in microbial genomes. J Mol Biol. 2007;368:1500–17.
Röttig M, Medema MH, Blin K, Weber T, Rausch C, Kohlbacher O. NRPSpredictor2 – a web server for predicting NRPS adenylation domain specificity. Nucleic Acids Res. 2011;39:W362–7.
Starcevic A, Zucko J, Simunkovic J, Long PF, Cullum J, Hranueli D. ClustScan: an integrated program package for the semi-automatic annotation of modular biosynthetic gene clusters and in silico prediction of novel chemical structures. Nucleic Acids Res. 2008;36:6882–92.
Walsh CT, Fischbach MA. Natural products version 2.0: connecting genes to molecules. J Am Chem Soc. 2010;132:2469–93.
Weber T, Rausch C, Lopez P, Hoof I, Gaykova V, Huson DH, Wohlleben W. CLUSEAN: a computer-based framework for the automated analysis of bacterial secondary metabolite biosynthetic gene clusters. J Biotechnol. 2009;140:13–7.
Yadav G, Gokhale RS, Mohanty D. Computational approach for prediction of domain organization and substrate specificity of modular polyketide synthases. J Mol Biol. 2003;328:335–63.
Yadav G, Gokhale RS, Mohanty D. Towards prediction of metabolic products of polyketide synthases: an in silico analysis. PLoS Comput Biol. 2009;5:e1000351.
Ziemert N, Podell S, Penn K, Badger JH, Allen E, Jensen PR. The natural product domain seeker NaPDoS: a phylogeny based bioinformatic tool to classify secondary metabolite gene diversity. PLoS One. 2012;7:e34064.

Approaches in Metagenome Research: Progress and Challenges

Heiko Nacke and Rolf Daniel
Institute of Microbiology and Genetics, Georg–August–University of Göttingen, Göttingen, Germany

Synonyms

Function-based screening, Metagenomic biomolecule, Metagenomic library, Metagenomics, Next-generation sequencing, Sequence-based screening

Definition

Metagenomics comprises the culture-independent and DNA-based analysis of entire microbial communities and complements cultivation-based analysis of microorganisms. Metagenomic approaches allow comprehensive insights into phylogenetic and functional diversity of complex microbial consortia present in moderate as well as extreme environments on Earth. The introduction of next-generation sequencing technologies enabled cost-effective high-throughput sequencing of metagenomic DNA molecules resulting in increased resolution of microbial community analysis. In addition, screening of metagenomic libraries led to the identification of numerous novel biomolecules from various environments such as soil, seawater, or glacial ice.

Introduction

The immensely manifold microbial niches on Earth comprise an extraordinarily high abundance and diversity of prokaryotic and eukaryotic microorganisms. The human body is colonized by a wide variety of microbes representing all three domains of life. The entirety of these microbial cells (the human microbiome) that is often described as an additional organ exceeds the number of human cells by at least an order of magnitude and outnumbers human genes by more than 100-fold (Li et al. 2012; Weinstock 2012). Also in extreme environments such as hydrothermal vents, sea ice, or deep inside the Earth's crust, various microorganisms could be detected. For example, a phylogenetically diverse and metabolically active microbial assemblage was identified in the brine of an ice-sealed Antarctic lake (Murray et al. 2012). The microorganisms existing in this aphotic ecosystem withstand a temperature of -13 °C, anoxic conditions, and high salinity.

Currently, less than 1 % of the microorganisms on Earth are readily culturable under laboratory conditions. To investigate the high percentage of uncultured microbes, different metagenomic approaches can be routinely applied. Metagenomics allows the direct study of the collective genomes present in microbial ecosystems (Handelsman 2004). This approach significantly expanded our knowledge on microbial phylogenetic and functional diversity and enabled the discovery of numerous previously unknown biomolecules. In the recent history of metagenomics, especially next-generation sequencing techniques, allowing cost-effective and rapid decoding of metagenomic DNA, were applied to analyze microbial populations. As a consequence, a number of bioinformatic tools to evaluate and compare comprehensive high-throughput metagenomic data have been developed in the last few years.

In this review, an overview of traditional and recent metagenomic research approaches, associated future challenges, and a short description of related meta-omic studies will be given.

Microbial Phylogenetic and Functional Diversity Determination

Small-subunit rRNA genes, universally distributed across prokaryotic and eukaryotic organisms, can be considered as evolutionary clocks enabling phylogenetic analysis. Most commonly, metagenome-derived 16S rRNA and 18S rRNA genes are used to phylogenetically characterize microbial communities. Furthermore, other conserved genes such as *recA*, *rpoB*, *HSP70*, or *EF-Tu* allow phylogenetic assignments (Ludwig and Klenk 2001). These genes can be investigated by applying traditional molecular approaches including fingerprinting methods such as denaturing gradient gel electrophoresis and terminal restriction fragment length polymorphism analysis or Sanger sequencing. A significant drawback of the Sanger sequencing-based analysis of microbial communities is the time-consuming and labor-intensive nature of this approach, as well as the required construction of clone libraries.

More recently, next-generation sequencing platforms were used to decode metagenomic DNA. Currently, the following next-generation sequencing technologies are available: sequencing by ligation (SOLiD – Applied Biosciences/Life Technologies), sequencing by synthesis (Solexa/Illumina), semiconductor chip sequencing (Ion Torrent/Life Technologies), pyrosequencing (454/Roche), and single-molecule sequencing (Oxford Nanopore Technologies, SMRT – Pacific Biosciences). Compared to Sanger sequencing, these cloning-independent techniques allow the generation of far more sequence data per run. Thus, microbial diversity comparisons between different environmental samples, requiring replicated data and statistical analysis, as well as deep analysis of highly complex microbial community structures, are possible. Currently, often tens to hundreds of thousands partial metagenomic small-subunit rRNA gene sequences are produced using next-generation sequencing platforms. In a recent pyrosequencing-based 16S rRNA gene survey, a total of 41,141 bacterial and 30,651 archaeal sequences were analyzed to investigate prokaryotic diversity in Yunnan and Tibetan hot springs (Song et al. 2013). To (pre-)process small-subunit rRNA gene sequence datasets, various tools, software packages, analytical web servers, and virtual instances can be used (Gonzalez and Knight 2012). The QIIME package (Caporaso et al. 2010) provides workflows to extensively analyze high-throughput amplicon-based sequence data starting with raw sequences. Nevertheless, the avoidance of marker gene amplification bias by applying direct sequencing of metagenomic DNA instead of amplicon-based sequencing allows the most exact taxonomic assessment (Simon and Daniel 2011). For further improvement of microbial diversity and abundance estimation, Kembel et al. (2012) recently introduced an approach, which incorporates 16S rRNA gene copy number information.

To identify the taxonomic affiliation of all sequences derived from metagenomic DNA, a process called binning can be carried out. Within binning procedures, sequences of a metagenomic dataset sharing the same taxonomic origin are “binned” (grouped). Composition-based binning is based on conserved genomic features such as dinucleotide frequencies, GC content, and synonymous codon usage, whereas similarity-based binning makes use of sequence homology. Among others, PhyloPythiaS, introduced by Patil et al. (2011), represents an appropriate application to perform composition-based binning. With respect to similarity-based binning, typically searches against reference databases (e.g., National Center for Biotechnology Information databases) are performed using alignment tools such as BLAST+ (Camacho et al. 2009). Subsequently, BLAST results can be interpreted by applying software such as MEGAN (Huson et al. 2011).

Due to the often very high diversity of microbial communities, assembly of metagenome-derived sequences is challenging. In a recent metagenomic survey of honey bee gut microbiota, de novo assembly of 81,343,096 Illumina paired-end reads resulted in 54,700 scaffolds of contigs (total length, 76.6 Mb) (Engel et al. 2012). Similar to the approach conducted by Engel et al. (2012), single-genome assemblers were used for metagenome assembly with modified settings. Recently, a single-genome assembler (Velvet) has been extended to enable the assembly of short metagenomic reads (Namiki et al. 2012). This new de novo assembler (MetaVelvet) generated significantly higher N50 scores, a parameter that evaluates assembly quality, than analyzed single-genome assemblers for simulated datasets.

Based on assemblies or individual metagenomic sequence reads, gene prediction, annotation, and reconstruction of pathways can be carried out to assess the functional potential encoded by metagenomes. Consecutive processing of these steps is provided by a number of web-based tools like MG-RAST (Meyer et al. 2008). These tools utilize resources of reference databases such as SEED (Overbeek et al. 2005) and KEGG (Kanehisa et al. 2008) to link biological information to predicted genes. In a recent survey including metagenomic methods, the functional potential of Arctic *Thaumarchaeota* was investigated (Alonso Sáez et al. 2012). By analyzing a metagenome derived from a Southeast Beaufort Sea sample collected during Arctic winter, Alonso Sáez et al. (2012) identified thaumarchaeal pathways for ammonia oxidation. A number of other *Thaumarchaeota* are also capable of ammonia oxidation, but unexpectedly these Arctic thaumarchaeal organisms harbored a high abundance of genes involved in urea transport and degradation.

Metagenomic Biomolecule Discovery

To access the large pool of unexplored biomolecules, microbial community DNA has been extracted and metagenomic libraries have been constructed. Small-insert and large-insert metagenomic libraries can be screened to identify novel biomolecules. For the construction of small-insert libraries containing metagenomic DNA $\leq$ 15 kb, plasmids are appropriate vectors, whereas cosmids, fosmids, and bacterial artificial chromosomes (BACs) can be used for cloning of large metagenomic DNA molecules (cosmids and fosmids, $\leq$40 kb; BACs, 100–200 kb). Metagenomic libraries from different microbial habitats such as glacier ice, digestive tracts of animals, soil, hot springs, or seawater have already been constructed and successfully screened for novel biomolecules (see, e.g., Nacke et al. 2012). Some of these biomolecules exhibit valuable characteristics for industrial applications such as thermal stability, halotolerance, and activity under acidic or alkaline conditions. In a recent metagenomic approach, Sulaiman et al. (2012) isolated a gene encoding a novel cutinase homolog designated LC-cutinase with polyethylene terephthalate-degrading activity from a leaf-branch compost fosmid library. The enzyme showed higher specific polyethylene terephthalate-degrading activity than previously reported bacterial and fungal cutinases. Thus, LC-cutinase is a potent candidate for industrial applications, i.e., in textile industry. In general, two different metagenomic screening approaches for the identification of novel biomolecules can be distinguished: function-based screening and sequence-based screening.

Principle and Variations of Function-Driven Screens

To perform function-driven screening, the construction of small-insert or large-insert metagenomic libraries is required. A broad array of different function-based screening approaches can be applied using these libraries. The phenotypic insert detection (PID) is the most frequently applied screening strategy. Metagenomic library-containing clones expressing target genes are identified based on phenotypic characteristics. This method has been applied to identify novel lipolytic genes and gene families from German forest and grassland soil samples using tributyrin as a screening substrate (Nacke et al. 2011). A total of 37 lipolytic clones, encoding novel lipases and esterases, which could be assigned to five different known families and two putatively new families of lipolytic enzymes, were identified by halo formation on indicator agar plates. The potential to identify entirely novel target genes is an important advantage of function-driven screening approaches. Modulated detection (MD) represents another commonly applied strategy to perform function-based screening. Only if a certain gene product is expressed by a metagenomic library-containing host strain, it can grow under selective conditions. Recently, novel acid resistance genes were derived from planktonic and rhizosphere microbial communities of the Tinto River (Spain) using this strategy (Guazzaroni et al. 2013). Fifteen genes, mainly encoding putative proteins of unknown function, conferred acid resistance to the host strain *Escherichia coli*. Moreover, substrate-induced gene expression (SIGEX), product-induced gene expression (PIGEX), and metabolite-regulated expression (METREX) screening strategies allow the identification of target genes from metagenomic libraries (Simon and Daniel 2009). Recently, Wang et al. (2012) suggested biosensor-based genetic transducer (BGT) systems as an alternative and sensitive approach to screen for gene clusters whose expression produce small molecules that activate the employed biosensors. Nevertheless, all of these function-based screening approaches share one significant disadvantage: the dependence of target gene production on the expression machinery of the metagenomic library host.

Principle and Variants of Sequence-Based Screening

Conserved regions of genes or proteins enable sequence-driven screening approaches. Based on these regions degenerate primers can be designed and fragments of target genes amplified. For example, novel biphenyl dioxygenase DNA segments encoding active site residues were obtained from polychlorobiphenyl-contaminated soils using this strategy (Standfuß-Gabisch et al. 2012). After sequencing of an amplified partial target gene, it can be decoded completely using primer walking and extracted environmental DNA or a metagenomic library as a template. In this way, an entire xylose isomerase gene (*xym1*) has been derived from a soil metagenomic library (Parachin and Gorwa-Grauslund 2011). The gene product of *xym1* consisted of 443 amino acids and was most similar (83 % identity) to a xylose isomerase from *Sorangium cellulosum*. Additionally, novel complex polyketide and nonribosomal peptide biosynthesis gene cluster that often exceed average insert sizes of large-insert metagenomic libraries can be discovered by using degenerate primers and subsequent chromosome walking (Piel 2011). The potential to identify genes of interest even if they are not expressed in a metagenomic library host represents a major advantage of sequence-based screening, but only novel variants of already-known gene or protein families can be detected by this method.

Future Challenges in Metagenomic Research and Related Meta-omic Approaches

One of the major requirements to combine and compare metagenomic studies conducted by research groups worldwide is the definition and acceptance of minimum standards in experimental design. The same applies to metatranscriptomics, metaproteomics, and metabolomics. In this way, comparison and combination of results obtained from the different meta-omic approaches are feasible. Metatranscriptomics, metaproteomics, and metabolomics comprise the study of the collective gene transcripts, expressed proteins, and metabolites, respectively, generated by the microorganisms within an ecosystem (Nacke et al. 2014; Hettich et al. 2012; Patti et al. 2012). The consequent application and combination of appropriate meta-omic approaches will lead to an enormous extension of knowledge on the gene structure, diversity, activity, and responses of microbial communities on an ecosystem level. Furthermore, the rapid growth of meta-omic technologies will continuously demand for progress in the field of bioinformatics. Thus, further development and linkage of meta-omic analysis tools will be important in the future. In addition, the application and improvement of culture-based methods will be still valuable in the future to extend the number of available reference genomes allowing mapping of metagenomic data. In this context, the young discipline of single cell genomics has potential to play a complementary role by continuously contributing novel reference genomes.

Summary

The introduction of metagenomics allowed culture-independent analysis of microbial populations in complex ecosystems. Subsequently, other culture-independent meta-omic disciplines including metatranscriptomics, metaproteomics, and metabolomics were established. Metagenomics provided insights into the enormous phylogenetic and functional diversity of microbial communities within various environments on Earth. The increasing number of next-generation sequencing technologies led to a more comprehensive and cost-effective assessment of the information encoded by metagenomic DNA. Metagenomic approaches comprising the construction and screening of metagenomic libraries resulted in identification of previously unknown biomolecules, including biomolecules with industrially relevant characteristics.

Cross-References

▶ A 123 of Metagenomics
▶ Extraction Methods, Variability Encountered in
▶ Fosmid System
▶ Genome Portal, Joint Genome Institute
▶ Microbial Diversity, Bar-Coding Approaches
▶ Microbial Ecosystems, Protection of
▶ Phylogenetics, Overview

References

Alonso Sáez L, Waller AS, Mende DR, et al. Role for urea in nitrification by polar marine Archaea. Proc Natl Acad Sci USA. 2012;109:17989–94.

Camacho C, Coulouris G, Avagyan V, et al. BLAST+: architecture and applications. BMC Bioinforma. 2009;10:421.

Caporaso JG, Kuczynski J, Stombaugh J, et al. QIIME allows analysis of high-throughput community sequencing data. Nat Methods. 2010;7:335–6.

Engel P, Martinson VG, Moran NA. Functional diversity within the simple gut microbiota of the honey bee. Proc Natl Acad Sci USA. 2012;109:11002–7.

Gonzalez A, Knight R. Advancing analytical algorithms and pipelines for billions of microbial sequences. Curr Opin Biotechnol. 2012;23:64–71.

Guazzaroni ME, Morgante V, Mirete S, et al. Novel acid resistance genes from the metagenome of the Tinto River, an extremely acidic environment. Environ Microbiol. 2013;15:1088–1102.

Handelsman J. Metagenomics: application of genomics to uncultured microorganisms. Microbiol Mol Biol Rev. 2004;68:669–85.

Hettich RL, Sharma R, Chourey K, et al. Microbial metaproteomics: identifying the repertoire of proteins that microorganisms use to compete and cooperate in complex environmental communities. Curr Opin Microbiol. 2012;15:373–80.

Huson DH, Mitra S, Ruscheweyh HJ, et al. Integrative analysis of environmental sequences using MEGAN4. Genome Res. 2011;21:1552–60.

Kanehisa M, Araki M, Goto S, et al. KEGG for linking genomes to life and environment. Nucleic Acids Res. 2008;36:D480–4.

Kembel SW, Wu M, Eisen JA, et al. Incorporating 16S gene copy number information improves estimates of

microbial diversity and abundance. PLoS Comput Biol. 2012;8:e1002743.
Li K, Bihan M, Yooseph S, Methé BA. Analyses of the microbial diversity across the human microbiome. PLoS ONE. 2012;7:e32118.
Ludwig W, Klenk HP. Overview: a phylogenetic backbone and taxonomic framework for procaryotic systematics. In: Garrity GM, Boone DR, Castenholz RW, editors. Bergey's manual of systematic bacteriology, Vol. 1. 2nd ed. New York: Springer; 2001. p. 49–65.
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 Bioinforma. 2008;9:386.
Murray AE, Kenig F, Fritsen CH, et al. Microbial life at −13 °C in the brine of an ice-sealed Antarctic lake. Proc Natl Acad Sci USA. 2012;109:20626–31.
Nacke H, Will C, Herzog S, et al. Identification of novel lipolytic genes and gene families by screening of metagenomic libraries derived from soil samples of the German biodiversity exploratories. FEMS Microbiol Ecol. 2011;78:188–201.
Nacke H, Engelhaupt M, Brady S, et al. Identification and characterization of novel cellulolytic and hemicellulolytic genes and enzymes derived from German grassland soil metagenomes. Biotechnol Lett. 2012;34:663–75.
Nacke H, Fischer C, Thürmer A, et al. Land use type significantly affects microbial gene transcription in soil. Microb Ecol. 2014;67:919–30.
Namiki T, Hachiya T, Tanaka H, et al. MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. Nucleic Acids Res. 2012;40:e155.
Overbeek R, Begley T, Butler RM, et al. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Res. 2005;33:5691–702.
Parachin NS, Gorwa-Grauslund MF. Isolation of xylose isomerases by sequence- and function-based screening from a soil metagenomic library. Biotechnol Biofuels. 2011;4:9.
Patil KR, Haider P, Pope PB, et al. Taxonomic metagenome sequence assignment with structured output models. Nat Methods. 2011;8:191–2.
Patti GJ, Yanes O, Siuzdak G. Innovation: Metabolomics: the apogee of the omics trilogy. Nat Rev Mol Cell Biol. 2012;13:263–69.
Piel J. Approaches to capturing and designing biologically active small molecules produced by uncultured microbes. Annu Rev Microbiol. 2011;65:431–53.
Simon C, Daniel R. Achievements and new knowledge unraveled by metagenomic approaches. Appl Microbiol Biotechnol. 2009;85:265–76.
Simon C, Daniel R. Metagenomic analyses: past and future trends. Appl Environ Microbiol. 2011;77:1153–61.
Song ZQ, Wang FP, Zhi XY, et al. Bacterial and archaeal diversities in Yunnan and Tibetan hot springs, China. Environ Microbiol. 2013;15:1160–75.
Standfuß-Gabisch C, Al-Halbouni D, Hofer B. Characterization of biphenyl dioxygenase sequences and activities encoded by the metagenomes of highly polychlorobiphenyl-contaminated soils. Appl Environ Microbiol. 2012;78:2706–15.
Sulaiman S, Yamato S, Kanaya E, et al. Isolation of a novel cutinase homolog wit polyethylene terephthalate-degrading activity from leaf-branch compost by using a metagenomic approach. Appl Environ Microbiol. 2012;78:1556–62.
Wang Y, Chen Y, Zhou Q, et al. A culture-independent approach to unravel uncultured bacteria and functional genes in a complex microbial community. PLoS ONE. 2012;7:e47530.
Weinstock GM. Genomic approaches to studying the human microbiota. Nature. 2012;489:250–6.

Arbuscular Mycorrhizal Fungi Assemblages in Chernozems

Chantal Hamel, Luke D. Bainard and Mulan Dai
Semiarid Prairie Agricultural Research Centre, Agriculture and Agri-Food Canada, Swift Current, SK, Canada

Synonyms

Diversity, arbuscular mycorrhizal fungi, Canadian Prairie, Chernozem, land use.

Definition

AM fungi are obligate plant symbionts that form the phylum Glomeromycota. These fungi contribute to plant nutrient uptake, influence soil biotic and abiotic environments, and provide important ecosystem services. 454-pyrosequencing of amplicons from metagenomic DNA revealed the distribution of AM fungi in major Canadian Chernozem great groups as influenced by land use and crop management.

Introduction

AM fungi form a mycorrhizal symbiosis with the roots of the majority of land plants. They have coevolved with plants over 450 Ma to produce today's mycorrhiza, which is an organ specialized in the extraction of soil nutrients. As such, AM fungi are seen as a key stone of agricultural sustainability (Garg and Chandel 2010).

World grain, pulse, and biofuel crop production mainly occurs on deep (typically >18–25 cm) warm-colored soils rich in humus (>0.6 % organic carbon) and weatherable minerals, with high levels of base saturation (>50 %) and calcium as the main exchangeable cation (Durán et al. 2011). These soils have similar properties but have different names in other soil classification systems. They are Chernozems in Canada, Ukraine, and Russia; Mollisols in the USA and South America; Isohumosols or Black Soils in China; and Chernozems, Kastanozems, and Phaeozems according to the FAO (Liu et al. 2012). These soils have typically developed under condition of moisture deficit and grassland vegetation in temperate regions around the globe. They mainly occur in a band across Eastern Europe and Central Asia, in northeast China, from south-central Canada down to the Gulf of Mexico, and over most of Uruguay and part of Argentina.

Tackling the Complexity of Soil Biodiversity

Soil hosts an extremely high level of microbial diversity (Young and Crawford 2004). However, high-throughput next-generation sequencing now allows generation of the massive sequence data required to characterize soil microbial diversity.

Amplicon sequencing is preferred over whole genome sequencing for the study of the taxonomic diversity of targeted microbial groups. The 454 FLX and 454 FLX + technologies allow the sequencing of DNA amplicons up to 400 and 800 bp in length, respectively. Such long sequences contain sufficient taxonomic information for the characterization of microbial communities and their use conveniently eliminates the need for sequence assembly.

Pyrosequencing of amplicons and bioinformatic analysis of sequence data yield the profile of operational taxonomic units (OTU) of the target microbial group in a soil sample. The concept of an OTU is useful in soil microbiology as the majority of microbial species are still undescribed. OTUs serve as a proxy for species making it possible to measure and describe soil microbial diversity. In addition, OTUs can be identified by comparison with known sequences in public databases such as GenBank and MaarjAM. AM fungi have been difficult to study due to their obligate biotrophy and inability to grow in pure culture. However, polymerase chain reaction (PCR) made possible the amplification of DNA from their spores and enabled the molecular characterization and classification of taxa within the Glomeromycota (Schuessler 2013).

Fungal diversity is commonly assessed based on the internal transcribed spacer (ITS) of the ribosomal RNA gene. However, abundant SSU rRNA gene sequences of AM fungi are found in databases due to the traditional use of this region for the Glomeromycota. Several primers sets producing taxonomically informative amplicons short enough for use with first- and next-generation molecular techniques have been used in ecological studies of AM fungi.

The AM fungi have a patchy distribution in soil (Hart and Klironomos 2003). Thus in order to capture their diversity, multiple samples must be taken at a study site. A composite sample is usually made by pooling and homogenizing all the samples from a sampling site. The distribution of organisms varies with soil depth, thus sampling depth also matters. The AM fungi are normally found within the rooting depth.

Arbuscular Mycorrhizal Fungi in the Canadian Chernozems

AM fungal communities in the Canadian Prairie Chernozem soils are composed of a few dominant and a large number of subordinate taxa. Less than 6 % of the AM fungal OTUs accounted for half of all AM fungal reads (Dai et al. 2013). Across the Canadian prairie landscape, the Glomeraceae were the most abundant family, accounting for

65 % of all AM fungal OTUs and 54 % of the AM fungal reads. The Claroideoglomeraceae is second in abundance with 25 % of all AM fungal OTUs and 39 % of the AM fungal reads. Diversisporaceae accounted for 8 % of the OTUs and 7 % of the AM fungal reads. Paraglomaceae, Gigasporaceae, and Archeosporaceae are poorly distributed across the prairie landscape, and Gigasporaceae and Archeosporaceae are rare.

In other regions, spore counts in grazed Kastanozems of Inner Mongolia revealed that the AM fungal communities resembled those observed in Canadian Chernozems (Tian et al. 2009). The Gigasporaceae are susceptible to disturbance and largely absent in croplands, which explains their greater abundance in the Kastanozems than in the Canadian Prairie Chernozems (Dai et al. 2012, 2013). Poorer AM fungal diversity is reported from American spore-based surveys of Mollisols under tallgrass prairie cover where Paraglomaceae and Archeosporaceae were undetected (Eom et al. 2001; Bentivenga and Hetrick 1992). Tallgrass prairies managed with fire were found to be very highly dominated by the Glomeraceae (Bentivenga and Hetrick 1992), underlining the importance of land use in the structuring of AM fungal communities.

AM fungi share root occupation with fungal endophytes belonging to different taxonomic groups. Non-AM fungal endophytes are particularly abundant in temperate grasslands (Porras-Alfaro et al. 2011). This observation triggered the question as to whether AM fungi are at the end of their range in dry areas.

This hypothesis was explored in the Canadian Prairie using primers Glo1/NS31, which produced 18S rDNA amplicons of about 230 bp (Yang et al. 2010). A succession of AM fungi was detected as the soil dried from early to late summer, suggesting that the adaptation of AM fungi to soil moisture availability varies with species. *Glomus viscosum*, *Funneliformis mosseae*, and *Glomus hoi* were dominant in early summer, under conditions of moisture sufficiency, whereas the dominant AM fungal OTUs in late season conditions (i.e., dry soil) belonged to *Glomus iranicum* and *Glomus macrocarpum*. This concurs with the previous observation of differences in the seasonal pattern of sporulation of different AM fungal species (Dhillion and Anderson 1993). Seasonal variation of AM fungi in the North American Great Plains was also described as the replacement of the fungi of the order Helotiales by AM fungi as the season unfolds in the North American Great Plains (Jumpponen 2011).

The Chernozem great groups are distributed along a gradient of precipitation radiating outward from the US border in eastern Alberta, i.e., from the Brown soil zone through Dark Brown and Black soils up to the Gray soil zone at the fringe of the boreal forest. The lowest abundance, richness, and diversity of AM fungi were observed in the driest soil zone (Brown Chernozem), which supported a negative impact of moisture deficit on these fungi.

Soil moisture appears to be just one of several factors that influence the composition of AM fungal communities in Chernozem soils. Despite the highest levels of precipitation in the Gray soil zone, the highly productive Black soils harbor the most abundant and diverse AM fungal communities (Dai et al. 2012). Black, Gray, Dark Brown, and Brown soils had an average of 10.2, 7.1, 7.0, and 6.2 AM fungal OTUs, respectively, and the Shannon diversity index of these soil groups follows a similar trend. AM fungal communities in Brown soils are characterized by a reduced relative abundance of Claroideoglomeraceae compared to Black and Dark Brown soils. Other important factors that influenced the abundance of AM fungal OTUs were A horizon thickness and physicochemical properties of the soils, such as bulk density, Zn level, pH, electrical conductivity, and sulfur level.

Soils are classified based on their physical and chemical properties. A soil type represents a living environment inhabited by different AM fungal communities. American Mollisols and Alfisols contain distinct AM fungal spore assemblages (Ji et al. 2012). Similarly, Canadian Chernozems and Podzols and even different great groups of Chernozems contained distinct assemblages of AM fungal rRNA gene sequences (Dai et al. 2013).

Land use modifies the conditions of the soil environment and the impact of land use on the structure of AM fungal communities exceeds that of soil type. In the Canadian Prairie, roadsides host a higher level of AM fungal diversity than cropland or natural areas (Dai et al. 2013). Roadsides have higher soil moisture levels than cropland and most natural areas, further indicating that water availability is an important determinant of the abundance and structure of AM fungal communities. Seven percent of the AM fungal OTUs found across the prairie soil zones are unique to croplands, whereas 14 % of the AM fungal OTUs are specific to roadsides. Roadsides and natural areas are dominated by an OTU closely related to *Claroideoglomus lamellosum, C. etunicatum, and C. claroideum*, which account for 14 % and 19 % of all AM fungal reads. In cropland, an OTU closely related to *Funneliformis mosseae* accounted for as much as 17 % of all AM fungal reads. The dominance of *F. mosseae* in croplands of the Canadian prairie is supported by studies based on metagenomic methods (Ma et al. 2005; Sheng et al. 2012; Dai et al. 2012, 2013) and on spore counts (Talukdar and Germida 1993).

Crop management systems also have a strong influence on the composition of AM fungal communities in Chernozem soils. Organic systems have been shown to support more abundant and diverse AM fungal communities compared to conventional systems (Dai et al. 2014). Organic systems also promote greater proliferation of *Claroideoglomus* and of *incertae sedis* taxa of the Glomeraceae, currently referred to as *Glomus iranicum* and *Glomus indicum*. However, these Glomeraceae *incertae sedis* are seemingly parasitic as they were associated with reduced crop growth and N and P uptake efficiency.

Summary

Metagenomic studies on the distribution of AM fungi in Chernozems are extremely useful to understand how the living soil provides ecological services and supports the production of food and bioproducts. Brown Chernozems are relatively poor in symbiotic AM fungi and are less hospitable to the *Claroideoglomus* than other Chernozems, whereas Black Chernozems are rich in AM fungal resources. The influence of soil type on the composition of AM fungal communities is relatively small compared to that of land use type. *Funneliformis* have a competitive edge and proliferate in conventional crop production systems, whereas *Claroideoglomus* and Glomeraceae *incertae sedis* are favored in organic production systems. These Glomeraceae *incertae sedis*, currently known as the *G. iranicum*/*G. indicum* group, are associated with reduced crop productivity and nutrient uptake.

References

Bentivenga SP, Hetrick BAD. The effect of prairie management practices on mycorrhizal symbiosis. Mycologia. 1992;84:522–7.

Dai M, Bainard LD, Hamel C, Gan Y, Lynch D. Impact of land use on arbuscular mycorrhizal fungal communities in rural Canada. Appl Environ Microbiol. 2013;79:6719–29. doi:10.1128/aem.01333-13.

Dai M, Hamel C, Bainard LD, St. Arnaud M, Grant CA, Lupwayi NZ, Malhi SS, Lemke R. Negative and positive contributions of arbuscular mycorrhizal fungal taxa to wheat production and nutrient uptake efficiency inorganic and conventional system in the canadian prairie. Soil Biol Biochem. 2014;74:156–166.

Dai M, Hamel C, St. Arnaud M, He Y, Grant C, Lupwayi N, Janzen H, Malhi SS, Yang X, Zhou Z. Arbuscular mycorrhizal fungi assemblages in chernozem great groups revealed by massively parallel pyrosequencing. Can J Microbiol. 2012;58:81–92.

Dhillion SS, Anderson RC. Seasonal dynamics of dominant species of arbuscular mycorrhizae in burned and unburned sand prairies. Can J Bot. 1993;71:1625–30.

Durán A, Morrás H, Studdert G, Xiaobing L. Distribution, properties, land use and management of Mollisols in South America. Chin Geogr Sci. 2011;21:511–30.

Eom AH, Wilson GWT, Hartnett DC. Effects of ungulate grazers on arbuscular mycorrhizal symbiosis and fungal community structure in tallgrass prairie. Mycologia. 2001;93:233–42.

Garg N, Chandel S. Arbuscular mycorrhizal networks: process and functions. A review. Agron Sustain Dev. 2010;30:581–99.

Hart MM, Klironomos JN. Diversity of arbuscular mycorrhizal fungi and ecosystem functioning. In: van der Heijden MGA, editor. Mycorrhizal ecology, Ecological studies, vol. 157. Berlin: Springer; 2003. p. 225–42.

Ji B, Bentivenga SP, Casper BB. Comparisons of AM fungal spore communities with the same hosts but different soil chemistries over local and geographic scales. Oecologia. 2012;168:187–97.
Jumpponen A. Analysis of ribosomal RNA indicates seasonal fungal community dynamics in *Andropogon gerardii* roots. Mycorrhiza. 2011;21:453–64.
Liu X, Lee Burras C, Kravchenko YS, Duran A, Huffman T, Morras H, Studdert G, Zhang X, Cruse RM, Yuan X. Overview of Mollisols in the world: distribution, land use and management. Can J Soil Sci. 2012;92:383–402.
Ma WK, Siciliano SD, Germida JJ. A PCR-DGGE method for detecting arbuscular mycorrhizal fungi in cultivated soils. Soil Biol Biochem. 2005;37:1589–97.
Porras-Alfaro A, Herrera J, Natvig DO, Lipinski K, Sinsabaugh RL. Diversity and distribution of soil fungal communities in a semiarid grassland. Mycologia. 2011;103:10–21.
Schuessler A. Glomeromycota. Taxonomy. 2013. Accessed 6 Nov 2013. http://schussler.userweb.mwn.de/amphylo/
Sheng M, Hamel C, Fernandez MR. Cropping practices modulate the impact of glyphosate on arbuscular mycorrhizal fungi and rhizosphere bacteria in agroecosystems of the semiarid prairie. Can J Microbiol. 2012;58:990–1001.
Talukdar NC, Germida JJ. Occurrence and isolation of vesicular-arbuscular mycorrhizae in cropped field soils of Saskatchewan, Canada. Can J Microbiol. 1993;39:567–75.
Tian H, Gai JP, Zhang JL, Christie P, Li L. Arbuscular mycorrhizal fungi in degraded typical steppe of Inner Mongolia. Land degrad dev. 2009;20:41–54.
Yang C, Hamel C, Schellenberg MP, Perez JC, Berbara RL. Diversity and functionality of arbuscular mycorrhizal fungi in three plant communities in semiarid Grasslands National Park. Can Microb Ecol. 2010;59:724–33.
Young IM, Crawford JW. Interactions and self-organization in the soil-microbe complex. Science. 2004;304:1634–7.