Bioinformation by Biomedical Informatics Publishing Group open access
www.bioinformation.net Views & Challenges

The next meta-challenge for Bioinformatics

Willy Valdivia-Granda[1, *]

[1]Orion Integrated Biosciences Inc., New York, United States of America;
Willy Valdivia-Granda* - E-mail: Willy.Valdivia@orionbiosciences.com; * Corresponding author

received May 15, 2008; revised May 22, 2008; accepted May 26, 2008; published May 29, 2008

Abstract:
The direct sequencing of uncultivable organisms present in complex biological and environmental samples has opportunities to discover new life forms and metabolic processes. This transformational field, known as metagenomics, is generating massive amounts of molecular information that can overwhelm the performance of conventional analysis and visualization algorithms. Here, I briefly highlight some of the emerging challenges this new discipline presents to the computational biology community and point some of the opportunities to develop applications that can translate metagenomic information into biomedical, agricultural, environmental, and industrial applications.

Keywords: metagenomics; high performance computing; enterprise database; genomic barcoding

Background:

There are a remarkably vast number of microorganisms on the planet. However, on both local and global scales, the wealth of this diversity is poorly appreciated. Not only is the number of described species a very small proportion of those existing in nature, but of this 0.1%, less than 1% have been cultivated and taxonomically classified using established morphological species concepts (MSC). According to the MSC, species are the smallest groups that are consistently and persistently distinct, as well as distinguishable by ordinary means. However, the plasticity of microbial genomes, and the rigidity of the parameters employed by the MSC has led to discrepancies in the taxonomical classification of different species when using molecular fingerprinting approaches **[1, 2]**.

In the last five years several groups have achieved the direct genome sequencing of organisms present in complex samples **[3-9]**. This process is completed mainly by three steps: 1) DNA is extracted directly from a *microbiome* consisting of mainly uncultivable organisms. 2) The DNA sample is sequenced directly or cloned in bacterial or viral plasmids and sequenced. 3) The direct sample or library DNA sequence file is assembled into putative genes and genomes and analyzed using different comparative genomics and phylogenomic tools. At difference of conventional sequencing projects, metagenomics generates massive amounts of genome information. This scale of data can easily overwhelm the capacity of existing software developed for data exchange, analysis, and visualization. For example, the DNA from a single organism's genome requires $1x10^7$ bytes of storage, increasing to $1x10^{10}$ bytes when the genome is annotated. By contrast, $1x10^7$ bytes of metagenome from a single sample and its associated metadata might require $1x10^{12}$ bytes or nine terabytes of disk space. In order to simultaneously compare metagenomic samples at different levels of resolution (e.g. sequence or systems biology levels) and discover new life forms as well as unravel their diversity, new computational biology tools are required. In consideration of these issues, here some computational biology and cyber-infrastructure challenges that need to be addressed in order to translate me tagenomic information into biomedical, agricultural, environmental, and industrial applications will be highlighted. In addition, issues of metagenomic raw- and metadata exchange formats, algorithms for sequence assembly, comparative analysis, pathway reconstruction, phylogenomics, molecular bar-coding, and diversity estimation will be presented. Finally, some emerging technologies that bioinformaticians and metagenomic researchers should consider during the implementation of new research programs will be summarized.

Challenges and opportunities:

Collecting, organizing, and integrating metagenomic data from multiple sources and making them easily accessible to different research communities is one of the first and most important tasks. However, linking metagenomic information with molecular adaptive patterns is not trivial and requires the simultaneous implementation of data exchange systems and metadata standards. This information should not only describe the sample (e.g. environmental, diagnostic, or ecological) but track its origins, as well DNA extraction protocols, computational methods used for sequence assembly, gene finding, and biological function annotation. Since most conditions from which metagenomic samples are extracted vary over time (e.g. temperature or host health status), it is necessary to record this information with much detail as possible using rich but standardized annotation pipelines that can be useful to different user needs. These workflow infrastructures should facilitate the use of machine-readable implementations including simple agents, semantic web tools, or artificial intelligence programs (e.g. natural language processing) that query, tag, evaluate, index, and integrate disparate databases. These architectures should interoperate with database schemas that define a set of metagenomics attributes and attribute semantics, the relationship between the attributes, and the syntax for attribute values.

Because metagenomic samples are collected from a variety of sources, the use of appropriate terminology describing details and specific features of the collection site (e.g. biogeographical information) or host habit condition should be included. This requires the development of semantic interoperable metagenomic ontologies which reuse existing biomedical, environmental, and geographical terminologies to construct tables for sample collection and experimental protocol description. These systems must allow both automatic and manual data error flagging of each descriptor as well as the biological function annotations of individual sequences. In parallel, it is necessary to develop algorithms based on the annotation content that can estimate the quality of metagenomic datasets. All this information should be part of a metagenomic catalog associated with a unique and persistent identifying code that integrates individual sequences from a sample and places them within the context of a specific habitat. This identifying code should be the primary key that facilitates the interoperability of different databases, analysis algorithms, and visualization tools. Considering the size and complexity of metagenomic databases, publishers must enforce the linking of standard sequence identifiers to publications using consortium initiatives such as the Genomic Standards Consortium (GSC) and following the Minimum Information about a Metagenomic Sequence (MIMS) **[10]**. This will not only improve the automated data curation of individual sequences within a metagenomic sample, but should allow data enterprises, federations, and warehouses to perform massive data exchange operations seamlessly (e.g. map a global list of genomes and metagenomes to individual taxonomies) and efficiently utilizing user specific filtering parameters (e.g. time). These practices must be adopted very early since genomic projects have demonstrated that metadata are still underutilized. Many researchers underestimate the value of this metadata or do not see any incentive in providing this level of detail. Therefore specialized metagenomic databases, publishers, and funding agencies must work with the community to promote the use of highly annotated standard sequence identifiers and detailed annotation practices that go beyond the GeneBank format.

Current sequence assembly and gene finder programs require training data from closely related species. These algorithms are designed to connect fragments and analyze single eukaryotic and prokaryotic genomes. Since viruses and phages represent a significant proportion of existing metagenomic samples **[11]**, assembly and gene finding algorithms need to address the issues associated with the assembling reads (e.g. 50 nucleotides) generated by the use of ultrahigh throughput sequencing technologies in small genomes. For example, analysis of bacterial genomes has pointed out that 13% of the genome of *Streptococcus pyogenes* is composed of prophages **[4]**. Therefore, metagenomic bioinformatic applications must be sensitive enough to detect these molecular events and chimerism due to erroneous assembly processes. At the same time, new applications must overcome the rate (10- 20%) of inaccurate gene predictions in metagenomic datasets **[12]**.

The development of new bioinformatic applications capturing polymorphisms of individual sequences across different metagenomic datasets represents another considerable challenge. Sparse sampling of DNA from many species in a more complex environmental sample coupled with high rates of polymorphism within specific species presents a significant obstacle to determine the true diversity of a sample. Comparative analysis of metagenomic phages shows that approximately 65% of the sequences have no homologues in the non-redundant databases. Furthermore, only less than 2% of the Sargasso Sea database can be overlapped at 90% identity with sequences from existing biological databases **[13, 14]**. On the one hand, this limited similarity might be due to the unique diversity of metagenomic datasets, on the other, it is possible that these results are due to limitations of existing implementations. Consequently, metagenomic analysis tools must overcome the sensitivity of traditional comparative analysis tools such as BLAST and its computational demand. While parallel and distributed implementations such as mpi-BLAST **[15]** and scala-BLAST have emerged to address this problem, these tools run a limited number of the BLAST options and remain a heuristic solution to sequence comparison.

Thanks to metagenomics, new species have been documented for the first time. However, discovering a particular species cannot possibly represent all the members of a phylotype. For example, hundreds of functionally important genes are not seen in the other strain of *Escherichia coli* K12 or *E. coli* O157:H7 **[16]**. While these comparisons are necessary at the metagenomic level, it is also important to discriminate single nucleotide polymorphisms (SNP) that define subsets of metagenomic data. This is particularly relevant for viruses. However, detection of SNP and their direct impact in the phenotypical makeup of a microbiome remains difficult because available gene assembly programs consider single-base mismatches as errors in the sequencing process. Nonetheless, in metagenomic datasets, these variations might represent sequences from different genomes. This situation is complicated by the fact that current comparative genomic tools use co-linear sequence alignment in order to infer global and local sequence similarities. However, these methods are impaired for performing inter- and intra-microbiomes comparisons because of the high proportion of fragmented sequences typical of metagenomic datasets. This data conformation by result in multiple sequence alignment reconstruction providing inaccurate inferences to an extent that species characteristics within a family are lost **[13]**. Pair-wise sequence comparisons requires compute the square of the number of sequences in the input file, under these circumstances, metagenomics researchers must access software optimized for highly parallel systems that leverage thousands of processors available either on a grid or on a massive parallel computer cluster. Therefore, the development of comparative metagenomic analysis tools requires the rethinking of algorithms that simultaneously make useful inferences about particular lineages considering both local and global motif shuffling in genes, genomes and microbiomes. This improvement requires additional comparative parameters such as oligonucleotide composition, codon usage preference, and motif distribution.

Phylogenetic classifications of many organisms can be achieved with the use of ribosomal RNA (rRNA), however, this technique is not applicable to viruses and highly fragmented metagenomic databases. To overcome this limitation, genomic bar-coding has been proposed as an emerging approach that utilizes a standardized genome segment as a discriminatory parameter for identifying species **[17, 18]**. The use of this technique to uniquely classify members of a given taxonomical group has attracted considerable attention as well as the cause of some controversy **[17, 18, 19]**. While it is recognized that each species has distinctive short genome segments that can be used for bar-coding, there is disagreement about the usefulness of this approach to study genomic complexity and to classify species **[20]**. Different genome regions evolve at dissimilar rates **[21]**; as a result, there is no simple technique or general rule to establish the number of genomes and the length of putative barcodes that must be analyzed to determine the molecular diversity of a species, genus, or metagenomic sample. To address this issue, our company is developing new high performance barcoding algorithms that focus on individual motifs rather than genes (unpublished results). Our research suggests that for many microbial pathogens, short segments less than 50 nucleotides contain sufficient information content to discriminate species within metagenomic sets. Placing individual genomic bar-codes in the context of individual genomes have yielded insightful results about evolution and adaptation. This genomic bar-coding approach not only overcomes the lack of sensitivity and specificity of conventional analysis tools, but allows taxonomical inferences of unknown sequences. As a result, we can quickly sort unknown specimens into genetically different categories. Similar implementations can yield genomic bar-codes that can be considered as “molecular- operational taxonomic units” (M-OTU). Each M-OTU encompasses sufficient variability to allow both inter- and intra-species discrimination. We believe that genomic bar-coding can be used to generate metagenomic-specific motif fingerprints that allow comparative analysis among databases and at the same time infer what proportion of the DNA sequence of these ancestral species are present in terrestrial or less extreme environments. Similar approaches can be utilized with seasonal changes of habitat condition, geographic distribution, and molecular dynamics of different microbiomes.

There is a remarkably dense and diverse microbial ecosystem where many species have yet to be discovered, or are known but have not been formally described. While it is estimated that the total number of species on Earth could range from about 3.6 million up to 117.7 million, 13 to 20 million is frequently used. As microorganisms become abundant due to changing environmental conditions or removal of competitors, it is of great importance to monitor the temporal distribution of different populations in a metagenomic location. Since metagenomics is disproving that the diversity of microbes is globally dispersed, it is necessary to determine if seasonal changes in different locations are intractably connected. These types of studies will provide the necessary information and parameters to develop complex mathematical models where software agents resembling specific diversity ranges can be subjected to alternative *in silico* conditions that might serve as predictors of inter- and intra-species prevalence and diversity. Modeling such as this can maximize our understanding of diversity composition within microbial communities that can be sampled but not completely characterized, the impact change effects in these communities, and the impact microbial communities have on macroorganisms. Consequently, the bioinformatics community must improve population estimations methods such as Bayesian, Yule's ‘characteristic, Horwtiz-Thompson and nonparametric estimation of Shannon’s index of diversity **[6, 22]**. However, this type of modeling should consider that while a microbiome might contain a larger number of species with relatively low abundances, random sampling or the physical sample processing techniques can considerably affect these estimations. It is plausible that during the sample manipulation process some rare species may be lost and not discovered because of preferential DNA amplification **[23]**.

The development of system biology techniques has demonstrated its potential to unravel genes with important biological functions within complex molecular interactions. Applied to metagenomics, systems biology techniques can be useful to reconstruction pathways, molecular interactions and gene circuits of microbial communities at the specie and ecosystem level. This can lead to the discovery of novel biocatalysts processes that fulfill energy production, environmental stewardship, and medicine. Synthetic biology techniques can be used to insert genes and enzymes needed for implementation in production processes to further prove the value of metagenome-derived sequences. The bioprospecting of pathways can be initially achieved by linking metagenomic sequences with the wealth of information from well characterized protein subfamilies and families interactions participating in specific molecular networks. However, this homology based pathway reconstruction and the identification of regulatory elements requires automated methods to propagate information across metagenomic datasets. At the same time, it is necessary to develop ultra-high modeling technologies to compare and construct 3D models of the metagenomic proteins and to build *in-silico* protein-protein and protein-DNA interactions. In hand with new fields such as proteogenomics - the combination of community genomics and proteomics - community functions to specific member microorganisms will be possible **[24-27]**. This entails overcome the limitations of existing genomic visualization tools of organism protein interaction representation and developing paradigm shift summarization technologies representing protein interactions at the microbiome levels.

In parallel to software development, metagenomics requires data compression and exchange technologies that support transfers of more than 5000 megabytes per second. Therefore, funding agencies must support institutional changes not only for the algorithms, but in the implementation of parallel fiber optic and data transfer technologies. For example, instead of using flat files and data-server approaches, metagenomics requires distributed virtual processing enterprise environments where the computational power for analysis and visualization are shared efficiently. This scale represents new challenges for analysis and visualization algorithms, since bacterial and viral communities might require different data representation standards. More specific challenges include finding improved techniques for the major hard optimization problems (maximum parsimony and maximum likelihood) in conventional phylogenetic inference, as well as dealing with higher level analysis of whole genome evolution, with insertions, deletions, duplications, rearrangements, and horizontal transfer of DNA segments.

While most attention is focused on the biology and ecology of large species, the greater part of Earth's species diversity is found in microbes. These entities are estimated to make up more than one-third of Earth’s biomass. Thanks to metagenomics our appreciation of heir diversity offers the opportunity of discovering new species in biofilms, soil, and marine environments. More importantly, metagenomics has far reaching implications. For example, metagenomics can be used to sample the blood of hunters potentially exposed to new exotic diseases that might cause a serious impact for global health. A metagenomic approach for sampling in microbial diversity in hospitals might avert emerging multidrug resistant pathogens, and enable the management of infections in patients and the development of strategies to decrease the movement of these organisms. This information can also be used to develop of new drugs using *genomic-barcode paradigm* where pathogen-specific genome regions can be targeted with a new generation of broad antimicrobials. Metagenomic profiles can uncover new virulence factors in human and animal diseases and the genomic shifting after vaccination or drug usage.

Metagenomics and genomic bar-coding can yield a universal catalogue of species and life. This will profoundly alter our understanding of the biosphere and is likely to lead to revision of concepts such as species, organism, adaptation, and evolution. This information and the use of synthetic biology can lead to the development a unique generation of antimicrobials, therapeutic compounds and enzymes with industrial applications. The mining of metagenomic datasets will allow us to answer fundamental questions of biology. What proportion of the metagenome is specific and unique for a particular species? What does the species abundance distribution of a clone library reveal about microbiome species abundance? How do the forces driving microbiome diversity vary within a temporal and spatial scale? Do microbial metagenomic sets correlate with specific genomic barcodes? Can the genomic barcodes patterns be associated with pandemic potential of unknown microbial species? Can we identify specific molecular signatures associated with microbiome geographical distribution? Is there a scaling relationship between metagenomic samples and macroorganisms? Answering these questions will initially prioritize the ecological justifications for selecting study sites, but at the end will result in products. This is an extraordinary time for biology. Used alongside new genome sequencing technologies, the potential rewards of sophisticated bioinformatic applications for analysis and visualization of metagenomic datasets are tantalizing. Nonetheless, it is now clear that metagenomics will be the next meta-challenge for bioinformatics.

Acknowledgement:
The author would like to thank Dr. Mya Breitbart and Francis Larson for the insightful comments and suggestions. The genomic barcoding technology is partially supported by the Defense Threat Reduction Agency.

References:

[01] P. Z. Goldstein, *et al. Exs.,* 92: 147 (2002) [PMID: 11924493]
[02] R. A. Samson, *et al. Med Mycol.,* 44: 133 (2006) [PMID: 17050433]
[03] G. A. Kowalchuk, *et al. Microb Ecol.,* 53: 475 (2007) [PMID: 17345132]
[04] R. A. Edwards, *et al. Nat Rev Microbiol.,* 3: 504 (2005) [PMID: 15886693]
[05] S. G. Tringe, *et al. Science,* 308: 554 (2005) [PMID: 15845853]
[06] M. Breitbart, *et al. Proc Biol Sci.,* 271: 565 (2004) [PMID: 15156913]
[07] C. von Mering, *et al. Science,* 315: 1126 (2007) [PMID: 17272687]
[08] S. Yooseph, *et al. PLoS Biol.,* 5: e16 (2007) [PMID: 17355171]
[09] M. Breitbart, *et al. J Bacteriol.,* 185: 6220 (2003) [PMID: 14526037]
[10] D. Field, *et al. Omics,* 10: 100 (2006) [PMID: 16901213]
[11] F. E. Angly, *et al. PLoS Biol.,* 4: e368 (2006) [PMID: 17090214]
[12] K. Mavromatis, *et al. Nat Methods,* 4: 495 (2007) [PMID: 17468765]
[13] M. L. Tress, *et al. BMC Bioinformatics,* 7: 213 (2006) [PMID: 16623953]
[14] K. U. Foerstner, *et al. Philos Trans R Soc Lond B Biol Sci.,* 361: 519 (2006) [PMID: 16524840]
[15] G. Aparicio, *et al. Stud Health Technol Inform.,* 120: 194 (2006) [PMID: 16823138]
[16] N. T. Perna, *et al. Nature,* 409: 529 (2001) [PMID: 11206551]
[17] P. D. Hebert, *et al. Proc R Soc Lond B Biol Sci.,* 270: 313 (2003) [PMID: 12614582]

[18] S. Beardsley, *Sci Am.*, 292: 26 (2005) [PMID: 15882014]
[19] E. Marshall, *Science*, 307: 1037 (2005) [PMID: 15718446]
[20] C. P. Meyer, *et al. PLoS Biol.*, 3: e422 (2005) [PMID: 16336051]
[21] J. Mrazek, *et al. Proc Natl Acad Sci U S A*, 104: 5127 (2007) [PMID: 17360339]
[22] A. Cann, *et al. Virus Genes*, 30: 151 (2005) [PMID: 15744573]
[23] M. Podar, *et al. Appl Environ Microbiol.*, 73: 3205 (2007) [PMID: 17369337]
[24] Lo, *et al. Nature*, 446: 537 (2007) [PMID: 17344860]
[25] A. Norbeck, *et al. J Microbiol Methods*, 67: 473 (2006) [PMID: 16919344]
[26] R. J. Ram, *et al. Science*, 308: 1915 (2005) [PMID: 15879173]
[27] R. Stepanauskas, *et al. Proc Natl Acad Sci U S A*, 104: 9052 (2007) [PMID: 17502618]

Edited by P. Kangueane
Citation: Valdivia-Granda, Bioinformation 2(8): 358-362 (2008)