In Silico Hybridization System for Mapping Functional Genes of Soil Microorganism Using Next Generation Sequencing

Guillermo G. Torres-Estupiñan and Emiliano Barreto-Hernández

Bioinformatics Center, Biotechnology Institute, Universidad Nacional de Colombia, Bogotá D.C. – Colombia
{Ggtorrese,ebarretoh}@unal.edu.co

Abstract. Nowadays a widely production and low cost of sequencing has allowed the extension of metagenomics in order to explore the genomic information from diverse environments. This offers the opportunity to examine new approaches for sequence binning and functional assignment. Driving of metagenomic studies using high throughput sequencing, usually follows the same pipeline used to analyze single genomes: sequence assembling, gene prediction, functional annotation and phylogenetic classification of reads, contigs or scaffolds, nevertheless, the accuracy of this approach is limited by the length of the reads or resulting contigs.

In Silico Hybridization System is an approach of functional and taxonomical assignment. It has default assignment at gender taxonomic level binning. This tool works with two general pipelines: Probes Creator (CrSo), ordered to design DNA probes (fingerprints) to gender taxonomic level and Sequential In silico Hybridator (HISS) which use the probes to make the hybridization with the community reads.

This bioinformatics tool allows characterization of the microbial metabolism in charge of biogeochemical cycles, tracking their key stages using debugged reference information. This strategy resulted in an increasing of binning accuracy. The simulated and real scenarios were better described using one probe and selection threshold fitted to a logarithmic distribution, with mean sensitivity of 85% and mean specificity of 83%.

Keywords: Metagenomics, binning, metatranscriptomics, soil, nitrogen, phosphorus, biogeochemical cycles.

1 Introduction

High-throughput sequencing (HTS) technologies currently have offered an unprecedented opportunity to examine the microbial ecology on a wide scale by three general approaches: 1) 18S/16S ADNr, fast and effective way to characterize communities structure [10, 23]; 2) Whole genomic sequencing (metagenomics), this approach is able to deal with the structure of communities along with the functional potential of them [12, 26] and 3) Whole transcriptome sequencing (metatranscriptomics), in order

L.F. Castillo et al. (eds.), *Advances in Computational Biology*,
Advances in Intelligent Systems and Computing 232,
DOI: 10.1007/978-3-319-01568-2_48,

to examine the gene potential and metabolic capabilities of the microbial communities under certain conditions [17, 27]. Metagenomic analysis follows a relative similar pipeline as working with a single organism: sequence assembling, gene prediction, functional annotation and phylogenetic classification (binning) of reads, contigs or scaffolds [15]. Nevertheless, the metagenomic assembling approach has had not so good result so far [16]. In the binning process we attempted phylotyping the sequences through composition-based or similarity-based strategies.

The composition-based strategy extracts information related with GC content [7], codon usage [19] or k-mer frequency [22] from the metagenome sequences and compares them with features calculated of reference sequences with known taxonomic origin. This strategy allowing achieves optimal assignments to sequence length larger than 800bp.

On the other hand, the similarity-based strategy relies on homology information obtained by database searches. The databases could contain nucleotide sequences (i.e. complete genome) or protein sequences with known taxonomic origin. However, commonly the bioinformatics tools use protein sequences as reference for metagenomic analysis, since protein sequences are more conserved than nucleotide sequences, they are better suited for detection of remote homologies in order to explore an ecosystem which mainly consists of uncultured microorganisms [2]. But, the usage of protein sequences as reference has the disadvantage that the metagenomic DNA fragments have to be translated into all six reading frames, which increases computation time of the homology search. This strategy can be sub-divided in two general methods: those to use Hidden Markov Models (HMM) [5] or BLAST-based [1] homology searches.

Despite of all the efforts, these bioinformatics tools have not achieved both the efficiency and accuracy level required by current metagenomics high-complexity data sets because of computational limitations, unable good accurate assignments for short DNA fragments (<400bp). Therefore, we present a new strategy to evaluate the taxonomic and functional features of soil microbial communities, an integrated approach that combines bioinformatic algorithms to probe (fingerprint) design (CrSo) from debugged coding gene sequences database and in silico hybridization (HISS), that allows to characterize the soil biogeochemical metabolism. Finally we discuss initial experimental results, which help evaluate our both fingerprint specificity and hybridization selection criteria.

2 Implementation

A specialized bibliographic review allowed identifying the metabolic processes, subprocesses and the reactions that compose them, in charge of Nitrogen and Phosphorus biogeochemical cycles. In order to distinguish microbial communities, the enzyme gene markers selected were not housekeeping genes, but them were differentially distributed in microbial communities. A PHP script was developed to retrieve the nucleotide, protein sequence and taxonomy identification via SOAP protocol (Simple Object Access Protocol) on EMBL -EBI database, this information was storage in local database named as SPH.

2.1 Probe Design

CrSo has been split the probe design problem in three general steps, however, these are not independent at all and indeed the input parameters in the first step have logical relationship with second and third step criteria. The first step reduces the redundancy of input information by clustering in an attempt to bring together DNA sequences of the same biological sequence, then extracting the consensus sequence. The second step, a probe design phase extracts from candidate consensus sequences only those subsequences that satisfy the experimental specificity conditions. In the third step a probe denoising process is executed, creating clusters with DNA probes to select just the singletons.

Step 1: Redundancy reduction. We start with large enzyme CDSs information allocated in SPH. However, this database presented two features: 1) several sequences covering the same biological sequence, and 2) sequence fragments of one biological sequence are globally alignable, so, it will be impractical testing each sequence from database for probe design because it will result in repeated probes. For this, we exploit the sequence similarities, clustering them, in order to estimate a biological sequence from consensus sequence derived from a multiple alignment using UCLUST [6]. But, clustering configuration depends on homology features of the target genes, so that we determine the level of taxonomic resolution of these functional genes. The gene sequences were grouped according to their taxonomic classification and aligned, later the sequence similarities were calculated with p-distance model using MEGA 5 [24].

Step 2: Probe design. At this stage we established a set of constraints to extract probes from candidate sequences. We have selected a probe design tool, OligoWiz 2.0 [28], that implements a complex score model to select the best probe. This model allows specificity constraints like: minimum homology, minimum length of homology stretch, maximum similarity, probes length and database. Probes satisfying these constraints are extracted and passed to the next step.

Step 3: Probe denoising. In this step, all probe set candidate are clustered [6] as a filter process to discard duplicates or close related ones, selecting just the singleton clusters for the hybridization process.

2.2 *In silico* Hibridization

The hybridization process follows a general rule: each read aligned significantly with a probe, should be considered originating or homology of the enzyme gene that probe represents, therefore HISS performs a nucleotide BLAST search for each probe against HTS reads database, and reads with non-significant alignments with target probes were not taken into account.

Due to the limitations of in silico hybridization models that determine which DNA alignments are significant, HISS examines alignment features to determine homology relationship between probe and reads. As the probes are designed to identify multiple target genes we have to consider multiple criteria to determine a homology selection

threshold, such as overall sequence identity, contiguous matches, mismatches, gaps and alignment length [13]. Analytically HISS incorporate a threshold selection function generated from series of continuous matches thresholds Xi with i mismatches. From basis on empirical threshold [14], we define 21 to X0, the longest stretch of contiguous matches between a probe and metagenomic sequence, the following thresholds were calculated such as Xi = X0 + W(i). Where W is BLAST input parameter, termed word size, involved in BLAST heuristic approach. In blastn program used for BLAST searches the default value of W is 11 and the smallest value is 7. Because every BLAST result has to include an exact match of length W, it becomes a bound on values of specificity thresholds, therefore, in order to increase the sensitivity of HISS, it implements a default W value of 7 [29].

3 Results

The entire pipeline was implemented on a HPC environment at Bioinformatics Center server of Biotechnology Institute of Universidad Nacional de Colombia, the cluster consists of 10 X 2.8 GHz Quad Core processors, running with SUSE10 with 32 GB of shared memory. The computational time of the algorithm depends on the number of probes and metagenomic sequences for the hybridization, hence the computational time of processing is directly dependent on the speedup achieved by BLAST, so the execution of HISS could be improved by using parallel versions of BLAST like mpiBLAST [4] or pioBLAST [9].

The SPH database, lodge 33536 sequences for Nitrogen cycle, associated with 39 reactions of nitrogen fixation, nitrification, mineralization, assimilation and denitrification process, and 13883 sequences for Phosphorus cycle, associated with 34 reactions of mineralization process of phosphomonoesters, phosphodiesteres and inositol phosphates.

The taxonomic resolution suggests a high variability, and confirms that amoA gene is a species specific marker but not higher taxonomical levels, furthermore, we found that species of the same gender clustering at an identity mean of 82% and 73% for genders of the same family (Table 1). Therefore, for the first stage of CrSo, the redundancy parameter was settled to 82% of identity to clustering, in order to make gender lever bins. The probes were designed with 75% of identity for minimum homology, 30 of minimum length of holomoly stretch [14], 83% of identity for maximum homology, probes length of 100bp and local database named as ProEnvFun compounded by EMBL sequences from Prokaryotes, Environmentals and Fungy groups to calculate cross hybridization (deeper information in [28]). The denoising was configured with 82 % of identity clustering.

CrSo triggers five sets of three best probes according to their length (25bp, 40bp, 60bp, 80bp and 100bp) that were evaluated with simulated metagenomes made with Grinder [3]. Ten Illumina sequencing metagenomes were simulated mimic to a low complexity community [25], 20 soil representatives genomes were selected from NCBI RefSeq. With the best probes we fitted the selection HISS parameters, that were started as linear function traced using a series of thresholds X_0, X_1, X_2, X_3, X_4

and X_5, analytically calculated, such as: 21, 28, 35, 42, 49 and 56 respectively. However we acquire better results with a logarithmic function (y = 9.9581ln(x) – 32.067), achieving a mean sensitivity of 85% and mean specificity of 83% for all five length sets of probes (The sensitivity was evaluated as the ratio of true positives and all Blast hits. The specificity was calculated as the ratio of true positives and the value of true positives plus false positives).

Table 1. Taxonomic resolution of functional genes. Data show the mean (±sd) of similarity values of the gene groups at different phylogenetic levels. N.A. not applicable

			Phylogenetic Hierarchies (% sequence similarity)		
Cycle	EC number	Gene	Strain	Specie	Gender
Nitrogen	1.14.99.39	*amo*A	0.98 ± 0.01	0.82 ± 0.01	0.78 ± 0.03
	1.18.6.1	*nif*D	0.95 ± 0.01	0.87 ± 0.01	0.82 ± 0.01
		*nif*K	0.91 ± 0.01	0.76 ± 0.01	0.73 ± 0.01
		*nif*H	0.92 ± 0.02	0.86 ± 0.01	0.83 ± 0.01
	1.7.2.1	*nir*K	0.91 ± 0.01	0.82 ± 0.01	0.72 ± 0.01
		*nir*S	0.93 ± 0.03	0.75 ± 0.02	0.66 ± 0.01
	1.7.99.4	*nap*A	0.96 ± 0.01	0.87 ± 0.01	0.76 ± 0.01
		*nar*G	0.93 ± 0.01	0.83 ± 0.02	0.79 ± 0.03
		*nar*H	0.90 ± 0.02	0.80 ± 0.01	0.71 ± 0.02
		*nar*I	0.92 ± 0.03	0.73 ± 0.03	0.66 ± 0.05
		*nar*Z	0.98 ± 0.02	0.90 ± 0.01	N.A.
		*nar*Y	0.98 ± 0.01	0.81 ± 0.01	0.81 ± 0.01
	1.7.1.4	*nas*B	0.95 ± 0.01	0.83 ± 0.02	0.72 ± 0.01
	1.7.7.1	*nas*A	0.92 ± 0.01	0.84 ± 0.01	0.69 ± 0.05
	1.7.2.5	*nor*B	0.94 ± 0.01	0.82 ± 0.02	0.70 ± 0.03
		*nor*C	0.93 ± 0.01	N.A.	N.A.
	6.3.1.2	*gln*A	0.98 ± 0.01	0.84 ± 0.01	0.73 ± 0.03
Phosphorus	3.1.3.11	fbp	0.95 ± 0.01	0.81 ± 0.02	0.76 ± 0.01
	3.1.3.8	phy	0.95 ± 0.01	0.87 ± 0.01	N.A.
	3.1.3.15	hisB	0.90 ± 0.02	0.75 ± 0.02	0.70 ± 0.02
		hisJ	0.95 ± 0.01	0.79 ± 0.02	N.A.
	3.1.3.18	cbbZ	0.92 ± 0.01	0.82 ± 0.02	0.64 ± 0.02

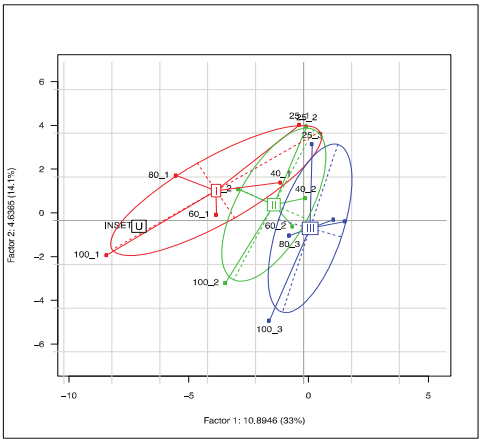

Fig. 1. Comparison of community characterization of different probe sets. We use an ACP to represent the synthetic metagenome INSET (U) as a dot and comparing with probe length and number of probes for gene descriptors.

The effect of multiple probes for gene was evaluated in order to figuring out the community structure. The ACP result suggests that the better way of characterize the community is with one probe of 100bp, 80bp or 60bp in length (Fig. 1), and there is not evidence to determine which is the best. With 100bp probes, we performed an evaluation of probes usage strategy against full-length gene to mapping the community. The full-length gene approach consider BLAST results with bitscore bigger than 55 and 100bp alignment length to assign reads [11]. The results suggest that using full-length genes overestimates the number of genes in the community (Fig. 2), similar scenery have been reported [8].

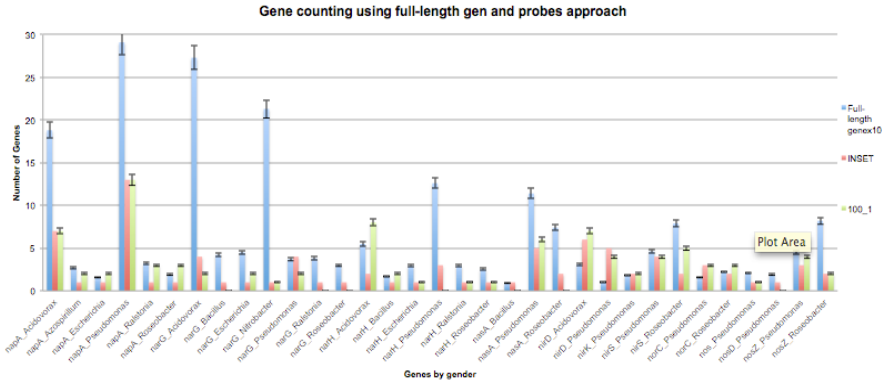

Fig. 2. Number of reads assigned with HISS one 100bp probe set and full-length gene mapping. INSET represents the expecting gen number. Full-lenth gene BLAST hits are represented 10 times less for representation simplicity. Error 5%.

There were generated two metagenomes at 1:5 information ratio (Lib1X and Lib5X) for evaluating sensitivity of HISS. We tracked denitrification process by characterizing the community gene abundance. HISS estimates the genes number close to the expected value, with a mean of 4 (stdev. 1.5 and mode of 4.3) times greater in Lib5X against Lib1X.

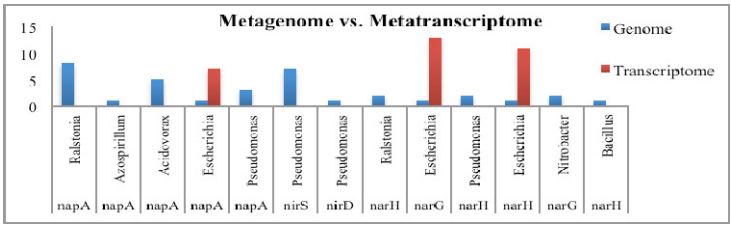

Fig. 3. Denitrification genes with different abundances detected in the metagenome and the metatranscriptome

Real Scenario. It was explored the functionality of HISS over real soil community sequences collected from potato (*Solanum phrueja*) crop, located in Cundinamarca department - Colombia. The system detects denitrification genes in metagenome and

some of these were confirmed on metatranscriptome but with different abundances (Fig. 3). HISS founds a high richness of functional genes associated to denitrification process of different genders, however just few were expressed. That was expected, since the ecosystem functional richness possesses an independent dynamics regarding to functional homogeny, nevertheless there is a close relationship with taxonomical diversity, such as taxa richness and abundance [18].

4 Discussion

CrSo probes and HISS detection system were developed for gene detection in short reads, they provide a versatile method to predict functional genes in soil metagenomes using probes. The number of probes for gene is critical to characterize them in the samples; the results exhibit a better characterization with one probe for gene. It is due to the hard task of retrieve multiple subsequences from the gene with low cross hybridization with no overlapping.

The HISS selection function is stringent enough to predict gene fragments of at least 21bp at a gender taxonomic level. In contrast to other approaches that are able to assign as short as 60bp [20]. CrSo and HISS have showed a high gene abundance sensitivity using simulated metagenomic data sets, in contrast to similarity-based binning approaches that have indicated that a significant amount of reads get unclassified or even misclassified [8, 11].

In the current study, a novel functional gene assignation with gender taxonomical level has been devised that attempts to characterize soil metagenomes according to its functional groups. The evaluation of real samples suggests that it is possible to track the behavior of soil community in order to develop comparable functional and metabolic pathway profiles of communities.

This work has calculate the similarity of Nitrogen and Phosphorus functional genes at different phylogenetic levels, it confirms amoA gene as specie-specific marker [21], however its behavior differs at higher taxonomical levels.

References

1. Altschul, S.F., et al.: Basic local alignment search tool. J. Mol. Biol. 215(3), 403–410 (1990)
2. Amann, R.I., et al.: Phylogenetic identification and in situ detection of individual microbial cells without cultivation. Microb. Rev. 59(1), 143 (1995)
3. Angly, F.E., et al.: Grinder: a versatile amplicon and shotgun sequence simulator. Nucleic Acids Res. 40(12), 94 (2012)
4. Darling, A.E., et al.: The design, implementation, and evaluation of mpiBLAST. Presented at the Conference on Linux Clusters: The HPC Revolution 2003 in ClusterWorld Conference & Expo and the 4th International (2003)
5. Eddy, S.R.: Profile hidden Markov models. Bioinfor. 14(9), 755–763 (1998)
6. Edgar, R.C.: Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26(19), 2460–2461 (2010)
7. Foerstner, K.U., et al.: Comparative Analysis of Environmental Sequences: Potential and Challenges. Philosophical Transactions: Biological Sciences 361(1467), 519–523 (2006)

8. Haque, M., et al.: SOrt-ITEMS: Sequence orthology based approach for improved taxonomic estimation of metagenomic sequences. Bioinformatics 25(14), 1722–1730 (2009)
9. Heshan Lin et al.: Efficient Data Access for Parallel BLAST. Presented at the 19th IEEE Internat. Parallel and Distributed Processing Symposium (2005)
10. Hugenholtz, P., Tyson, G.W.: Microbiology: metagenomics (2008)
11. Huson, D.H., et al.: MEGAN analysis of metagenomic data. Genome Res. 17(3), 377–386 (2007)
12. Huson, D.H., et al.: Methods for comparative metagenomics. BMC Bioinformatics 10(suppl. 1), S12 (2009)
13. Kane, M.D., et al.: Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. Nucleic Acids Res. 28(22), 4552–4557 (2000)
14. Liebich, J., et al.: Improvement of oligonucleotide probe design criteria for functional gene microarrays in environmental applications. Appl. Environ. Microbiol. 72(2), 1688–1691 (2006)
15. Mavromatis, K., et al.: Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. Nat. Methods 4(6), 495–500 (2007)
16. Mende, D.R., et al.: Assessment of metagenomic assembly using simulated next generation sequencing data. PLoS ONE 7(2), e31386 (2012)
17. Moran, M.: Metatranscriptomics: eavesdropping on complex microbial communities. Microbe (2009)
18. Mouillot, D., et al.: Functional regularity: a neglected aspect of functional diversity. Oecologia 142(3), 353–359 (2004)
19. Noguchi, H., et al.: MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. Nucleic Acids Res. (2006)
20. Rho, M., et al.: FragGeneScan: predicting genes in short and error-prone reads. Nucleic Acids Res. 38(20), e191 (2010)
21. Rotthauwe, J.H., et al.: The ammonia monooxygenase structural gene amoA as a functional marker: molecular fine-scale analysis of natural ammonia-oxidizing populations. Appl. Environ. Microbiol. 63(12), 4704–4712 (1997)
22. Sandberg, R.R., et al.: Capturing whole-genome characteristics in short sequences using a naïve Bayesian classifier. Genes Dev. 11(8), 1404 (2001)
23. Schloss, P.D., Handelsman, J.: Introducing SONS, a Tool for Operational Taxonomic Unit-Based Comparisons of Microbial Community Memberships and Structures. Appl. Environ. Microbiol. 72(10), 6773–6779 (2006)
24. Tamura, K., et al.: MEGA5: Molecular Evolutionary Genetics Analysis Using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. Mol. Biol. E 28(10), 2731–2739 (2011)
25. Tringe, S.G., et al.: Comparative metagenomics of microbial communities. Science 308(5721), 554–557 (2005)
26. Urich, T., et al.: Simultaneous assessment of soil microbial community structure and function through analysis of the meta-transcriptome. PLoS ONE 3(6), e2527 (2008)
27. Vila-Costa, M., et al.: Transcriptomic analysis of a marine bacterial community enriched with dimethylsulfoniopropionate. ISME J., 1–11 (2010)
28. Wernersson, R., Nielsen, H.B.: OligoWiz 2.0–integrating sequence feature annotation into the design of microarray probes. Nucleic Acids Res. 33(Web Server issue), W611–W615 (2005)
29. Ye, J., et al.: Primer-BLAST: A tool to design target-specific primers for polymerase chain reaction. BMC Bioinformatics 13(1), 134 (2012)