A Statistical Framework for the Functional Analysis of Metagenomes

Itai Sharon[1], Amrita Pati[2], Victor M. Markowitz[3] and Ron Y. Pinter[1]

[1] Department of Computer Science, Technion, Haifa, Israel
[2] Genome Biology Program, DOE Joint Genome Institute, 2800 Mitchell Dr, Walnut Creek, CA 94598
[3] Biological Data Management and Technology Center, Lawrence Berkeley National Laboratory, Berkeley, CA 94720

{itaish, pinter}@cs.technion.ac.il, {apati, VMMarkowitz}@lbl.gov

Abstract. Metagenomic studies consider the genetic makeup of microbial communities as a whole, rather than their individual member organisms. The functional and metabolic potential of microbial communities can be analyzed by comparing the relative abundance of gene families in their collective genomic sequences (metagenome) under different conditions. Such comparisons require accurate estimation of gene family frequencies. We present a statistical framework for assessing these frequencies based on the Lander-Waterman theory developed originally for Whole Genome Shotgun (WGS) sequencing projects. We also provide a novel method for assessing the reliability of the estimations which can be used for removing seemingly unreliable measurements. We tested our method on a wide range of datasets, including simulated genomes and real WGS data from sequencing projects of whole genomes. Results suggest that our framework corrects inherent biases in accepted methods and provides a good approximation to the true statistics of gene families in WGS projects.

Keywords: metagenomics, functional analysis, function comparison, Lander-Waterman.

1 Introduction

It has been estimated that less than 1% of the microbial species living on earth have been cultivated in the laboratory. Our inability to culture the vast majority of the rest is partially due to the complex conditions in which these organisms live, conditions that cannot be fully reconstructed in the lab. Metagenomics is a new and rapidly developing field that makes it possible to study uncultured organisms and their ecological systems. Several important discoveries have been made in recent years by extracting data directly from the environment, including the discovery of proteorhodopsin [1]. Metagenomic surveys, in which both sampling and sequencing are done randomly, provide a useful way to study microbial communities directly from the environment. In this type of projects researchers are usually interested in

determining parameters related to community structure, such as the number of species and their diversity, as well as parameters related to functional capabilities of organisms in the environment. Several metagenomic surveys were performed to-date in diverse environments such as the sea [2, 3, 6, 13, 14], acid mine drainage [4], human distal gut [5], and more.

Metagenomic studies present us with new computational challenges that are quite different from those of classical genomics. In most cases the amount and coverage of sequence data is insufficient to ensure assembly and classification of sequences into different microbial populations, thus preventing even limited population-specific genomic and metabolic reconstruction. Consequently, a prevalent method for analyzing metagenomic datasets is to compare the relative frequencies of gene families between datasets to highlight over- and under-represented functions in a given microbial community [11, 12]. Such comparisons require a measure of confidence in the observed differences in gene family frequencies between metagenomic datasets which are usually based on statistical tests. Statistical tests that have been applied to metagenome dataset comparisons are based on re-sampling [11, 12, 26]. These methods may produce reliable results but re-sampling does not scale well computationally with increased dataset sizes.

Recall that DNA sequencing – for both metagenomic and conventional genomic studies – is done using a process known as Whole Genome Shotgun (WGS) sequencing [20, 21, 22]. First, the extracted DNA is sheared randomly using physical or chemical means to fragments ranging in size between a few kilo base pairs (Kbps) to a few dozen Kbps. Next, the fragments are inserted into vectors which are used for constructing a genomic library. Last, clones from the library go through pair-end sequencing in which approximately 800bps from each side of each clone are sequenced using Sanger sequencing[1]. The outcome of this process is a set of many pair-end reads that can be used for all types of analysis including assembly, binning, and gene finding. Lander and Waterman [16] developed a statistical model for single-genome WGS sequencing projects that makes it possible to estimate parameters such as the expected number of contigs (continuous assembled reads) as a function of coverage depth, where coverage is the average number of reads covering each position in the genome.

Functional analysis of metagenomic data provides valuable insight into the kind of biological functions that are predominantly performed by organisms in a given environment E. One way of performing such an analysis is by aligning the metagenomic sequence data (also referred to as a "the metagenome") against databases such as COG [8] or Pfam [9] and identifying the most abundant protein families in it (note that we use the terms "gene family" and "protein family" interchangeably, as implied by context). Once the frequency of each protein family in the metagenome has been estimated, it is possible to identify those families most essential to the survival of certain species in E by comparing them with the frequencies observed in other metagenomes taken from similar or different environments. This process, termed *function comparison*, has become a common

[1] Alternatively it is possible to sequence DNA directly from the raw sample using pyrosequencing. While Sanger sequencing is considered in this paper, the framework described is also applicable to pyrosequencing with slight modifications.

routine in metagenomic works [6, 12, 25, 26] and has proven to be very useful. For example, DeLong *et al.* [6] collected metagenomic samples from seven depths ranging between 10 to 4000 meters from a Pacific Ocean station near Hawaii. By performing function comparison on the seven samples using the COG database, the authors were able to identify gene families that are characteristic of certain depths, such as photosynthetic-related genes that are most abundant in shallow water. Other families had significant representation in all depths, suggesting that they represent functions required by a wide range of microbes regardless of their environment. This illustrates the importance of comparing metagenomes: high abundance of a gene family does not necessarily imply relevance to specific conditions in the environment from which it was obtained. By performing function-comparison it is possible to differentiate between environment-specific and environment-independent functions.

Frequency estimation of gene families is usually done using a process we term the *read-counts approach*. The frequency of a family P is given by the number of reads carrying members of P divided by the number of reads carrying members of any other family P' (see Section 1.1 below). While being simple and straightforward, this approach "favors" families composed of longer genes and assigns them higher frequencies. In order to see why, consider two equally abundant gene families P_1 and P_2 in which members of P_1 are on average twice as long as members of P_2. Assuming that reads are sampled and sequenced uniformly across all positions in all genomes in the environment, then the above process should yield more reads containing portions of P_1 than reads containing portions of P_2 as a direct effect of the difference in lengths. In the hypothetical event of reads of length 1 (single base pair) the estimated frequency of P_1 should be twice the estimated frequency of P_2. This bias, known as the *read-counts bias*, is not negligible, as the length of genes may vary between a few dozen base pairs (*e.g.* tRNA genes) and a few thousand base pairs (*e.g.* the photosynthetic genes *psaA* and *psaB* and many other genes).

The scope of our work is function analysis based on low-level function databases, in which each family represents a single functionality such as protein (*e.g.* COG, [8]) or domain (*e.g.* Pfam [9] and TIGRfam [10]). We present a statistical framework for estimating family frequencies that is based on the abovementioned Lander-Waterman model and explain how derived frequencies may be used for functional comparison of metagenomic datasets. In addition, we provide a novel method for assessing reliability of computed frequencies which can be used for removing seemingly unreliable measurements. We have tested our method on both simulated and real WGS sequencing data from projects of whole genomes. Our tests indicate a substantial improvement of our method over existing ones.

Our contribution is 3-fold: (i) correction of the read-counts bias which is present in all methods that have been suggested to-date; (ii) for the first time a complete theoretical statistical framework with reasonable assumptions is being suggested, and (iii) the most extensive testing performed to-date, both on synthetic as well as (again, for the first time) on real data.

1.1 Previous Work

Functional characterization of metagenomic data involves (i) identifying protein CoDing Sequences (CDSs) in unassembled or partially assembled metagenomic sequences using an ab initio or evidence-based gene finder, then (ii) associating these CDSs with protein families, such as COGs, Pfams, and TIGRfams, and subsequently (iii) comparing the relative abundance of protein families. Protein coding sequences are associated with protein families using BLAST against sequence databases such as COG, reverse position-specific BLAST (RPS-BLAST) against position specific scoring matrices (PSSMs, e.g. the CDD database [27]), or using Hidden Markov Models (HMMs, e.g. the Pfam and TIGRfam databases).

Several methods for function comparison have been proposed to date [6, 11, 17, 26]. They differ in the way they perform function comparison, but compute the frequency for a protein family P from the proportion of reads in a metagenomic sample M that are associated with P when M is compared to a function database using BLAST [24]. Observe that in such an approach, each read may be associated with multiple protein families, and hence, counted several times. We refer to this approach as the read-counts frequency computation approach; as mentioned above, this method tends to overestimate the frequencies of longer genes. The method described in [6] and [11] begins with the assessment of frequencies for each family as described above, and then computes the distance between the frequencies of each protein family in the two metagenomes using simulations. The computation of the p-value for this distance is also done using simulations. The method is supposed to be assumption-free, but the simulation process described is – in fact – a binomial one. In addition, the generation of the random distribution is somewhat ad-hoc. Finally, the simulations may make the process computationally intensive, possibly unnecessarily since the distributions may be computed directly considering the fact that the process is binomial.

The method employed in the IMG/M system [11] is also based on the computation of family-frequencies using the read-counts approach, but differs in its function comparison procedure. Given a protein family P, and two metagenomes M_1 and M_2 for which the respective frequencies f_1 and f_2 of association of P with reads has been computed, the method first computes the distance between f_1 and f_2 and an associated p-value for the distance. The p-value computation is based on the null hypothesis that the raw counts of occurrence of P among reads in both M_1 and M_2 can be approximated by a binomial distribution whose Bernoulli probability is computed as a pooled probability from the counts of occurrence of P in both metagenomes. Such a computation is biased towards the frequency of P in the larger metagenome, which is undesirable. Also the use of a common source distribution is likely to reduce the significance of the difference between the two metagenomes. The work described in [17] is a set of new statistical methods for comparing communities. Rather than using databases such as COG or Pfam, the authors employ clustering algorithms for putting together related proteins. This approach may extend functional analysis beyond known genes at the expense of the reliability of the results.

2 Methods

In this section, we propose a statistical model for computing the frequency of a protein family among reads for a given genome. This will serve as a more accurate replacement of the binomial proportion that has been widely used in the scientific literature to model protein family frequencies. The proposed statistical model derives its roots from the Lander-Waterman theory [16] on the statistics of WGS sequencing projects. The model is described with respect to a single genome; minor adjustments which are required in order to make the theory suitable for metagenomes are described later.

2.1 Estimating the Frequency of a Protein Family

Consider a WGS sequencing project in which N independent clones ($2N$ pair-end reads) of average read length L are sequenced randomly from a genome of length Γ. Assume that each clone may start at any position in the genome with equal probability $\alpha = \frac{N}{\Gamma}$. This description is consistent with [16] and may be extended for metagenomes as will be shown later in Section 2.3. The number of clones starting at each position is a Poisson-distributed random variable with mean α. The probability of j clones starting at a given position is:

$$f(j;\alpha) \sim Poisson(\alpha) = \frac{\alpha^j \cdot e^{-\alpha}}{j!} \qquad (1)$$

Let G be a gene with length g. Assume that a read must contain at least T base pairs, $L>T$, of G in order for G to be detected. First, we estimate the number of reads containing a detectable part of G. Then, there are $2(L+g-2T)$ positions at which a clone might begin (see Fig. 1) in order for G to be detected on one of its pair-end reads.

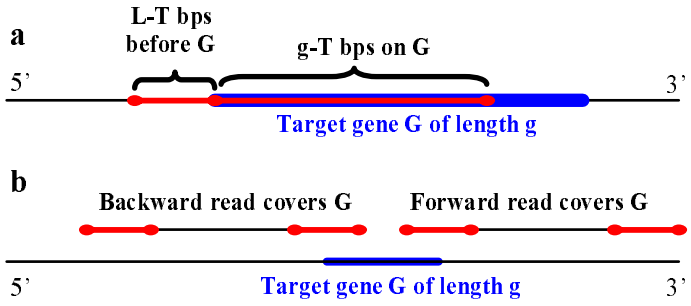

Figure 1. (a) Any read of length L that begins within the L-T base pairs before the target gene or somewhere on the first g-T base pairs on the gene will cover at least T base pairs of the gene. Overall there are L+g-2T positions in which such reads may begin. (b) Each clone may cover G with either its forward or backward reads. Combined with (a) this gives a total of 2(L+g-2T) positions.

The number of reads containing a detectable part of G can be represented by a random variable R_G that is the sum of *2(L+g-2T)* independent Poisson variables with mean α. R_G is therefore Poisson distributed with mean $\lambda_g = 2\alpha(L + g - 2T)$.

The function analysis process is based on identifying genes that belong to gene families that are captured as COGs or Pfams. Each such family contains several genes whose lengths are usually similar. Let P be a protein family composed of genes whose average length is g. In the genome of length Γ, suppose that C_P genes are associated with P. Assuming that the occurrences of the genes associated with P are independent of each other, the number of reads associated with such genes can be modeled by a random variable R_P, which is the sum of C_P independent Poisson variables, each with mean λ_g. As deduced above, R_P is Poisson distributed with mean

$$\lambda_P = 2\alpha(L + g - 2\mathrm{T})\cdot C_P \tag{2}$$

The numbers resulting from BLASTing the metagenome reads against the protein families database yield an empirical estimate for R_P, while we are interested in estimating C_P for computing the frequency of P from all genes in the genome of length Γ. From Equation 2 we have,

$$\mathrm{C_P} = \frac{\lambda_P}{2\alpha\cdot(L + g - 2\mathrm{T})} \tag{3}$$

Substituting λ_P with R_P, we get an estimator $\hat{C}_P$ for C_P when the value of R_P is known as follows:

$$\hat{\mathrm{C}}_\mathrm{P} = \frac{R_P}{2\alpha\cdot(L + g - 2\mathrm{T})} \tag{4}$$

In the case of metagenomes, all the required parameters for computing C_P are available, except for the total length Γ of the genome, required for the computation of α.

Let D be the set of protein families such that reads are associated with members of D. Then, the proportion of reads, F_P, associated with a protein family P can be approximated by dividing $\hat{C}_P$ by the total number of genes belonging to any gene family:

$$F_P = \frac{\hat{C}_P}{\sum_{Q\in D}\hat{C}_Q} = \frac{R_P}{2\alpha\cdot(L + g - 2\mathrm{T})}\Big/\sum_{Q\in D}\frac{R_Q}{2\alpha\cdot(L + g_Q - 2\mathrm{T})} = \frac{R_P}{(L + g - 2\mathrm{T})}\Big/\sum_{Q\in D}\frac{R_Q}{(L + g_Q - 2\mathrm{T})} \tag{5}$$

where g_Q is the average length of genes in protein family Q.

Once computed, F_P can be used for function analysis. As seen in Equation 5, α is eliminated from the expression for F_P resulting in an expression comprising only known parameters. The denominator in this case covers all genes in the genome and should be accurate enough for WGS projects of reasonable size; the numerator

depends on the observed number of reads for the gene family where higher read count means more accurate frequency estimation. A high read count is the result of longer genes and a higher number of occurrences of the gene family in the genome. Next we provide a method for estimating the accuracy of the observed frequency.

2.2 Computing Confidence Bounds

The estimation of F_P is based on the observed number of reads covering members of protein family P. In the event of significant inaccuracy in this estimation, later stages such as function comparison will also be affected and will yield incorrect conclusions. Here, we provide a method for estimating a range of possible frequencies for the gene family which may have generated the observed counts with probability higher than some user-defined threshold ε. For abundant families this range is expected to be narrow, while for rare families this range is going to be wide.

Specifically, we need to compute $F_P^{\min}$ and $F_P^{\max}$ such that for every $f \in [F_P^{\min}, F_P^{\max}]$, $\Pr[R_P \mid f] \geq \varepsilon$, and there exists no other f' such that $f \notin [F_P^{\min}, F_P^{\max}]$ and $\Pr[R_P \mid f] \geq \varepsilon$. These upper and lower bounds on the frequencies can then be used for filtering gene families for which frequencies have ranges that are too wide.

We compute the interval in the following manner. Start with a Maximum Likelihood estimation $\hat{\lambda}_P = Observed(R_P)$ for the parameter of the Poisson distribution in Equation 3. Iteratively look for factors $c^{\min}$ and $c^{\max}$ such that $\Pr\left(R \geq R_P \mid c^{\min} \cdot \hat{\lambda}_P\right) < \varepsilon$ and $\Pr\left(R \leq R_P \mid c^{\max} \cdot \hat{\lambda}_P\right) < \varepsilon$. From Equation 2 it follows that

$$C_P \alpha = \frac{\lambda_P}{2(L + g - 2\mathrm{T})} \tag{6}$$

Therefore, multiplying λ_P by a constant c is equivalent to multiplying C_P by c, since α is not changed, and a simple multiplication of the frequencies as described here is sufficient.

2.3 Transition from Genomes to Metagenomes

So far we have discussed genomes, while, in fact, we are interested in analyzing metagenomes. In this section we show that Equation 5 can be used for metagenomes as well. In order to see why, consider a metagenomic sample S containing m different species, where species i has a genome of length Γ_i and is represented by n_i members. As a first step in the WGS sequencing process all genomes in S are sheared into clones; next, N clones are chosen at random and undergo pair-end sequencing. Overall, there are $n_i\Gamma_i$ base pairs in S associated with the genome of species i, and the

total length of all the genomes in S is $\sum_{j=1}^{m} n_j \Gamma_j$. As in the case of a single genome it is assumed that a clone may begin at each position on the genome of any organism in S; therefore, assuming that a total of N clones undergo pair-end sequencing, the expected number of clones extracted from the genome of species i is $N \cdot n_i \Gamma_i \Big/ \sum_{j=1}^{m} n_j \Gamma_j$. From Equation 1 it follows that the expected number of clones beginning at each position on the genome of species i is

$$\alpha_i^S = \left(\frac{N \cdot n_i \Gamma_i}{\sum_{j=1}^{m} n_j \Gamma_j} \right) \cdot \frac{1}{\Gamma_i} = \frac{N \cdot n_i}{\sum_{j=1}^{m} n_j \Gamma_j} \tag{7}$$

Given species i with genome of length Γi and C_p^i genes of average length g on this genome that are associated with protein family P, it follows from Equation 2 that the number of reads that cover genes associated with P in genome i (R_p^i) is a Poisson random variable with mean

$$\lambda_P^i = 2\alpha_i^S \left(L + g - 2\mathrm{T}\right) \cdot C_P^i \tag{8}$$

The total number of reads covering genes associated with P anywhere in the sample, R_P^S, can now be expressed as the sum of m Poisson random variables. Using Equation 7, this is a Poisson random variable with mean

$$\lambda_p^s = \sum_{i=1}^{m} \lambda_p^i = 2(L + g - 2\mathrm{T}) \cdot \sum_{i=1}^{m} \alpha_i^s \cdot C_p^i = \frac{2N(L + g - 2\mathrm{T})}{\sum_{j=1}^{m} n_j \Gamma_j} \sum_{i=1}^{m} n_i \cdot C_p^i = \frac{2N(L + g - 2\mathrm{T})}{\sum_{j=1}^{m} n_j \Gamma_j} C_p^S \tag{9}$$

where C_p^S is the total number of genes associated with P in S. By replacing λ_P^S with R_P^S, which is the observed number of reads covering genes associated with P in S, we obtain an estimator $\hat{C}_P^S$ for C_P^S:

$$\hat{C}_P^S = \frac{R_P^S \cdot \sum_{j=1}^{m} n_j \Gamma_j}{2N \cdot \left(L + g - 2\mathrm{T}\right)} \tag{10}$$

Thus, the proportion of reads in S that are associated with P is given by

$$F_P^S = \frac{\hat{C}_P^S}{\sum_{Q \in D} \hat{C}_Q^S} = \frac{R_P^S \cdot \sum_{j=1}^{m} n_j \Gamma_j}{2N \cdot (L + g - 2\mathrm{T})} \Bigg/ \sum_{Q \in D} \frac{R_Q^S \cdot \sum_{j=1}^{m} n_j \Gamma_j}{2N \cdot (L + g_Q - 2\mathrm{T})} = \tag{11}$$

$$= \frac{R_P^S}{(L + g - 2\mathrm{T})} \Bigg/ \sum_{Q \in D} \frac{R_Q^S}{(L + g_Q - 2\mathrm{T})}$$

For all practical purposes, Equation 11 is the same as Equation 5 since it does not contain any information related to S, such as the number of species and their genome lengths.

2.4 Performing Function Comparison

Given a reference metagenome B and a metagenome of interest A, for a protein family P we are interested in estimating the significance of the difference between the estimated frequencies of P in B and A, F_P^B and F_P^A, respectively. Here we propose a novel approach for assessing the significance of the difference between the two frequencies.

Recall that F_P^B and F_P^A are derived from the observed number of reads containing some part of P in A and B, R_P^B and R_P^A, respectively. We are interested in computing $equivalent(R_P^A, B)$, the number of reads in B that would yield the frequency F_P^A. Considering Equation 5 (or its metagenomic equivalent Equation 11), this goal can be formulated by

$$F_P^A = \frac{equivalent(R_P^A, B)}{(L + g - 2\mathrm{T})} \Bigg/ \sum_{Q \in D} \frac{R_Q^B}{(L + g_Q - 2\mathrm{T})} \tag{12}$$

This provides an explicit expression for $equivalent(R_P^A, B)$:

$$equivalent(R_P^A, B) = F_P^A \cdot (L + g - 2\mathrm{T}) \cdot \sum_{Q \in D} \frac{R_Q^B}{(L + g_Q - 2\mathrm{T})} \tag{13}$$

Since all components of Equation 13 are available, $equivalent(R_P^A, B)$ can be computed. Taking R_P^B as the maximum likelihood estimation for λ_P^B, the parameter of the Poisson variable describing the distribution of reads covering members of P in B, it is possible to compute the probability to observe at least (or at most) $equivalent(R_P^A, B)$ reads covering P in B. This probability may be interpreted as a significance measure for the difference between the frequencies F_P^B and F_P^A.

3 Results

In order to analyze the performance of our model we have used three types of data: (i) simulated data of single genomes, (ii) real data from genome sequencing projects, and (iii) real metagenomic data. Simulated data is generated according to our assumptions, and therefore provides an opportunity to evaluate the method under controlled conditions. In addition, it is possible to generate different datasets with varying parameters such as gene family distributions, coverage, and population structure. As explained above, single genomes may be regarded as a special case of metagenomes; they are simpler to analyze and generate and were therefore used for evaluation. Genome sequencing projects provide the opportunity to test our framework on real data (generated in a similar way to metagenomes) with known gene family statistics. Real metagenomes lack information regarding the underlying statistics but provide us with the opportunity to check real-world cases. Evaluation done using such data is mostly qualitative.

3.1 Evaluation of Simulated Data

We have written a program that generates simulated genomes and WGS sequencing projects, including genes and their families and reads' distribution across the genome. The simulator also implements our statistical framework, as well as the read-counts based frequency estimator for protein families. In order to compare the performance of the two frequency estimators we have used the following quality measure:

$$Q(M(P)) = \log_2 \frac{M(P)}{F_P^{True}} \tag{14}$$

in which F_P^{True} is the true frequency of protein family P in the (meta)genome and $M(P)$ is the estimation. When $M(P) = F_P^{True}$ it is the case that $Q(M(P)) = 0$, while $Q(M(P))<0$ when $M(P)$ underestimates F_P^{True}, and $Q(M(P))>0$ when $M(P)$ overestimates F_P^{True}.

As a first step we have tested the read-counts bias. For this purpose we have synthesized three genomes of length 10 Mbps each: (i) a genome that contains genes associated with 200 protein families of lengths ranging between a few dozens to a few thousands bps and a similar number of genes associated to each family, (ii) a genome containing genes associated with several hundreds of protein families of constant length but a different number of genes associated with each family, and (iii) a combination of (i) and (ii) – genomes with protein families of varying lengths and varying number of associated genes. Other than the length and the number of genes associated with the protein families all other parameters remained constant across all simulated genomes, including the number of reads (1M) and their lengths (942 bps) and the distance between genes (60 bps). Note that the resulting coverage is extremely high; this was done in order to test the behavior of each method when all data is available.

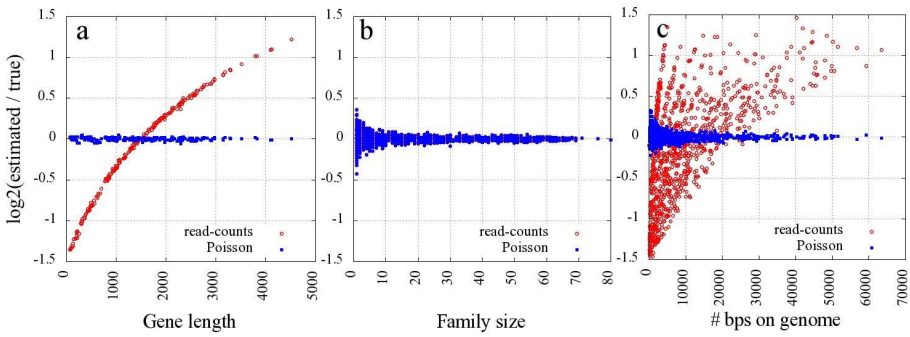

Figure 2. Quality of frequency estimations as a function of (a) gene size, (b) family size, and (c) a combination of both. In all cases coverage is extremely high (94.2X) in order to demonstrate the inherent differences between the methods even when redundant data is available.

Fig. 2 summarizes the results of these three tests. As can be seen from Fig. 2a, our framework scales with any gene length while read-counts based estimations tend to overestimate the frequencies of long genes and to underestimate the frequencies of short genes. When all genes across the genome are of exactly the same length both methods yield the same estimations, regardless of differences in the number of genes associated with each family (Fig. 2b). This is not surprising considering the fact that when $g_i=g_j$ for any two families i and j, Equation 5 is reduced to simple division of read-counts with no other scaling required. As can be seen from Fig. 2b, frequency estimations are less accurate for small families, which is also reasonable. Our method remains reliable also when both the number of genes associated with a family and their lengths vary (Fig. 2c), while the read-counts approach produces inaccurate estimations. The accuracy of the method increases when more bps are associated with family P. A large number of bps associated with family P may be the result of both long genes, or a large amount of genes associated with the family.

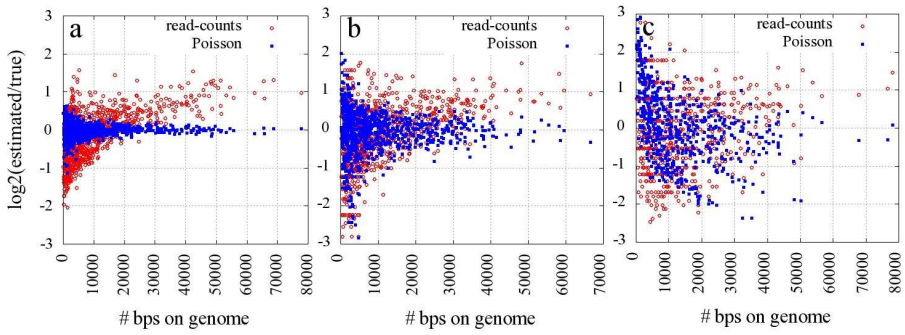

Figure 3. The influence of coverage on frequency estimation for (a) coverage=9.42X, (b) 0.942X and (c) 0.0942X.

Next, we were interested in analyzing the influence of different coverage levels on the quality of the estimations. For this purpose we have used the setup of the genome whose behavior is described in Fig. 2c and tried a different number of reads each time (resulting in different coverage levels). At high coverage levels typical to genome sequencing projects (Fig. 3a) our method remains relatively accurate. As expected, the quality of the estimations decreases with lower coverage levels (Fig. 3, b and c). In all cases our method outperforms the read-counts estimator and provides significantly better results (see Fig. 4b).

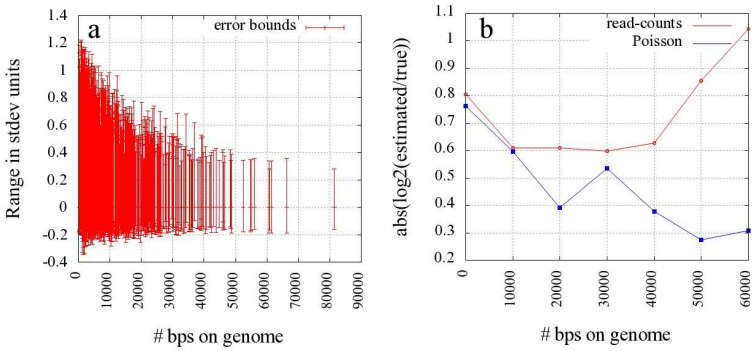

Figure 4. (a) Reliability bounds as a function of the number of base pairs associated with protein families, ε=0.1 used. (b) Log-ratio error as a function of the number of base pairs associated with different families. For each range [x·10Kbps, (x+1)·10Kbps] median log-ratio error for all families in the range is shown. Confidence bounds become narrow with the increase in the family's share on the genome, in line with the increased accuracy of frequency estimation.

Fig. 4a shows confidence bounds computed for different families as a function of the number of bps associated with the family. As can be clearly seen, the confidence range decreases (in terms of standard deviation units) with the more associated bps available. This property of the confidence bounds scheme is desirable, considering the fact that our method derives better frequency estimations for protein families to which many bps are associated across the genome (Fig. 4b). This can also be used for filtering families with a wide confidence range.

3.2 Evaluation of Data Derived from WGS Sequencing Projects

Raw data of WGS sequencing projects of single genomes is available through NCBI's Trace DB. Using data from [18] we were able to reconstruct read coordinates and also to identify pairs of reads that belong to the same clones. Overall we were able to reconstruct information for 44 genomes (see Appendix for complete list of genomes). COG families were assigned to these genomes by BLASTing (blastx) their genes (extracted from the IMG system [7]) against the COG database with an e-value threshold of 1e-50. Next the number of reads containing members of each COG

family was decided and used for producing frequency estimations by both our and the read–counts approach. Evaluation of estimations was done using Equation 14 as before. Results are summarized in Fig. 5 (see figure legend for description of figure generation process). As can be observed from the figure, behavior of both estimation methods fit nicely the behavior that was predicted in our simulations. While our method generated predictions that are, on average, close to the real frequencies and are not affected by the length of family members, the read-counts approach produces estimations that are biased by gene length. Both methods are not affected by the number of family members.

It is important to note that while our method produces unbiased estimations, its frequency estimations are, on average, 20-30% off the real value for all gene lengths. The estimation error is affected by the number of base pairs associated with a family on the genome (more base pairs yield better estimations). This error rate is smaller than or equal to the error of the read-counts approach for all cases.

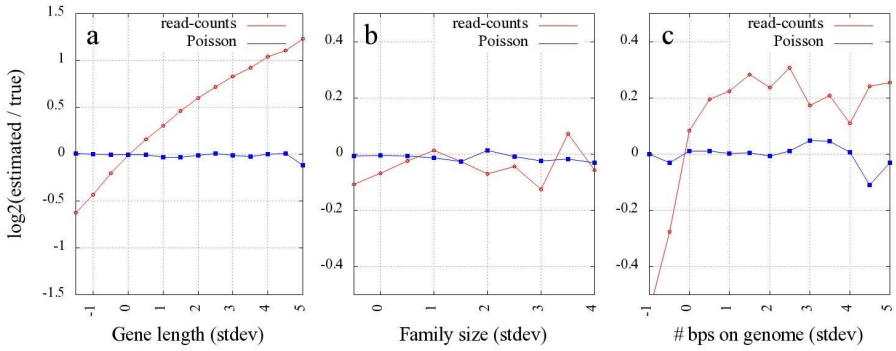

Figure 5. Summary of data analysis for 44 single genome WGS projects.(a) Log-ratio between estimations and true frequencies, sorted by family length. For each genome, average and standard deviation of gene lengths were computed and used for binning evaluations into x0.5 standard deviation bins. Next the median for each bin was computed and kept with medians of the same bin from other genomes. Last the median over all medians for each bin was computed and used for generating the figure. (b) Same process as before using number of members for each family as the key. (c) Same process, using number of base pairs occupied by each family on the genome as the key.

3.3 Evaluation of Real Data

While metagenomes are the goal of our method, it is impossible to validate the results generated from them. However, in order to get a sense of what the results look like we have compared two depths datasets (10 and 500 meters) from [6]. In order to assign genes to protein families we have used the COG database: each read was BLASTed against the COG database with the best hit being assigned to the read. The 10 meters dataset is composed of 7,842 reads whose average length is 954 bps, while the 500 meters dataset consists of 9,027 reads with average length of 971 bps. Overall a total

of 2,426 genes from the 10 meters dataset have been assigned to COG families and 2,997 genes from the 500 meters were similarly assigned.
While family frequencies usually show correlation with the number of hits, there are many exceptions: for example, COG0187 was found 16 times in the 10 meters dataset but was assigned a frequency of 0.53%, less than the frequency of COG1024 with only 13 hits (0.59%). While COG1024's average length is 270 bps the length of COG0187 is 685 bps which explains why it received more hits. The frequency of COG0085 with 17 hits in the 500m dataset is 0.35%, less than half the frequency of COG0316 (0.77%) with almost the same number of hits (18). Again, the difference is the result of gene lengths. Using the 500 meters dataset for the background distribution we were able to discover COG families whose frequency significantly differs between the two datasets. The results are not always observable by simply considering the read-counts; for example, the *p*-value assigned to the difference between the frequencies of COG0719 (ABC-type transport system involved in Fe-S cluster assembly, permease component) was quite significant ($7 \cdot 10^{-5}$) despite a non-impressive difference in the read-counts (9 *vs.* 3 in the 10 and 500 meter datasets, respectively). Such cases convincingly demonstrate the need for a reliable statistical model when doing function analysis.

4 Discussion

All the methods suggested to-date for the functional comparison of metagenomes suffer from a read-counts bias. Our method is the first one to correct this problem, allowing for more reliable and credible analyses of this kind of data. This is achieved by offering a comprehensive statistical framework, based on sound theoretical foundations, that – with reasonable assumptions – allows us to assess the significance of the evidence for the presence of a protein family of interest in a given genome. Moreover, we present the most extensive evaluation performed to-date of such methods, which includes – for the first time – validation on real data. Our results suggest that the proposed framework provides a good approximation to the statistics of gene families in WGS projects. Observed estimation errors may be explained by reasons that are related to the definition of gene families. For example, the assignment of a gene to a family is hardly ever to the whole length of the family but rather to some part of it. This may result from the target family carrying several domains, while the assigned gene may carry only one of them. However, when we compute the expected frequency we consider the whole family's length, which is clearly unrealistic. The use of domain-specific databases, such as Pfam, may address this problem and improve frequency estimations.
Whereas our framework is most suitable for gene family based functional analysis, we believe that it is not suitable for databases organized according to high-level functionalities such as pathways (*e.g.* KEGG, [23]) or subsystems (*e.g.* SEED, [15]). This is also true for other methods described in the past, including read-counts based approaches, and is the result of the complex nature of these databases. Each family-equivalent in these databases is composed of several low-level functions, some of which may be shared by several pathways or subsystems. Moreover, the meaning of

the results is unclear since we do not know how to interpret likely cases in which only some of the members of some high-level organization are present. Therefore it is necessary to better define the problem for these cases and develop an appropriate framework.

Acknowledgments. We would like to thank Rotem Sorek for sharing his data and to Zohar Yakhini, Ernest Szeto, and Konstantinos Mavromatis for fruitful discussions. The work presented in this article was partially supported by the Director, Office of Science, Office of Biological and Environmental Research, Life Sciences Division, US Department of Energy under Contract No. DE-AC03-76SF00098.

References

1. Beja, O., Aravind, L., Koonin, E.V., Suzuki, M.T., Hadd, A. *et al.*: Bacterial Rhodopsin: Evidence for a New Type of Phototrophy in the Sea. Science 289(5486), 1902--1906 (2000)
2. Venter, J.C., Remington, K., Heidelberg, J.F., Halpern, A.L., Rusch, D. *et al.:* Environmental Genome Shotgun Sequencing of the Sargasso Sea. Science 304(5667), 66--74 (2004)
3. Angly, E.A., Felts, B., Salamon, P., Edwards, E.A., Carlson, C. *et al.:* The Marine Viromes of Four Oceanic Regions. PLoS Biol. 4(11) (2006)
4. Tyson, G.W., Chapman, J., Hugenholtz, P., Allen, E.E., Ram, R.J. *et al.*: Community Structure and Metabolism through Reconstruction of Microbial Genomes from the Environment. Nature 428(6978), 37--43 (2004)
5. Gill, S.R., Pop, M., Deboy, R.T., Eckburg, P.B., Turnbaugh, P.J. *et al.*: Metagenomic Analysis of the Human Distal Gut Microbiome. Science 312(5778), 1355--1359 (2006)
6. DeLong, E.F., Preston, C.M., Mincer, T., Rich, V., Hallam, S.J. *et al.*: Community Genomics among Stratified Microbial Assemblages in the Ocean's Interior. Science 311(5760), 496--503 (2006)
7. Markowitz, V.M., Szeto, E., Palaniappan, K., Grechkin, Y., Chu, K. *et al.*: The Integrated Microbial Genomes (IMG) System in 2007: Data Content and Analysis Tool Extensions. Nucleic Acids Res. 36 (Database Issue), DS528--DS533 (2008)
8. Tatusov, R.L., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Kiryutin, B. *et al.*: The COG Database: an Updated Version Includes Eukaryotes. BMC Bioinformatics 4, 41 (2003)
9. Finn, R.D., Tate, J., Mistry J., Coggill, P.C., Sammut, J.S. *et al.*: The Pfam Protein Families Database. Nucleic Acids Res. 36 (Database Issue), D281--D288 (2008)
10. Haft D.H., Selengut J.D., White, O.: The TIGRFAMs Database of Protein Families. Nucleic Acids Res. 31, 371--373 (2003)
11. Rodriguez-Brito, B., Rohwer, F., Edwards, R.A.: An Application of Statistics to Comparative Metagenomics. BMC Bioinformatics 20(7), 162 (2006)
12. Tringe, S.G., von Mering, C., Kobayashi, A., Salamov, A.A., Chen, K. *et al.*: Comparative Metagenomics of Microbial Communities. Science 308(5721), 554--557 (2005)
13. Rusch, D.B., Halpern, A.L., Sutton, G., Heidelberg, K.B., Williamson, S. *et al.:* The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. PLoS Biol. 5(3), e77 (2007)
14. Yooseph, S., Sutton, G., Rusch, D.B., Halpern, A.L., Williamson, S.J. *et al.*: The Sorcerer II Global Ocean Sampling Expedition: Expanding the Universe of Protein Families. PLoS Biol. 5(3), e16 (2007)
15. Overbeek, R., Begley, T., Butler, R.M., Choudhuri, J.V., Chuang, H.Y. *et al.*: The Subsystems Approach to Genome Annotation and its Use in the Project to Annotate 1000 Genomes. Nucleic Acids Res. 33, 5691--5702 (2005)

16. Lander, E.S., Waterman, M.S.: Genomic Mapping by Fingerprinting Random Clones: a Mathematical Analysis. Genomics 2(3), 231--239 (1988)
17. Schloss, P.D., Handelssman, J.: A Statistical Toolbox for Metagenomics: Assessing Functional Diversity in Microbial Communities. BMC Bioinformatics 9(34) (2008)
18. Sorek, R., Zhu, Y., Creevey, C., Francino, M.P., Bork, P., Rubin, E.M.: Genome-wide Experimental Determination of Barriers to Horizontal Gene Transfer. Science 318(5855), 1449--1452 (2007)
19. Mavromatis, K., Ivanova, N., Barry, K., Shapiro, H., Goltsman, E. *et al.*: Use of Simulated Data Sets to Evaluate the Fidelity of Metagenomic Processing Methods. Nature Methods 4, 495--500 (2007)
20. Sanger, F., Coulson, A.R., Hong, G.F., Hill, D.F., Petersen, G.B.: Nucleotide Sequence of Bacteriophage Lambda DNA. J. Mol. Biol. 162, 4 (1982)
21. Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F. *et al.*: Whole-genome Random Sequencing and Assembly of *Haemophilus influenzae Rd*. Science 269(5223), 496--512 (1995)
22. Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J. *et al.:* The Sequence of the Human Genome. Science 291(5507), 1304-1351 (2001)
23. Kanehisa, M., Goto, S.: KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 28, 27--30 (2000)
24. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J.: Basic Local Alignment Search Tool. J. Mol. Biol. 215, 403--410 (1990)
25. Martín-Cuadrado, A.B., López-García, P., Gottschalk, G., Rodríguez-Valera, F.: Metagenomics of the Deep Mediterranean, a Warm Bathypelagic Habitat. PLoS ONE 2, 914 (2007)
26. Warnecke, F., Luginbuhl, P., Ivanova, N., Ghassemian, M., Richardson, T.H. *et al.*: Metagenomic and Functional Analysis of Hindgut Microbiota of a Wood Feeding Higher Termite. Nature 450, 560-565 (2007)
27. Marchler-Bauer, A., Anderson, J.B., Chitsaz, F., Derbyshire, M.K., DeWeese-Scott, C. *et al.*: Specific Functional Annotation with the Conserved Domain Database. Nucleic Acids Res. 37(Database Issue), D205—D210

Appendix: Genomes Used for Real Data Analysis

Table 1. Information for participating genomes in the real data analysis.

Organism	Genome Size	Coverage	# Clones	Avg. Read Length
Acidobacterium sp. Ellin 345	5650368	11.1	27373	1035.9
Alkalilimnicola ehrlichei MLHE-1	3275944	12.8	16913	1014.4
Anabaena variabilis ATCC29413	6365727	17.1	36406	934.8
Anaeromyxobacter dehalogenans 2CP-C	5013479	12.0	15691	1007.5
Bacillus anthracis str ames	5227293	20.6	38924	856.4
Baumannia cicadellinicola	686194	29.0	2394	980.0
Campylobacter jejuni rm1221	1777831	12.2	9316	881.1
Carboxydothermus hydrogenoformans z-2901	2401520	14.8	12188	910.7
Chlorobium tepidum tls	2154946	13.8	10995	814.2
Colwellia psychrerythraea 34h	5373180	11.5	27652	902.2
Dechloromonas aromatica RCB	4501104	24.9	59247	850.7
Dehalococcoides ethenogenes 195	1469720	16.5	4861	907.3
Ehrlichia canis Jake	1315030	23.1	8229	968.3
Ehrlichia chaffeensis str Arkansas	1176248	14.6	7631	925.8
Frankia sp. CcI3	5433628	15.5	29142	990.6
Geobacter sulfurreducens pca	3814139	14.2	18579	916.6
Listeria monocytogenes str 4b f2365	2905187	17.7	21550	929.7
Methanococcoides burtonii DSM6242	2575032	24.0	34876	916.8
Methanospirillum hungatei JF-1	3544738	13.0	18776	958.9
Methylococcus capsulatus str bath	3304561	11.2	16517	884.3
Moorella thermoacetica ATCC 39073	2628784	23.5	21032	989.5
Neorickettsia sennetsu str miyayama	859006	13.7	4996	917.8
Nitrobacter hamburgensis X14	4406967	12.5	20259	935.3
Nitrobacter winogradskyi Nb-255	3402093	11.3	14703	958.4
Nitrosococcus oceani ATCC 19707	3481691	14.4	20047	970.0
Prochlorococcus marinus sp. NATL2A	1842899	11.0	10561	960.4
Prochlorococcus marinus str. MIT 9312	1709204	13.6	9834	965.9
Prochlorococcus sp. CC9605 (oligotrophic)	2510659	17.0	16269	969.6
Pseudoalteromonas atlantica T6c	5187005	10.0	21681	1035.6
Psychrobacter cryohalolentis K5	3059876	12.2	13397	953.8
Rhodopseudomonas palustris BisB18	5513844	11.1	25015	1023.6
Rhodopseudomonas palustris BisB5	4892717	11.4	20142	1024.8
Rhodopseudomonas palustris HaA2	5331656	11.7	25012	959.5
Rubrobacter xylanophilus DSM 9941	3225748	14.5	14321	1193.9
Shewanella denitrificans OS217	4545906	16.6	27125	967.2
Shewanella frigidimarina NCMB400	4845257	11.9	23999	978.6
Shewanella sp. MR-4	4706287	11.5	21464	1020.4
Streptococcus agalactiae a909	2127839	13.9	12135	914.9
Synechococcus sp. PCC 7942 (elongatus)	2695903	13.0	15863	891.1
Thermobifida fusca YX	3642249	31.1	39988	850.3
Thiomicrospira crunogena, XCL-2	2427734	18.5	18675	955.0
Thiomicrospira denitrificans ATCC 33889	2201561	12.0	9992	958.4
Treponema denticola atcc 35405	2843201	15.6	16935	901.9
Trichodesmium erythraeum	7750108	17.3	49207	966.3