Self-organizing Approach for the Human Gut Meta-genome

Jianfeng Zhu[1], Songgang Li[1] and Wei-Mou Zheng[*,1,2]

[1]*Beijing Genomics Institute, Shenzhen (BGI-SZ), Shenzhen 518083, China*

[2]*Institute of Theoretical Physics, Academia Sinica, Beijing 100190, China*

Abstract: We extend the self-organizing approach for annotation of a bacterial genome to analyzing the raw sequencing data of the human gut metagenome without sequence assembling. The original approach divides the genomic sequence of a bacterium into non-overlapping segments of equal length and assigns to each segment one of seven 'phases', among which one is for the noncoding regions, three for the direct coding regions to indicate the three possible codon positions of the segment starting site, and three for the reverse coding regions. The noncoding phase and the six coding phases are described by two frequency tables of the 64 triplet types or 'codon usages'. A set of codon usages can be used to update the phase assignment and vice versa. After an initialization of phase assignment or codon usage tables, an iteration leads to a convergent phase assignment to give an annotation of the genome. In the extension of the approach to a metagenome, we consider a mixture model of a number of categories of genomes. The Illumina Genome Analyzer sequencing data of the total DNA from faecal samples are then examined to understand the diversity of the human gut microbiome.

Keywords: Human gut meta-genome, codon usages, self-organizing genome annotation.

1. INTRODUCTION

The majority of microbes in our body resides in the gut. They are crucial for human life. Metagenomic sequencing is a powerful tool for analyzing the diversity of bacterial populations in various environments [1]. Targeted sequencing of 16S ribosomal RNA gene (rRNA) revealed that two bacterial divisions, the Bacteroidetes and the Firmicutes, constitute over 90% of the known phylogenetic categories and dominate the distal gut microbiota [2]. Substantial diversity of the gut microbiome is seen between individuals, with infants in particular [3]. Changes of gut microbiome may be associated with bowel diseases and obesity [4, 5]. A shift towards Firmicutes can be observed in obese individuals. The ratio between Firmicutes and Bacteroidetes dynamically reflects the overall weight condition of an individual.

Reductions in sequencing costs allow wider scale surveys by shotgun-sequencing an entire bacterial population in the human gastro-intestinal tract [6, 7]. The Illumina-based gut metagenome sequencing of 124 individuals from Denmark and Spain generated a 576.7 Gb sequence, from which assembly and characterization of 3.3 million nonredundant microbial ORFs were derived. The ORF set probably covers most of the prevalent human intestinal microbial genes. The gene pool is largely shared among individuals of the cohort, which includes healthy, overweight and obese individuals, as well as inflammatory disease patients.

At least 80% of 3.3M ORFs map to the 0.32M genes (target genes) of the 89 frequent reference microbial genomes in the human gut. When aligning reads onto a nonredundant set of 650 sequenced bacterial and archaeal genomes representative of 932 publicly available genomes, at a 1% coverage (~40 kb for a typical gut bacterial genome), 18 species are detected in all individuals, and 75 in half of the individuals.

Bacterial populations generally consist of an un-even mixture of organisms. The uneven level of coverage renders useless many statistical approaches. New algorithms are therefore necessary to deal with the sequence data of such mixed populations.

A deep sequencing of the total DNA from faecal samples of 77 Asian (Han) adults has been conducted in BGI, Shenzhen, from which 0.2Tb of sequence data were generated. In this paper we extend the self-organizing approach for annotation of a bacterial genome to the analysis of the raw data of the gut meta-genome without sequence assembling. The original approach divides the genomic sequence of a bacterium into non-overlapping segments of equal length and assigns each segment to one of the seven 'phases', among which one is for the noncoding regions, three for the direct coding regions to indicate the three possible codon positions of the segment starting site, and three for the reverse coding regions. The noncoding phase and the six coding phases are described by two frequency tables of the 64 triplet types or 'codon usages'. A set of codon usages can be used to update the phase assignment and vice versa. After an initialization of phase assignment or codon usage tables, an iteration leads to a convergent phase assignment to give an annotation of the genome. The extension of the approach to a meta-genome is to consider a mixture model of a number of categories of genomes. The Illumina Genome Analyzer sequencing data of the total

*Address correspondence to this author at the Beijing Genomics Institute, Shenzhen (BGI-SZ), Shenzhen 518083, China; Tel: 86-10-62541820; Fax: 86-10-62562587; E-mail: zhengwm@genomics.org.cn

1875-0362/12

DNA from faecal samples are then examined to understand the diversity of the human gut microbiome.

2. SELF-ORGANIZING APPROACH FOR *R. PROWAZEKII* GENOME

Genome annotation by statistical methods is based on various statistical models of genomic sequences, one of the most popular being the inhomogeneous, three-period Markov chain model for protein-coding regions with an ordinary Markov model for noncoding regions [8, 9]. The 'codon usage' model discussed here is the independent random chain model of non-overlapping triplets, and corresponds to an inhomogeneous Markov model of order 2 [10]. Most of the current computer methods for locating genes require some prior knowledge of statistical properties of the genome sequence, particularly codon frequencies, from a sizable training set. An automatic modeling procedure to partition genome sequences into coding and non-coding segments is desirable [11]. Such a procedure using the codon usage measure can be proposed as follows [12]. To be explicit, we take *Rickettsia prowazekii* genome as an example. (Its size is about 1.11 Mb.) We first divide the genome sequence into non-overlapping windows with length being a multiple of 3, say 99. Roughly speaking, we may classify these windows into seven categories or 'phases'. The first one (phase 0) consists of those windows falling in non-coding regions. The windows belonging to the other six fall in coding regions, either direct (phases 1 -- 3) or reverse (phases 4 -- 6). We assign their phases according to the codon position of their first nucleotide (site). For example, the phase of a direct coding window $W = s_i s_{i+1} \cdots s_{i+98}$ is 1 (2 or 3) if triplet $s_i s_{i+1} s_{i+2}$ ($s_{i-1} s_i s_{i+1}$ or $s_{i-2} s_{i-1} s_i$) forms a codon. (We ignore that a few windows may fall between coding and noncoding regions).

The codon usage model has two sets of triplet frequencies, one for coding regions and the other for noncoding regions. We shall call them 'triplet tables' or simply tables. The noncoding table is obtained by counting triplets of each type in windows of phase 0. When obtaining the coding table, besides shifting windows to the left or right by one site, for a reverse coding window we have to take its Crick-Watson dual (by the operation of interchange $A \leftrightarrow T$, $G \leftrightarrow C$ and then reverse). If we know the two tables, for a given sequence we can calculate seven probabilities, each of which corresponds to one phase. The one of phase 0 is obtained as a product of factors, and each factor is obtained by looking up each successive non-overlapping triplet of the sequence in the noncoding table. The values of the other six phases are all obtained from the coding table. For phases 4, 5 and 6, we have to take the Crick-Watson dual of the sequence; for phases other than 1 and 4, we have to drop possible incomplete codons at both ends. (To make the seven probabilities to be directly comparable, we may also drop one triplet for phases 0, 1 and 4.) Using these seven values, we may infer the phase of sequence W by the greedy approach as

$$\theta^* = \arg\max_\theta P(W \mid \theta), \quad (1)$$

where $P(W \mid \theta)$ is the probability for W at phase θ, or infer the probability for W to have phase θ as

$$P(\theta \mid W) \propto P(W \mid \theta) P(\theta), \quad (2)$$

where $P(\theta)$ is a prior probability for phase θ.

At the beginning, we are given an unannotated genome sequence, and we know neither the two tables, nor the phases of windows. Initially, we assign the windows in an arbitrary way, either randomly or periodically, to seven phases. (This will imply that the proportion of each phase is even.) We may then estimate the two tables. Using the tables, the probabilities of seven phases are calculated for each window. According to which of the seven probabilities is the largest, the phase assignment of each window is updated. We then repeat the iteration until a convergent assignment of windows is reached. The convergence is rather fast. Compared with the known annotation of the genome, the accuracy rate for window phases to be correctly identified is 85.4%. The flowchart of the approach is shown in Fig. (**1**). We may as well start the iteration by initializing the two frequency tables. A simple way to do this is to extract the tables using an annotated genome, say that of *Escherichia coli*. This is also shown in the figure.

Shifting all the windows by three nucleotides, we again repeat the iteration to obtain a new convergent phase assignment. Repetition of 32 such shifts covers the window width. In this way, we obtain 33 phase assignments for every triplet (except for a few triplets at the two ends). The triplets with the 33 identical assignments cover 57.4% of the genome, and the accuracy rate reaches 98.3%.

We have examined the above self-organizing approach for the independent triplet model on several prokaryotic genomes. Its effectiveness has been verified.

We introduce the following measure for the distance between two distributions p and q

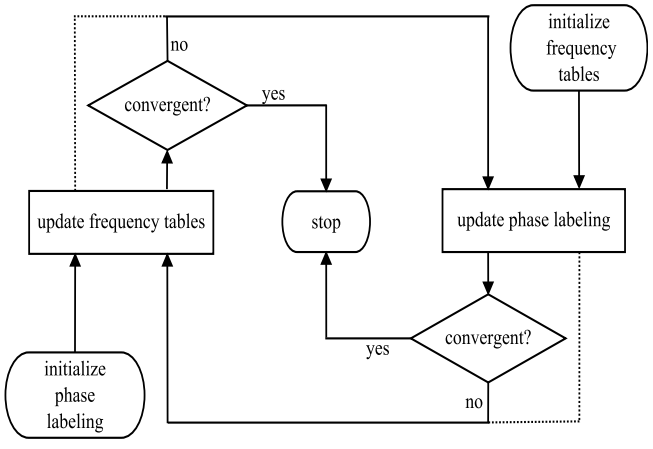

Fig. (1). Flowchart of the self-organizing approach. The algorithm starts with initializing either the two frequency tables for coding and noncoding regions or the label assignment for segments or reads. One of the two convergence verifications may be bypassed (shown by the dashed line).

$$d(p,q) = \sum_i \frac{2(p_i - q_i)^2}{p_i + q_i} \quad (3)$$

(which is the leading term of the Kullback-Leibler distance when expanded around $p_i = q_i$). For convenience, here we multiply d by 32, and use $D(p,q) = 32d(p,q)$. The distance between the noncoding and coding table estimated from the annotated *E. coli* genome is 9.24, while the distance between the predicted coding distribution and the one extracted from the known annotation is less than 0.15.

3. SELF-ORGANIZING APPROACH FOR THE HUMAN GUT METAGENOME

Deep metagenomic sequencing provides the opportunity to explore the existence of a common set of microbial species in the human gut metagenome, as well as the variability in an abundance of microbial species across individuals. The Illumina Genome Analyzer (GA) technology has been used to perform deep sequencing of the total DNA from faecal samples of 77 Han Chinese adults, including 38 normal and 39 diabetic samples, in BGI, Shenzhen. The read length is 90 bp. The number of reads of each sample is between 16M and 46M with the mean being 31M (except for one sample of 8M reads). Diabetic sample DLF014 and normal sample NLF008, both containing 30 M reads, are chosen for a thorough inspection on the self-organizing approach in our study of the human gut metagenome. The remaining 75 samples are used for further testing the utility of the approach.

Gut flora is the largest reservoir of human microbiota. Somewhere between 300 and 1000 different species live in the gut. It is almost a hopeless task to investigate each species of the metagenome from the sequencing data. A recent study claimed that bacterial ecosystems divide individuals into three groups [13,14]. In many cases, some general characteristics of the metagenome should play an important role. We shall regard a metagenome as a mixture of finite number of submicrobiota or coarse categories. The total number of parameters for a mixture model of k categories is $[(64-1)\times 2 + (2-1)]\times k + (k-1) \approx 128k$, where we have required that all the six coding phases have the same fraction, and the number of independent phase fractions is then only $2-1=1$ instead of $7-1=6$. Taking advantage of the minimal number of parameters to reduce the interference of meta-optimal solutions, we shall first consider the simplest case of $k=1$. Since the contrast between coding and noncoding regions is extraordinary, even this simple model is still informative, and also helpful for further refined modeling.

3.1. Single-category Model

In this model the whole metagenome is viewed as if it were just a single genome. We are facing short reads instead of an assembled long genomic sequence. To initiate the self-organizing iteration, we may either randomly assign a phase to each read, or use some reasonable tables of triplet frequencies. For the latter, we may extract such tables from a known genome. A good candidate could be a typical genome of species present in the gut flora. We have extracted the tables of *Escherichia coli*, *Bacteroides vulgatus*, and *Firmicutes ruminococcus* from their genomes. The distribution distances D between their tables are listed in Table **1**. It is clear that any coding tables differ notably from noncoding ones, and different species show their diversity mainly in their coding tables.

We test sample DLF014 for the initialization with the three sets of extracted tables and two random assignments of the read phase. For the convergence criterion we require that the distance D between two successive iterations for each table should be less than 0.05 (which is about one tenth of the smallest distance listed in Table **1**). It takes only four iterations for an initialization using tables extracted from either of the three genomes to obtain convergence while it takes ten iterations for initializations using random phase assignment. As an example, the convergence of iterations after two initializations for DLF014 is shown in Fig. (**2**).

The five different initializations all lead nearly to a single solution. The difference between solutions using known tables and random phases, measured in the distribution distance D, is about 0.20 for coding tables and 0.02 for noncoding tables; the coincidence rate of their final read phase assignments is 0.927 ± 0.005. (If the requirement that the score of the inferred phase should be at least twice that of other phases is considered, the rate of coincidence becomes one percent higher, and the coverage is above 97.3%).

The difference between solutions using known tables of various genomes is notably smaller than the difference from the solutions using random initializations; the coincidence rate of their final read phase assignments is above 0.988. Thus, even the single-category model for metagenomes is meaningful in telling the difference between coding and noncoding reads. The proportions of reads assigned to the six coding phases are rather even ($\sim 0.118 \pm 0.007$). By taking the initialization with the tables of *E. coli* as a representative, the distances of the convergent tables from those of the three known genomes are also listed in Table **1**. The GC contents

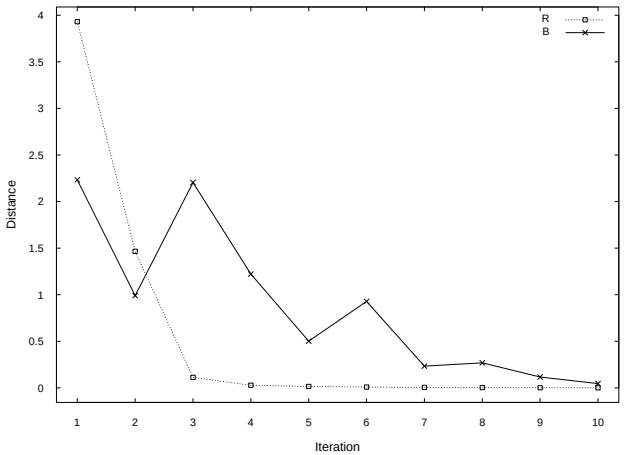

Fig. (2). The convergence of the distribution distance in iterations for sample DLF014 when a random phase assignment (R) or the triplet frequency tables of *B. vulgatus* (B) are used for initialization.

Table 1. Distribution Distances D between Triplet Frequency Tables of *E. coli* (E), *B. Vulgatus* (B), *F. Ruminococcus* (F), and the Metagenome Samples NLF014 (N), DLF008 (D). Coding and Noncoding Tables are Indicated by Letters c and n, Respectively

	B-c	F-c	D-c	N-c	E-n	B-n	F-n	D-n	N-n
E-c	6.93	11.69	3.81	4.36	9.24	14.97	13.76	19.21	22.00
B-c		6.33	6.39	2.59	10.50	8.55	8.97	10.42	12.18
F-c			7.91	6.12	14.05	12.99	10.89	14.86	17.01
D-c				1.58	9.82	14.18	12.65	19.19	22.41
N-c					8.94	11.04	10.05	15.12	17.80
E-n						3.80	2.85	8.09	10.65
B-n							1.12	1.31	2.59
F-n								3.07	4.82
D-n									0.35

of coding and non-coding tables are 0.50 and 0.34, respectively.

A parallel test has been conducted also on sample NLF008. The results are similar, but the convergent solutions using different initializations are even closer, and the coincidence rates of phases are higher. The correspondence of tables between samples NLF008 and DLF014 is evident. The distances related to the two samples are also listed in Table **1**. Thus, the two samples share similar 'codon usages'.

3.2. Two-category Model

A straightforward way to extend the single-category model to a multi-category model is to consider a mixture model, in which the probability of sequence W is

$$P(W) = \sum_{\gamma,\theta_\gamma} P(W \mid \theta_\gamma, \gamma) P(\theta_\gamma \mid \gamma) P(\gamma), \quad (4)$$

where γ indicates the composite category or component, and θ_γ the phase viewed in the category. We denote the maximal term in the summation:

$$Q(W) = P(W \mid \theta^*, \gamma^*), \quad (\theta^*, \gamma^*) = \arg\max_{\theta_\gamma, \gamma} P(\theta_\gamma, \gamma, W). \quad (5)$$

The log-likelihood $L_P = \sum_W \log P(W)$ serves as the objective function of the Expectation-Maximization algorithm for the model parameter training. Similarly, $L_Q = \sum_W \log Q(W)$ serves as the objective function for the greedy algorithm. (The monotonicity is only guaranteed for L_P.) Equation (2) then becomes

$$P(\theta_\gamma, \gamma \mid W) \propto P(W \mid \theta_\gamma, \gamma) P(\theta_\gamma \mid \gamma) P(\gamma), \quad P(\gamma \mid W) \propto \sum_{\theta_\gamma} P(\theta_\gamma, \gamma \mid W). \quad (6)$$

We have seen in the above that the difference between any two noncoding tables is generally much smaller than that between a coding and a noncoding table, so an alternative model is to consider six phases (other than the 0) instead of seven and regard all noncoding reads (phase 0) as a new independent category ($\gamma = 0$). With the only modification that no phase assignment is for $\gamma = 0$ (or $P(W, \theta_0, \gamma = 0) = P(W, \gamma = 0)$) and the summation over phases is restricted to phases other than 0, then Eq. (6) is still valid. (In fact, in the mixture model here the association of a noncoding table with its coding table is rather weak. For example, if two categories have similar $P(\gamma)$, then exchanging their noncoding table results in no significant change.) We shall call this reduced model the 1n2c model. Correspondingly, the original mixture model of two 'complete' categories is called the 2n2c model, and the previous single-category model the 1n1c model.

For the 2n2c model, a convenient initialization is to use tables extracted from two known genomes, say *E. coli* and *B. vulgatus*. As for the 1n2c model, the noncoding table may be taken to be that of either genome (or that combining both of them). To be least dependent on any knowledge (which could be a source introducing bias), another initialization uses random assignment of categories and phases. A more efficient way for doing this is to consider the following 'entropic clustering' [15]. We start with those reads identified as phase 1 in the above single-category model. Splitting the set of sequences into two subsets or clusters we can calculate two triplet tables, and then update the cluster assignment. This procedure of iteration is similar to the self-organizing approach, but even simpler.

We first investigate the 2n2c model on sample DLF014 using the tables of *E. coli* and *B. vulgatus*. The iteration converges after four steps using the criterion that for every table the distance between two successive steps falls within 0.05. The convergent tables are denoted as EB-1c, EB-1n for category 1, and EB-2c, EB-2n for category 2, with c being for coding and n for noncoding. We then conduct a similar scrutiny using the tables of *F. ruminococcus* and *B. vulgatus*. The convergent tables are marked with FB. The tables of EB and FB are very close, and the comparison is given in Table **2**.

Table 2. Distribution Distances D between the Convergent tables of Three Models: the 2n2c Model EB Using *E. coli* and *B. Vulgatus*, the 2n2c Model FB Using *F. Ruminococcus* and *B. Vulgatus*, and the 1n2c model EB' Using *E. coli* and *B. vulgatus*. Coding and Noncoding Tables are Indicated by Letters c and n, Respectively

	EB-2c	FB-1c	FB-2c	EB'-1c	EB'-2c	EB-1n	EB-2n	FB-1n	FB-2n	EB'-n
EB-1c	19.01	0.038	17.76	0.08	17.13	16.22	38.22	15.89	38.54	38.17
EB-2c		20.44	0.03	19.20	0.15	14.03	13.39	14.46	13.52	13.33
FB-1c			19.16	0.12	18.51	17.15	39.65	16.80	39.97	39.59
FB-2c				17.94	0.07	13.41	13.88	13.81	14.02	13.83
EB'-1c					17.21	15.03	37.80	14.69	38.13	37.75
EB'-2c						11.92	13.71	12.30	13.86	13.65
EB-1n							18.77	0.02	19.12	18.73
EB-2n								19.64	0.01	0.00
FB-1n									20.00	19.60
FB-2n										0.00

We further combine the two noncoding tables EB-1n and EB-2n (by taking the average with their fractions being weights) as one table, which is then used together with EB-1c and EB-2c to train the 1n2c model. The convergent tables of the 1n2c model are denoted by EB'-1c, EB'-2c, and EB'-n. The comparison between the 1n2c and 2n2c models is also given in Table **2**. The coincidence rate in the coding phase assignment between the 1n2c and 2n2c models is 95.9%. In fact, the coding phase assignment is rather robust. Even between the 1n1c and 1n2c models the coincidence rate in the coding phase assignment is still as high as 88.7%. The coding fractions belonging to EB'-1c and EB'-2c are 0.39 and 0.51. The GC content of non-coding EB'-n, and coding EB'-1c, EB'-2c are 0.32, and 0.59, 0.44, respectively.

To avoid any risk of bias, starting with the convergent tables of the 1n1c model, we continue training for the 1n2c model. The single noncoding table is ready for use. Since the coding phase assignment is steady, we use the entropic clustering approach to simply split annotated coding reads into two clusters without changing their phases assigned in the 1n1c model, and then obtain a coding table from each cluster. The purpose of the entropic clustering here is only to obtain two initial coding tables. We then may use either the whole set of coding reads, or more efficiently use only a subset, say reads of a given phase. The convergent tables are indicated with R for the clustering of the whole set and R' for that of a subset. The comparison between their tables and the above EB' tables is given in Table **3**.

A parallel scrutiny has been conducted also on sample NLF008. However, it is much more efficient to use a set of the tables for the 1n2c model trained on sample DLF014 as input for building up the 1c2n model on sample NLF008. There is no significant difference between the solutions of the two approaches. The correspondence between the tables of NLF008 and DLF014 is easily recognizable. Thus, as in the coarser 1n1c model, the two samples still share similar 'codon usages' under the 1n2c model. The fractions of

Table 3. Distribution Distances D between the Convergent Tables of Three 1n2c Models: the Above EB' Model, and Two Models Derived from the 1n1c Model Using the Whole set (R) and a Subset (R') of Annotated Coding Reads. Coding and Noncoding Tables are Indicated with Letters c and n, Respectively

	EB'-2c	R-1c	R-2c	R'-1c	R'-2c	EB'-n	R-n	R'-n
EB'-1c	17.21	0.01	16.60	0.01	16.65	37.75	38.27	38.39
EB'-2c		17.89	0.01	17.84	0.01	13.65	13.92	13.99
R-1c			17.27	0.00	17.33	38.57	39.09	39.21
R-2c				17.23	0.00	13.81	14.09	14.17
R'-1c					17.29	38.53	39.05	39.17
R'-2c						13.76	14.04	14.12
EB'-n							0.01	0.01
R-n								0.00

coding reads annotated by the two coding tables are 0.37, 0.52 for DLF014, and 0.33, 0.56 for NLF008.

Combining samples DLF014 and NLF008, and using the 1n2c tables of sample DLF014 as input, we have trained a common set of tables for annotating both samples. The comparison between tables of three 1n2c models trained on sample DLF014 (D), on sample NLF008 (N), and on their combination (DN) is given in Table **4**, where in the last column the fractions are identified by the corresponding frequency tables.

3.3. Three-category 1n3c Model

We shall focus only on the 1n3c model. A primitive survey shows that the tables of *E. coli*, *B. vulgatus* and *F. ruminococcus* are not appropriate for the training of the 1n3c model. We start with the tables obtained by training the 1n1c model, and split the single coding table 1c into three tables by the entropic clustering approach. In the three runs of random initializations (marked as R_a, R_b, and R_c, respectively) we use all the reads annotated as coding, those annotated as phase 1, and those of phase 4 of the model 1n1c, then carry out the entropic clustering to establish three coding tables in each run, and examine the sensitivity of solutions to the initial conditions. The comparison between them is given in Table **5**. It is seen that the three solutions are very close (although R_b is a little less close to R_a and R_c in their third coding table, R_a and R_c are almost identical).

Table 4. Distribution Distances D between the Convergent Tables of Three 1n2c Models Trained on Sample DLF014 (D), on Sample NLF008 (N), and on their Combination (DN). The Last Column (f) Indicates the Corresponding Fraction of Reads. Coding and Noncoding Tables are Indicated by Letters c and n, Respectively

	D-2c	N-1c	N-2c	DN-1c	DN-2c	D-n	N-n	DN-n	f
D-1c	17.27	2.43	20.55	0.65	18.83	39.09	40.48	40.18	0.37
D-2c		10.96	1.45	13.73	0.39	14.09	15.09	14.83	0.52
N-1c			13.69	0.60	12.13	30.73	32.14	31.80	0.33
N-2c				16.81	0.34	12.90	13.74	13.53	0.56
DN-1c					15.14	34.73	36.14	35.82	0.35
DN-2c						13.06	13.98	13.75	0.54
D-n							0.13	0.04	0.11
N-n								0.03	0.11

Table 5. Distribution Distances D between the Convergent Tables of the 1n3c Model Trained on Sample DLF014 Using Three Random Initializations (Marked as R_a, R_b, and R_c). Coding and Noncoding Tables are Indicated by Letters c and n, Respectively

	R_a-2c	**R_a-3c**	**R_b-1c**	**R_b-2c**	**R_b-3c**	**R_c-1c**	**R_c-2c**	**R_c-3c**	**R_a-n**	**R_b-n**	**R_c-n**
R_a-1c	23.04	10.33	0.03	21.54	13.94	0.00	23.34	10.92	13.60	13.74	13.67
R_a-2c		10.20	21.97	0.06	14.77	23.09	0.01	9.91	41.68	42.04	41.82
R_a-3c			9.57	8.97	1.35	10.31	10.47	0.13	20.97	21.33	21.10
R_b-1c				20.49	13.21	0.03	22.28	10.16	13.79	13.95	13.87
R_b-2c					13.49	21.59	0.09	8.70	40.00	40.36	40.14
R_b-3c						13.89	15.05	1.58	19.83	20.20	19.96
R_c-1c							23.39	10.92	13.54	13.69	13.62
R_c-2c								10.20	42.06	42.42	42.20
R_c-3c									21.42	21.78	21.54
R_a-n										0.00	0.00
R_b-n											0.00

3.4. Annotating Samples with Trained Triplet Tables

For the 1n2c model we have obtained three sets of tables, one by using sample DLF014 (D), another using NLF008 (N), and the third (DN) by combining D and N. We scan the 77 samples for the fractions of the two coding types with these three sets of tables. The cumulated number of samples with their fractions belonging to 1c not greater than a given value f is shown in Fig. (**3**). It is seen that the diabetic samples tend to have a high fraction of 1c (or a low fraction of 2c). Since the three sets of triplet frequency tables are quite close to each other, the differences in the annotated fractions are not significant. In fact, the three fraction annotations are highly correlated (with a correlation coefficient $r \sim 99\%$).

We have also scanned the 77 samples for the fractions of the three coding types with the tables of the 1n3c model trained on sample DLF014. The result is shown in Fig. (**4**), where a sample is represented by a point inside an equilateral triangle, and the distance from the point to the side 1 (2, or 3) gives its fraction of coding type 1c (2c, or 3c). It is seen that the diabetic samples tend to have a low fraction of 1c.

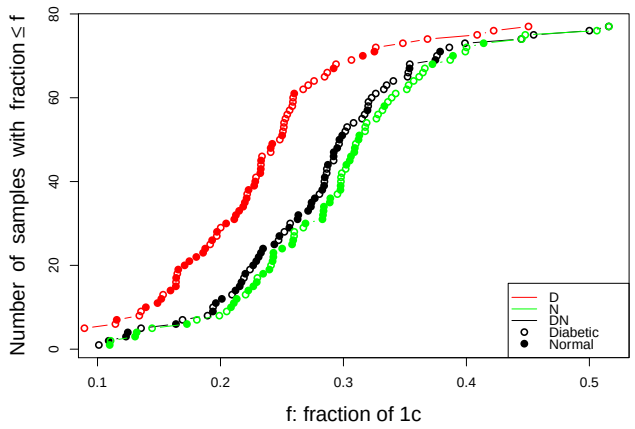

Fig. (3). The cumulated number of samples with their 1c fractions $\leq f$ in the annotations with the three sets of the 1n2c frequency tables obtained by training on sample DLF014 (D), training on NLF008 (N), and combining D and N (DN).

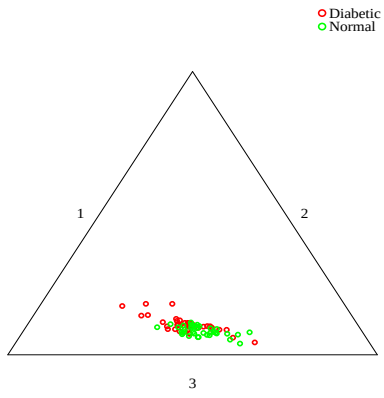

Fig. (4). Annotation of the 77 samples with the frequency tables of the 1n3c model trained on sample DLF014. A sample is represented by a point inside the equilateral triangle, and the distance from the point to the side marked 1 (2, or 3) gives its fraction of coding type 1c (2c, or 3c), respectively.

By noticing that coding table 1c of the 1n3c model is rather close to table 2c of model 1n2c ($D = 0.42$), this is consistent with the annotation by model 1n2c. The 77 samples carry a label of 'lean' or 'obese'. The relation between this label and the annotation is shown in Fig. (**5**).

Two species of *Roseburia* (*R. intestinalis* and *R. inulinivorans*) were found to be enriched in nondiabetic individuals. They are close relatives of *Firmicutes*, and belong to the traditionally studied butyrate-producing bacteria [6]. Using the coding table of *Roseburia intestinalis* genome, we scan the 77 samples. With the 25% quantile value of the simulated likelihood distribution taken as the threshold, the cumulated number of samples as a function of the annotated fraction is shown in Fig. (**6**).

4. CONCLUDING REMARKS

Most of the prevalent human intestinal microbial genes have been identified by assembling from the Illumina-based gut metagenome sequencing data. However, the sequence assembly is computationally costly. Furthermore, the human gut metagenome is an uneven mixture of organisms. This

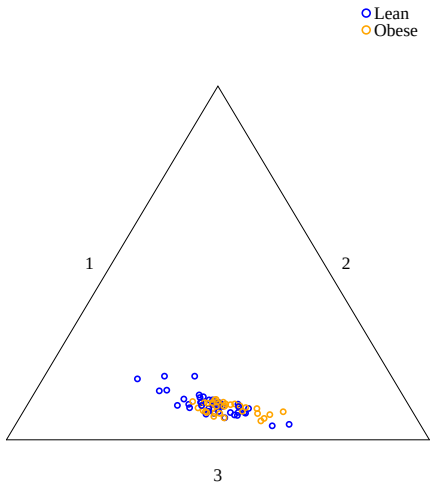

Fig. (5). Annotation of the 77 samples with the frequency tables of the 1n3c model trained on sample DLF014. In contrast to Fig. 4, here the labeling of samples is with respect to 'Lean' and 'Obese'.

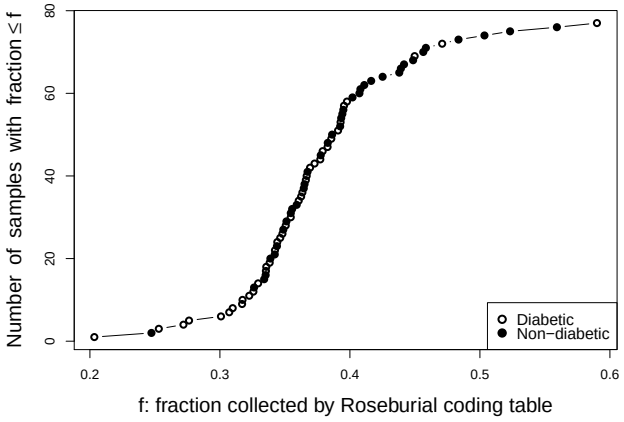

Fig. (6). Cumulated number of samples as a function of the fraction collected by the coding frequency table of the *Roseburia intestinalis* genome.

makes many statistical approaches useless when dealing with the sequence data of such mixed populations. In many cases, knowing some general characteristics of the metagenome could be more important than knowing individual genes therein. Our self-organizing approach for the human gut metagenome without requiring assembling meets such needs.

In the simplest model (1n1c) the whole metagenome is viewed as if it were a genome of a single organism. Even this unrealistic model is still useful since it is relatively easy to infer the coding phase of reads, while the phase annotation of the 1n1c model can be useful. At least we may use the phase information to pick out the reads of a single phase for sequence assembling. In some ideal cases, the phase information can even be used to infer the order of several 'unitigs'.

The approaches to use the 'codon usages' is a supplement to other methods. It might be useful in detecting changes in the metagenome composition [16]. When we are interested in a specific species in a metagenome, we may scan the sequencing reads of the metagenome with its codon usages for concentrating the relevant reads and even for extracting some qualitative information about the species.

According to the self-organizing approach, the fractions of the six coding phases should be fairly even. Since the direction of the reads is unknown, there should be some symmetry in the noncoding triplet frequencies, i.e. the frequencies for the two triplets in the Crick-Watson pair should be very close. A key issue for the self-organizing approach is to avoid the sensitivity of solutions on the initial conditions. The simpler the model, the less the risk to be trapped in an unwanted solution. Our reduced models have been carefully designed, and their sensitivity to the initial conditions checked.

Our initial investigations on the human gut metagenome show that normal and diabetic individuals share, to a large extent, common general characteristics in their gut metagenomes. However, we do see some diversity in a few samples of obese individuals. For example, in the annotation of samples with coding tables of the 1n2c model, diabetic samples tend to have a high fraction of 1c. Our observations agree with the results obtained based on sequence assembling [17]. So far, the results obtained are very preliminary; further study is in progress.

This work is supported by the National Natural Science Foundation of China.

ACKNOWLEDGEMENTS

None declared.

CONFLICT OF INTERESTS

The author(s) confirm that this article content has no conflicts of interest.

REFERENCES

[1] F. Guarner and J.-R. Malagelada, "Gut flora in health and disease", *Lancet*, vol. 361, pp. 512-519, 2003.

[2] P. B. Eckburg, E. M. Bik, C. N. Bernstein, E. Purdom, L. Dethlefsen, M. Sargent, S. R., Gill, K. E. Nelson, and D. A. Relman, "Diversity of the human intestinal microbial flora", *Science*, vol. 308, pp. 1635-1638, 2005.

[3] S. Fanaro, R. Chierici, P. Guerrini, and V. Vigi, "Intestinal microflora in early infancy: composition and development", *Acta Paediatrica Supplementum*, vol. 91, pp. 48-55, 2003.

[4] R. E. Ley, P. J. Turnbaugh, S. Klein, and J. I. Gordon, "Microbial ecology: human gut microbes associated with obesity", *Nature*, vol. 444, pp. 1022-3, 2006.

[5] P. J. Turnbaugh, R. E. Ley, M. A. Mahowald, V. Magrini, E. R. Mardis, and J. I. Gordon, "An obesity-associated gut microbiome with increased capacity for energy harvest", *Nature*, vol. 444, pp. 1027-31, 2006.

[6] J. Qin, R .Li, J. Raes, M. Arumugam, K.S. Burgdorf, C. Manichanh, T. Nielsen, N. Pons, F. Levenez, T .Yamada, D.R. Mende, J. Li, J. Xu, S. Li, D. Li, J. Cao, B. Wang, H. Liang, H. Zheng, Y. Xie, J. Tap, P. Lepage, M. Bertalan, J.M.. Batto, T. Hansen, D. Le Paslier, A. Linneberg, H.B. Nielsen, E. Pelletier, P. Renault, T. Sicheritz-Ponten, K. Turner, H. Zhu, C. Yu, S. Li, M. Jian, Y. Zhou, Y. Li, X. Zhang, S. Li, N. Qin, H. Yang, J. Wang, S. Brunak, J. Doré, F. Guarner, K. Kristiansen, O. Pedersen, J. Parkhill, J. Weissenbach, C. MetaHIT, P. Bork, S.D. Ehrlich and J. Wang, "A human gut microbial gene catalogue established by metagenomic sequencing", *Nature*, vol. 464, no. 7285, pp. 59-65, 2010.

[7] J.Tap, S. Mondot, F. Levenez, E. Pelletier, C. Caron, J.P. Furet, E. Ugarte, R. Muñoz-Tamayo, D.L. Paslier, R. Nalin, J. Dore and M. Leclerc, "Towards the human intestinal microbiota phylogenetic core", *Environmental Microbiology*, vol. 11, pp. 2574-2584, 2009.

[8] C. Burge and S. Karlin, "Prediction of complete gene structures in human genomic DNA", *Journal of Molecular Biology*, vol. 268, pp. 78-94, 1997.

[9] C. B. Burge and S. Karlin, "Finding the genes in genomic DNA", *Current Opinion in Structural Biology*, vol. 8, pp. 346-354, 1998.

[10] R. Staden and A. D. McLachlan, "Codon preference and its use in identifying protein coding regions in long DNA sequences", *Nucleic Acids Research*, vol. 10, pp. 141-156, 1982.

[11] S. Audic and J.-M. Claverie, "Self-identification of protein-coding regions in microbial genomes", *Proceeding of the National Academy Sciences of the United States of America*, vol. 95, pp. 10026-10031, 1998.

[12] W.-M. Zheng and F. Wu, "In-phase implies large likelihood for independent codon model: distinguishing coding from non-coding sequences", *Journal of Theoretical Biology*, vol. 223, pp. 199-203, 2003.

[13] C. Zimmer, "Bacteria divide people into 3 types, scientists say," *The New York Times*. 2011. Available: http://www.nytimes.com/2011/04/21/science/21gut.html.

[14] M. Arumugam, J. Raes, E. Pelletier, D. Le Paslier, T. Yamada, D.R. Mende, G.R. Fernandes, J. Tap, T. Bruls, J.M. Batto, M. Bertalan, N. Borruel, F. Casellas, L. Fernandez , L. Gautier, T. Hansen, M. Hattori, T. Hayashi, M. Kleerebezem, K. Kurokawa, M. Leclerc, F. Levenez, C. Manichanh, H.B. Nielsen, T. Nielsen, N. Pons, J. Poulain, J. Qin, T. Sicheritz-Ponten, S. Tims, D. Torrents, E. Ugarte, E.G. Zoetendal, J. Wang, F. Guarner, O. Pedersen, W.M. de Vos, S. Brunak, J. Doré, C. MetaHIT, M. Antolín, F. Artiguenave, H.M. Blottiere, M. Almeida, C. Brechot, C. Cara, C. Chervaux, A. Cultrone, C. Delorme, G. Denariaz, R. Dervyn, K.U. Foerstner, C. Friss, M. Van de Guchte, E. Guedon, F. Haimet, W. Huber, J. Van Hylckama-Vlieg, A. Jamet, C. Juste, Kaci, J. Knol, O. Lakhdari, S. Layec, K. Le Roux, E. Maguin, A. Mérieux, R. Melo Minardi, C.G. M'rini, Muller, R. Oozeer, J. Parkhill, P. Renault, M. Rescigno, N. Sanchez, S. Sunagawa, A. Torrejon, J. K. Turner, G. Vandemeulebrouck, E. Varela, Y. Winogradsky, G. Zeller, J. Weissenbach, S.D. Ehrlich and P. Bork, "Enterotypes of the human gut microbiome", *Nature*, vol. 4, pp. 550-553, 2011.

[15] W.-M. Zheng, "Entropic approach for reduction of amino acid alphabets", *Preprint*, 2001. [Online] Available: http://arxiv.org/abs/physics/0106074.

[16] G.D. Wu, J. Chen, C. Hoffmann, K. Bittinger, Y.Y. Chen, S.A. Keilbaugh, M. Bewtra, D. Knights, W.A. Walters, R. Knight, R. Sinha, E. Gilroy, K. Gupta, R. Baldassano, L. Nessel, H. Li, F.D. Bushman and J.D. Lewis, "Linking long-term dietary patterns with gut microbial enterotypes", *Science*, vol. 105, pp. 105-8, 2011.

[17] T. Wang, "A metagenome-wide association study of gut microbiota identifies markers associated with type 2 diabetes", In: *International Human Microbiome Congress*, March 19-21, Paris, France, 2012.

Received: April 05, 2012 Revised: May 19, 2012 Accepted: May 21, 2012