Eu-Detect: An algorithm for detecting eukaryotic sequences in metagenomic data sets

MONZOORUL HAQUE MOHAMMED, SUDHA CHADARAM, DINAKAR KOMANDURI, TARINI SHANKAR GHOSH and SHARMILA S MANDE*

Bio-Sciences R&D Division, TCS Innovation Labs, Tata Consultancy Services Limited, Hyderabad 500 081, India

**Corresponding author (Fax, +91-40-66672222; Email, sharmila@atc.tcs.com)*

Physical partitioning techniques are routinely employed (during sample preparation stage) for segregating the prokaryotic and eukaryotic fractions of metagenomic samples. In spite of these efforts, several metagenomic studies focusing on bacterial and archaeal populations have reported the presence of contaminating eukaryotic sequences in metagenomic data sets. Contaminating sequences originate not only from genomes of micro-eukaryotic species but also from genomes of (higher) eukaryotic host cells. The latter scenario usually occurs in the case of host-associated metagenomes. Identification and removal of contaminating sequences is important, since these sequences not only impact estimates of microbial diversity but also affect the accuracy of several downstream analyses. Currently, the computational techniques used for identifying contaminating eukaryotic sequences, being alignment based, are slow, inefficient, and require huge computing resources. In this article, we present Eu-Detect, an alignment-free algorithm that can rapidly identify eukaryotic sequences contaminating metagenomic data sets. Validation results indicate that on a desktop with modest hardware specifications, the Eu-Detect algorithm is able to rapidly segregate DNA sequence fragments of prokaryotic and eukaryotic origin, with high sensitivity. A Web server for the Eu-Detect algorithm is available at *http://metagenomics.atc.tcs.com/Eu-Detect/*.

[Mohammed MH, Chadaram S, Komanduri D, Ghosh TS and Mande SS 2011 Eu-Detect: An algorithm for detecting eukaryotic sequences in metagenomic data sets. *J. Biosci.* **36** 709–717] **DOI** 10.1007/s12038-011-9105-2

1. Introduction

Metagenomic samples are composed of microbes belonging to either the prokaryotic or the eukaryotic sub-kingdoms. Researchers typically use physical partitioning methods (e.g. filtration, differential centrifugation, differential lysis, density gradients, pulsed-field electrophoresis, etc.) for first segregating the prokaryotic and eukaryotic cellular fractions of a metagenomic sample. Subsequently, the entire genomic content of the metagenomic sample is extracted, sequenced and analysed. However, given that the physical dimensions of cells belonging to the micro-eukaryotic community is roughly similar to those of prokaryotes, the process of obtaining an exclusive prokaryotic cellular fraction using physical partitioning approaches is untenable. Consequently, metagenomic data sets (sequenced with the objective of studying the prokaryotic fraction) are frequently observed to be contaminated with DNA sequences originating from the genomes of eukaryotic cells (Lopez-Garcia *et al.* 2001; Moon-van der Staay *et al.* 2001; Piganeau *et al.* 2008; Scanlan and Marchesi 2008; Willner *et al.* 2009). It is to be noted that eukaryotic sequences in metagenomic data sets may originate not only from the genomes of micro-eukaryotic species but also from accidental contamination from the genome of eukaryotic host cells. The latter scenario generally occurs in the case of host-associated metagenomes, e.g. microbes living in human gut. In such cases, one needs to identify and remove contaminating host DNA sequences, since the latter sequences will not only lead to incorrect estimates of microbial diversity but will also affect the accuracy of several other downstream analyses.

Given the limitations of physical partitioning methods, researchers frequently use *in silico* approaches for identifying

Keywords. Alignment-free; feature vector space; metagenomics; micro-eukaryotes; oligonucleotide composition

Supplementary materials pertaining to this article are available on the *Journal of Biosciences* Website at *http://www.ias.ac.in/jbiosci/Sep2011/pp709–717/suppl.pdf*

and segregating prokaryotic and eukaryotic sequences in metagenomic data sets. Existing *in silico* approaches are either composition based or similarity based. Composition-based approaches (Diaz *et al.* 2009) classify sequences as prokaryotic/eukaryotic by comparing their 'compositional' characteristics (e.g oligonucleotide frequency patterns) with those of known prokaryotic/eukaryotic genomes. Being alignment-free, composition-based approaches are capable of analysing thousands of sequences within reasonable limits of time. However, given that metagenomic sequences obtained using current generation sequencing technologies have short lengths, the compositional signal derived from these sequences is generally insufficient for obtaining reliable/accurate taxonomic assignments. On the other hand, similarity-based approaches utilize BLAST searches (Altschul *et al.* 1990) of each sequence (in the metagenomic data set under study) against a database containing eukaryotic sequences. Sequences in the metagenomic data set showing significant similarity (at least 80% identity over 80% of query length) to known eukaryotic sequences (or to host DNA sequences in case of host-associated metagenomes) are identified and separated. Based on project objectives, the identified sequences are either analysed separately or are completely discarded.

Although sequence-similarity-based approaches are capable of identifying eukaryotic DNA sequences with high sensitivity, practical utility of such methods for metagenome settings is limited for the following reasons. Firstly, since genome sequences of most eukaryotes are currently unavailable, the fraction of eukaryotic sequences in the metagenome data set originating from genomes of hitherto unknown eukaryotes will remain unidentified. Secondly, since metagenomic data sets typically contain millions of sequences (e.g. 7.5 million sequences in the Global Ocean Sampling Expedition Microbial Metagenomic data sets of Venter *et al.* 2004; Rusch *et al.* 2007; Yooseph *et al.* 2007), a BLAST search of all these sequences against a database containing sequences of a host genome (e.g. human genome) will take enormous amount of time and computing resources.

In this article, we present a novel alignment-free algorithm, called Eu-Detect, that can detect eukaryotic sequences in metagenomic data sets. The principle behind this algorithm is as follows. The oligonucleotide composition of prokaryotic genomes is distinct as compared to oligonucleotide composition of DNA sequences originating from eukaryotic cells. Consequently, if a mixture of sequences originating from both these groups are clustered based on oligonucleotide usage patterns, DNA sequences of prokaryotic origin are expected to cluster together and spatially well separated (in feature vector space) from the clusters containing DNA sequences of eukaryotic origin. This spatial separation is utilized by Eu-Detect to classify a given query sequence to either groups. Query sequences that associate themselves (i.e. have the least distance) with clusters containing a majority of prokaryotic or eukaryotic sequences are classified by Eu-Detect as prokaryotic or eukaryotic sequences, respectively. Given that the runtime steps of the Eu-Detect algorithm only involves finding the closest clusters (in terms of oligonucleotide composition) for a given query sequence, and subsequently checking the numerical ratio of prokaryotic to eukaryotic sequences in these clusters, the Eu-Detect algorithm is quicker than existing alignment-based methods by several orders of magnitude. In addition, unlike alignment-based algorithms, Eu-Detect is able to identify the fraction of eukaryotic sequences originating not only from known eukaryotic genomes but also from the genomes of hitherto unknown eukaryotes.

2. Methods

2.1 *Clustering prokaryotic and eukaryotic DNA sequences*

As a one-time pre-processing step, DNA fragments (generated by splitting known eukaryotic and prokaryotic genomes) were first clustered based on oligonucleotide usage patterns. This resulted in the following three types of clusters

a) Clusters composed of only/majority of eukaryotic sequences
b) Clusters composed of only/majority of prokaryotic sequences
c) Clusters composed of a comparable number of prokaryotic and eukaryotic sequences.

Besides differing in their composition, these clusters were observed to occupy spatially distinct regions in feature vector space, indicating that the oligonucleotide composition of prokaryotic genomes is distinct as compared with that of eukaryotic genomes. Capturing these patterns of spatial localization forms the rationale/objective of the one-time pre-processing/clustering step. During actual runtime, these spatial localization pattens are utilized by Eu-Detect as 'reference patterns' for classifying a given query sequence as either a eukaryotic or a prokaryotic sequence. For this purpose, sequences of 237 prokaryotic and 27 eukaryotic genomes were downloaded from NCBI database (*ftp://ftp.ncbi.nih.gov/genomes/*). The 27 eukaryotic genomes (listed in supplementary table 1) represented micro-eukaryotes, fungi, plants and a higher eukaryote (human). All the downloaded genome sequences were split into non-overlapping fragments, each having a length of 1000 bp. Frequencies of all possible tetranucleotides in each fragment were computed and stored in the form of a 256 dimensional vector. Using k-means clustering approach (Hartigan and Wong 1979), all vectors

corresponding to prokaryotic and eukaryotic genome fragments were clustered. The number of cluster centers (k) used for initiating the clustering process was calculated using the following formula (Mardia *et al.* 1979).

$$k \approx \sqrt{n/2}$$

where n is the number of objects (data points).

This formula is generally used as a thumb rule for determining the initial number of clusters. The Manhattan distance (L1 norm) between individual vectors was used as the similarity measure for clustering. A total of 1983 clusters were formed. Besides obtaining the numerical count of prokaryotic and eukaryotic sequences in each cluster, vectors corresponding to the cluster centroids for each of the individual clusters were computed and stored.

2.2 *Steps for classifying a given sequence as prokaryotic or eukaryotic*

The following two steps are followed by Eu-Detect for classifying input query sequences as prokaryotic or eukaryotic.

Step 1. Identifying clusters having least distance to the query sequence: A vector representing the frequencies of all 256 tetra-nucleotides in each input query sequence is generated. The Manhattan distance (L1 Norm) of this query vector to the vectors corresponding to each of the precomputed 'cluster centroids' is calculated. A set of clusters having the least distance to the query vector, and having a 'cumulative sequence count' (a predetermined value), are identified. The methodology to obtain the predetermined value of the cumulative sequence count is described in section 2.3.

Step 2. Classifying a given sequence as prokaryotic or eukaryotic: If the percentage of eukaryotic sequences in the set of closest clusters exceeds a predetermined 'coverage threshold', the query sequence is classified as a probable eukaryotic sequence. Sequences which do not satisfy this 'coverage threshold' criterion are tagged as probable prokaryotic sequences. The methodology used for identifying optimal value of 'coverage threshold' is given in the next section.

2.3 *Obtaining optimal threshold values*

The Eu-Detect algorithm requires two runtime values. The first value represents the 'optimal number' of closest clusters (in the pre-clustered database) that are to be considered (in the subsequent classification steps) for a given query sequence. In this study, the cumulative number of sequences within the set of closest clusters was used as a parameter (referred to as 'cumulative sequence count') for identifying an 'optimal number' of closest clusters. Increasing the value of 'cumulative sequence count' would result in more number of closest clusters to be considered for a given query sequence. The second runtime value represents the threshold proportion (referred to as 'coverage threshold') of eukaryotic sequences within the identified set of closest clusters. If the proportion of eukaryotic sequences (in the identified set of closest clusters) exceeds the threshold value, the query sequence is classified as an 'eukaryotic sequence'. Else, it is classified as an 'prokaryotic sequence'. The objective was to identify an optimal combination of these two run-time values, where Eu-Detect is able to achieve highest classification accuracy.

In order to obtain (the above mentioned) optimal threshold values for 'cumulative sequence count' and 'coverage threshold', five training data sets were generated using MetaSim software (Richter *et al.* 2008). These data sets contained sequences randomly generated from genomes of 140 prokaryotes (belonging to diverse taxonomic clades), 13 fungi, 3 micro-eukaryotes, 10 plants and a higher eukaryote (human). The eukaryotic genomes used for generating these training data sets are listed in supplementary table 1. The prokaryotic training data set consisted of 560000 sequences of varying lengths. Each of the four eukaryotic training data sets consisted of 140000 sequences. Based on the lengths of the sequences, these training sets were further divided into four data sets, termed as Sanger data set, 454-400 data set, 454-250 data set and 454-100 data set. Sequences constituting these four data sets simulated the typical sequence lengths and errors models associated with the four commonly used sequencing technologies, namely Sanger (read length centered around 800 bp), 454-GS-FLX-Titanium (400 bp), 454-GS-FLX-Standard (250 bp) and Roche-454-GS20 (100 bp), respectively (Sanger *et al.* 1977; Margulies *et al.* 2005). Using the steps described in section 2.2, sequences in each data set were classified by Eu-Detect as probable prokaryotic or eukaryotic sequences. Classification accuracy of Eu-Detect algorithm (with all training data sets) using various combinations of 'cumulative sequence count' (40K, 50K, 60K and 70K sequences) and 'coverage threshold' values (20%, 30%, 40%, 50%, 60% and 70%) was tabulated. Optimal threshold values for Eu-Detect were identified by analysing the tabulated results. For a data set containing prokaryotic and eukaryotic sequences, classification accuracy was defined as the percentage of sequences correctly classified by Eu-Detect as prokaryotic and eukaryotic, respectively.

2.4 *Validation of the Eu-Detect approach*

The performance of the Eu-Detect algorithm (in terms of classification accuracy and time taken for classifying query sequences as prokaryotic or eukaryotic) has been validated

using 20 test data sets. These data sets (having 35000 sequences each), corresponding to a higher eukaryote (mouse), micro-eukaryotes, plants, fungi and prokaryotes, were generated using an approach similar to that used for generating training data sets (explained in section 2.3). In order to test the classification accuracy of the Eu-Detect algorithm in scenarios where test sequences originate from new organisms, it was ensured that sequences in test data sets were derived from genomes that were not considered during the clustering step (section 2.1).

2.5 *Comparison with other methods*

In order to compare the performance of Eu-Detect in detecting eukaryotic and prokaryotic sequences originating from unknown organisms, results obtained with Eu-Detect (for all 20 test data sets) were compared with those obtained using a composition-based classification method, namely TACOA (Diaz *et al.* 2009). Results of TACOA were obtained by querying sequences in test data sets against a reference database built using the same 237 prokaryotic and 27 eukaryotic genomes thar were used by Eu-Detect for creating a clustered database (section 2.1)

2.6 *Testing of Eu-Detect algorithm on real metagenomic data sets*

The performance of the Eu-Detect algorithm was tested on five metagenomic data sets (comprised of 1405980 sequences) reported earlier by Willner *et al.* (2009). These data corresponded to the metagenomes obtained from the respiratory tract of humans. The objective of analysing these real metagenomes was to quantify the percentage of overlap between sequences identified as eukaryotic by the Eu-Detect algorithm and those identified by a similarity-based approach (MegaBLAST) and a composition-based approach (TACOA). For this purpose, sequences in these five metagenomic data sets were first compared against human genome sequences using MegaBLAST. The number of sequences (in each data set) having significant hits (minimum 80% identity over at least 80% of query length) to human genome sequences were identified. Results of TACOA were obtained by querying all the sequences in each data set against a modified database that contained the same set of genomes used by Eu-Detect during the validation process

3. Results

3.1 *Optimal threshold values for Eu-Detect*

The results of Eu-Detect using various combinations of 'cumulative sequence count' and 'coverage threshold' values are summarized in supplementary table 2 and supplementary figure 1. In the present study, a query sequence is classified as a 'eukaryotic sequence' only if the proportion of eukaryotic sequences in the identified set of closest clusters exceeds the 'coverage threshold' value. Else, it is classified as a prokaryotic sequence. Given that the cluster database consists of sequences thar are either prokaryotic or eukaryotic, a progressive increase in the value of 'coverage threshold' would logically result in fewer eukaryotic sequences being classified as eukaryotic. This is reflected as a progressive decrease in the classification accuracy for eukaryotic sequences with increasing coverage threshold value (supplementary table 2). On the other hand, a progressive increase in the value of 'coverage threshold' is observed to result in a greater number of prokaryotic sequences being classified as prokaryotic (supplementary table 2). These results indicate that the accuracy rates for classifying prokaryotic and eukaryotic sequences vary inversely with the increasing 'coverage threshold' value. Results in supplementary table 2 also indicate that a progressive increase in the value of cumulative sequence count (keeping the value of coverage threshold constant) does not significantly alter the classification accuracy for both prokaryotic and eukaryotic sequences. This indirectly indicates a clear spatial separation of prokaryotic and eukaryotic sequences in the cluster database.

In summary, results given in supplementary table 2 (wherein Eu-Detect's classification accuracy with various data sets was obtained using a clustered database generated using 1000 bp genome fragments) indicate that for data sets with sequence lengths greater than 250 bp, an average classification accuracy of 83–91% (for both prokaryotic and eukaryotic data sets) is obtained using a cumulative sequence count of 40000 and a coverage threshold of 60%. On the other hand, for data sets having short sequences (around 100 bp), an average classification accuracy (for both prokaryotic and eukaryotic data sets) of 80% is obtained, using cumulative sequence count and coverage threshold values of 40000 and 70%.

In order to check if the size of the fragments used for building the clustered database has any impact on the classification accuracy of Eu-Detect, two separate cluster databases were created using genome fragments of length 500 bp and 250 bp. The results of Eu-Detect using these two cluster databases are given in supplementary tables 3 and 4, respectively. A comparison of the average classification accuracy obtained with various clustered databases (generated using fragments of length 1000 bp, 500 bp and 250 bp, respectively) is summarized in supplementary table 5. Results in this table indicate that the classification accuracy of Eu-Detect obtained with clustered databases created using shorter fragments (500 bp and 250 bp) is significantly lower as compared with that obtained using the clustered

database built using fragments of length 1000 bp. These results thus indicate that the clustering pattern obtained using shorter fragments is not very robust. In other words, clustering using shorter fragments is unable to efficiently partition prokaryotic and eukaryotic sequences based on oligonucleotide usage patterns. Given these observations, the clustered database built using genome fragments of length 1000 bp was retained for testing Eu-Detect. The optimized values of 'cumulative sequence count' and 'coverage threshold' (where Eu-Detect achieves highest classification accuracy for both prokaryotic and eukaryotic training data sets) obtained with the clustered database built using genome fragments of length 1000 bp were subsequently retained as runtime parameters for both the test data sets as well as for the five real metagenomic data sets.

3.2 *Results with test data sets*

Results given in table 1 indicate that the average classification accuracy of Eu-Detect ranges between 80–90% as compared to only 45–53% by TACOA. It is to be noted that the sequences in the test data were derived from genomes that are distinct from those in the training data sets.

Results of Eu-Detect obtained with test data sets are comparable with those obtained with training data sets. This indicates the validity of the premise (i.e. spatial separation of prokaryotic and eukaryotic sequences in feature vector space) on which the Eu-Detect algorithm is based. High classification accuracy with test data sets also indicates the suitability of the Eu-Detect algorithm for analysing real metagenomic sequence data sets, wherein majority of sequences are known to originate from hitherto unknown organisms.

Results in table 2 indicate that the cumulative time taken by Eu-Detect for analysing 1120000 sequences in test data sets was approximately 84 min (a processing rate of approximately 212 sequences per second). In comparison, the cumulative time taken by TACOA was in excess of 151 h (more than 6 days). Although a similarity based search (e.g. BLAST) would have accurately classified all sequences in test data sets, an estimated 21 days would be required to complete the analysis.

3.3 *Results with real metagenomic data sets*

Table 3 shows the results obtained using Eu-Detect, MegaBLAST (Zhang *et al.* 2000), and TACOA with five real metagenomic data sets. Results of MegaBLAST indicated that 7–35% of sequences in each data set have significant hits with human genome sequences. Approximately 91–96% of the sequences identified by MegaBLAST (as sequences of human origin) are also predicted by Eu-Detect as eukaryotic sequences. In contrast, only 31–48% of these sequences are predicted by TACOA to be of eukaryotic origin. This indicates that the Eu-Detect algorithm can be used for identifying the fraction of eukaryotic sequences in real metagenomic data sets.

In addition to the human sequences identified by both MegaBLAST and Eu-Detect, 9–13% of sequences in the real metagenomic data sets are additionally predicted by Eu-Detect as sequences of eukaryotic origin (table 3). Similarly, about 2–5% of sequences are additionally predicted as eukaryotic sequences by TACOA. A MegaBLAST analysis of these additionally predicted eukaryotic sequences indicated that 99.8% of sequences failed to obtain any significant hits with known prokaryotic and eukaryotic genome sequences. Since these sequences are compositionally similar to eukaryotic genome sequences, it is likely that they may have originated from the genomes of hitherto unknown eukaryotes.

Furthermore, results in table 3 indicate that the Eu-Detect algorithm is able to identify the taxonomic affiliation (prokaryotic or eukaryotic) of all sequences in a given data set. In contrast, the algorithm TACOA is unable to identify the taxonomic affiliation of 18–26% percentage of sequences in each data set. Consequently, 52–69% of sequences identified by MegaBLAST (as originating from human genome sequences) remain unclassified by TACOA (table 3).

4. Discussion

Characterizing the prokaryotic fraction of microbial communities is the prime objective of most metagenomic projects. A number of techniques are thus routinely employed (during the sample preparation stage) for enriching and finally segregating the prokaryotic and eukaryotic fractions of the metagenomic sample. In spite of these efforts, several metagenomic studies have reported the presence of contaminating eukaryotic sequences in metagenomic data sets. For example, analysis of contigs obtained from hind-gut microbiota of a wood-feeding higher termite (Warnecke *et al.* 2007) indicated the presence of arthropod DNA. Similarly, an *in silico* analysis of metagenome data obtained from the respiratory tract (Willner *et al.* 2009) have indicated that the extent of eukaryotic (i.e. host sequence) contamination is as high as 34%. These observations indicate the limitations of existing physical partitioning methods for segregating the prokaryotic and eukaryotic microbial fractions of a metagenomics sample. Identification and removal of contaminating eukaryotic sequences are important, since contaminating sequences not only lead to incorrect estimates of microbial diversity, but also affect the accuracy and efficiency of several downstream analyses.

In this article, a novel alignment-free algorithm has been presented that can rapidly analyse metagenomic

Table 1. (**A**) Classification accuracy of Eu-Detect and TACOA with various test data sets[a] and (**B**) average classification accuracy of Eu-Detect and TACOA across various test data sets[a]

(A) Taxonomic group to which test sequences belong	Number of test sequences classified as eukaryotic sequences		Number of test sequences classified as prokaryotic sequences		Number of test sequences classified as sequences of unknown origin	
	Eu-Detect	TACOA	Eu-Detect	TACOA	Eu-Detect	TACOA
Sanger data set						
Fungi	77.94	18.58	22.06	43.22	0	38.2
Plants	90.59	12.2	9.41	52.96	0	34.84
Higher Eukaryotes	91.89	61.98	8.11	9	0	29.02
Micro-eukaryotes	92	26.44	8	49.94	0	23.62
Prokaryotes	7.3	0.26	92.7	77.36	0	22.38
454-400 data set						
Fungi	71.31	12.32	28.69	55.64	0	32.04
Plants	88.33	4.92	11.67	64.4	0	30.68
Higher Eukaryotes	92.83	37.84	7.17	19.82	0	42.34
Micro-eukaryotes	90.48	26.46	9.52	49.76	0	23.78
Prokaryotes	13.94	0.26	86.06	72.76	0	26.98
454-250 data set						
Fungi	64.45	15.1	35.55	51.22	0	33.68
Plants	86.17	2.76	13.83	70.18	0	27.06
Higher Eukaryotes	88.55	32.86	11.45	26	0	41.14
Micro-eukaryotes	87.44	29.68	12.56	44.42	0	25.9
Prokaryotes	17.54	1.68	82.46	73.34	0	24.98
454-100 data set						
Fungi	56.11	21.64	43.89	43.2	0	35.16
Plants	85.3	0.62	14.7	69.64	0	29.74
Higher Eukaryotes	80.79	29.98	19.21	19.3	0	50.72
Micro-eukaryotes	82.91	28.64	17.09	39.38	0	31.98
Prokaryotes	15.38	0.94	84.62	69.06	0	30

(B) Test data sets	Average classification accuracy with eukaryotic test data sets (A)		Classification accuracy with prokaryotic test data sets (B)		Average classification accuracy with all test data sets (A+B)/2	
	Eu-Detect	TACOA	Eu-Detect	TACOA	Eu-Detect	TACOA
Sanger	88.11	29.8	92.7	77.36	90.41	53.58
454-400	85.74	20.39	86.06	72.76	85.9	46.58
454-250	81.65	20.1	82.46	73.34	82.06	46.72
454-100	76.28	20.22	84.62	69.06	80.45	44.64
Average	**82.95**	**22.63**	**86.46**	**73.13**	**84.7**	**47.88**

[a] All sequences in test data sets were derived from genomes which were absent from the reference cluster database.

data sets (having millions of sequences) and segregate DNA sequence fragments of prokaryotic or eukaryotic origin with high accuracy. A distinct advantage of Eu-Detect lies in the fact that the only runtime requirement

Table 2. Time taken by Eu-Detect and TACOA for analysing various test data sets

Taxonomic group to which test sequences belong	No. of sequences in data set	Time taken (minutes)	
		Eu-Detect	TACOA
Sanger data set			
Fungi	35000	3	324
Plants	35000	3	316
Higher Eukaryote	35000	3	311
Micro-eukaryotes	35000	3	315
Prokaryotes	140000	13	1254
454-400			
Fungi	35000	3	294
Plants	35000	3	283
Higher Eukaryote	35000	3	289
Micro-eukaryotes	35000	3	293
Prokaryotes	140000	13	1209
454-250			
Fungi	35000	2	276
Plants	35000	2	264
Higher Eukaryote	35000	2	234
Micro-eukaryotes	35000	2	275
Prokaryotes	140000	9	1179
454-100			
Fungi	35000	2	234
Plants	35000	2	229
Higher Eukaryote	35000	2	231
Micro-eukaryotes	35000	2	240
Prokaryotes	140000	9	1008
Total	**1120000**	**84 min**	**9058 min**

of the Eu-Detect program is a small file (less than 2 Mb) containing information about the numerical count of prokaryotic and eukaryotic sequences in each cluster, and the vectors corresponding to the cluster centroids of each individual cluster. Moreover, the processing time taken by the present algorithm for identifying 'probable eukaryotic sequences' in five respiratory tract metagenomic data sets (1.4 million sequences) on a modest desktop (2 GB RAM, 2.33 GHz processor) was roughly 1 h. This indicates that the Eu-Detect algorithm can be used by any metagenomic research group having access to a simple desktop. It is interesting to note that performing an analysis of the same data sets (on the same desktop) using TACOA and MegaBLAST took approximately 185 h (>7 days) and 317 h (i.e. in excess

Table 3. Results obtained using MegaBLAST, Eu-Detect and TACOA on five metagenomics data sets from the respiratory tract of humans. (Willner *et al.* 2009)

Data set name	Total Number of Sequences	Number (percentage) of 'eukaryotic sequences' (identified using MegaBLAST)	Number (percentage) of MegaBLAST identified 'eukaryotic sequences' classified as 'Eukaryotic'		Number (Percentage) of sequences predicted as 'Eukaryotic' (in addition to eukaryotic sequences identified using MegaBLAST)		Number (Percentage) of sequences classified as 'sequences of unknown origin'	
			By Eu-Detect	By TACOA	Eu-Detect	TACOA	Eu-Detect	TACOA
Non CF1 Asthma	273139	48874 (17.9)	45334 (92.8)	21953 (48.4)	27122 (9.9)	10714 (3.9)	0	50813 (18.6)
Non CF2	260800	92072 (35.3)	83960 (91.2)	26735 (31.8)	24625 (9.4)	5437 (2.1)	0	67501 (25.9)
Non CF3	232096	15498 (6.7)	14562 (93.9)	6147 (42.2)	26549 (11.4)	7568 (3.3)	0	41701 (17.9)
Non CF4 Spouse	319889	99296 (31.0)	95569 (96.3)	36631 (36.9)	34718 (10.8)	9398 (2.9)	0	81010 (25.3)
Non CF5	320056	42504 (13.3)	39899 (93.7)	18187 (45.6)	41768 (13.1)	15160 (4.7)	0	67767 (21.2)
Total	1405980	298244 (20.8)	279324 (93.7)	109653 (36.8)	154782 (10.9)	48277 (3.4)	0	308792 (21.9)

of 13 days), respectively. Furthermore, it is observed that alignment-based algorithms, including the recent published Deconseq algorithm (Schmieder and Edwards 2011), need enormous computational power (in terms of memory requirements) for performing this analysis. For instance, approximately 4 GB of RAM is needed by the BWA-SW algorithm (used internally by the Deconseq algorithm) for performing alignment-based searches against the human genome. Moreover, the utility of alignment-based methods (including DeconSeq) is limited to detecting contaminating sequences originating only from known eukaryotic genomes. On the other hand, the Eu-Detect algorithm is able to rapidly identify contaminating sequences originating, not only from known eukaryotic genomes, but also from the genomes of hitherto unknown eukaryotes.

Using a pre-clustering step for obtaining spatial localization patterns of prokaryotic and eukaryotic sequences (in feature vector space) forms the underlying premise of the Eu-Detect algorithm. A k-mer size of 4 was used in the present study for generating vectors corresponding to prokaryotic and eukaryotic sequences. This choice of k-mer size was driven by the following observations. Several earlier studies (Pride *et al.* 2003; Teeling *et al.* 2004) had indicated that the taxonomic discrimination capability of tetra-nucleotides is higher than that obtained using lower k-mer values. Moreover, the size of sequences generated using existing sequencing technologies is generally between 100–1000 bp. For sequences in this length range, the frequencies of various oligonucleotides generated using a k-mer value greater than 4 are expected to be low and statistically insignificant. Choosing a higher k-mer is thus expected to reduce the overall sensitivity/specificity of the classification approach. In addition, a higher k-mer size would also results in an increase in the overall time taken for clustering and subsequent analysis due to the following reason. A vector generated using a 5-mer would have 1024 dimensions as opposed to just 256 dimensions using 4-mer frequencies. Therefore, the time required for comparing 1024 dimensional vectors is expected to be higher as compared with 256 dimensional vectors.

In addition to the choice of k-mer size, an important factor that determines the classification accuracy of Eu-Detect is the quality of the cluster database generated during the pre-processing step. Any clustering method (supervised or un-supervised) that can efficiently partition prokaryotic and eukaryotic genome fragments, in principle, can be used for creating the initial cluster database. The k-means method was chosen (in this study) as the clustering method, since it was observed to efficiently achieve this partitioning. Other reasons behind using the k-means clustering approach pertain to its simplicity, ease of implementation and low time complexity. On a simple desktop, the complete clustering process (including the time taken for vector generation) took less than 4 h.

5. Conclusion

Eu-Detect is a novel alignment-free algorithm that can rapidly identify eukaryotic sequences contaminating metagenomic data sets. Validation results indicate that, even on a desktop with modest hardware specifications, this algorithm is able to rapidly identify contaminating sequences originating, not only from known eukaryotic genomes, but also from the genomes of hitherto unknown eukaryotes with high sensitivity.

References

Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ 1990 Basic local alignment search tool. *J. Mol. Biol.* **215** 403–410

Diaz N, Krause L, Goesmann A, Niehaus K and Nattkemper T 2009 TACOA-Taxonomic classification of environmental genomic fragments using a kernelized nearest neighbor approach. *BMC Bioinfo* **10** 56

Hartigan JA and Wong MA 1979 A K-Means Clustering Algorithm. *App. Stat.* **28** 100–108

Lopez-Garcia P, Rodriguez-Valera F, Pedros-Alio C and Moreira D 2001 Unexpected diversity of small eukaryotes in deep-sea Antarctic plankton. *Nature (London)* **409** 603–607

Mardia KV, Kent JT and Bibby JM 1979 *Multivariate analysis* (Academic Press)

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, *et al.* 2005 Genome sequencing in micro-fabricated high-density pico-litre reactors. *Nature (London)* **437** 376–380

Moon-Van Der Staay SY, Wachter RD and Vaulot D 2001 Oceanic 18S rDNA sequences from picoplankton reveal unsuspected eukaryotic diversity. *Nature (London)* **409** 607–610

Piganeau G, Desdevises Y, Derelle E and Moreau H 2008 Picoeukaryotic sequences in the Sargasso Sea metagenome. *Genome Biol.* **9** R5

Pride DT, Meinersmann RJ, Wassenaar TM and Blaser MJ 2003 Evolutionary implications of microbial genome tetranucleotide frequency biases. *Genome Res.* **13** 145–158

Richter DC, Ott F, Auch AF, Schmid R and Huson DH 2008 MetaSim – A sequencing simulator for genomics and metagenomics. *PLoS One* **3** e3373

Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S, Wu D, Eisen JA, *et al.* 2007 The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biol.* **5** e77

Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes CA, Hutchison CA, Slocombe PM, *et al.* 1977 The nucleotide sequence of bacteriophage phi X174 DNA. *Nature (London)* **265** 687–695

Scanlan PD and Marchesi JR 2008 Micro-eukaryotic diversity of the human distal gut microbiota: qualitative assessment using culture-dependent and independent analysis of faeces. *ISME J.* **2** 1183–1193

Schmieder R and Edwards R 2011 Fast identification and removal of sequence contamination from genomic and metagenomic data sets. *PLoS One,* **6** e17288

Teeling H, Meyerdierks A, Bauer M, Amann R and Glockner FO 2004 Application of tetranucleotide frequencies for the assignment of genomic fragments. *Environ. Microbiol.* **6** 938–947

Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, *et al.* 2004 Environmental genome shotgun sequencing of the Sargasso sea. *Science* **304** 66–74

Warnecke F, Luginbühl P, Ivanova N, Ghassemian M, Richardson TH, Stege JT, Cayouette M, McHardy AC, *et al.* 2007 Metagenomic and functional analysis of hindgut micro-biota of a wood-feeding higher termite. *Nature(London)* **450** 560–565

Willner D, Furlan M, Haynes M, Schmieder R, Angly FE, Silva J, Tammadoni S, Nosrat B, *et al.* 2009 Metagenomic analysis of respiratory tract DNA viral communities in cystic fibrosis and non-cystic fibrosis individuals. *PLoS One* **4** e7370

Yooseph S, Sutton G, Rusch DB, Halpern AL, Williamson SJ, Remington K, Eisen JA, Heidelberg KB, *et al.* 2007 The Sorcerer II global ocean sampling expedition: expanding the universe of protein families. *PLoS Biol.* **5** e16

Zhang Z, Schwartz S, Wagner L and Miller W 2000 A greedy algorithm for aligning DNA sequences. *J. Comput. Biol*. **7** 203–214

MS received 01 February 2011; accepted 31 May 2011

*e*Publication: 16 August 2011

Corresponding editor: REINER A VEITIA