BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 1 2014, pages 17–23
doi:10.1093/bioinformatics/btt256

Sequence analysis

Advance Access publication May 3, 2013

Further Steps in TANGO: improved taxonomic assignment in metagenomics

Daniel Alonso-Alemany[1], Aurélien Barré[2], Stefano Beretta[3], Paola Bonizzoni[3], Macha Nikolski[2,4] and Gabriel Valiente[1,*]

[1]Department of Software, Technical University of Catalonia, E-08034 Barcelona, Spain, [2]Université Bordeaux, Bordeaux Bioinformatics Center (CBiB), F-33000 Bordeaux, France, [3]Dipartimento di Informatica Sistemistica e Comunicazione, Università Degli Studi di Milano-Bicocca, I-20125 Milan, Italy and [4]Université Bordeaux, Laboratoire Bordelais de Recherche en Informatique (CNRS/LaBRI), F-33405 Talence, France

Associate editor: Michael Brudno

ABSTRACT

Motivation: TANGO is one of the most accurate tools for the taxonomic assignment of sequence reads. However, because of the differences in the taxonomy structures, performing a taxonomic assignment on different reference taxonomies will produce divergent results.

Results: We have improved the TANGO pipeline to be able to perform the taxonomic assignment of a metagenomic sample using alternative reference taxonomies, coming from different sources. We highlight the novel pre-processing step, necessary to accomplish this task, and describe the improvements in the assignment process. We present the new TANGO pipeline in details, and, finally, we show its performance on four real metagenomic datasets and also on synthetic datasets.

Availability: The new version of TANGO, including implementation improvements and novel developments to perform the assignment on different reference taxonomies, is freely available at http://sourceforge.net/projects/taxoassignment/.

Contact: valiente@lsi.upc.edu

Received on March 18, 2013; revised on April 23, 2013; accepted on April 30, 2013

1 INTRODUCTION

Analysis of microbial communities has been until recently a complicated task because of the high diversity and the fact that a large number of these organisms cannot be cultured. Current next-generation sequencing technologies have provided an opportunity for doing this analysis routinely (Petrosino *et al.*, 2009). However, the unprecedented amount of generated data represents a major challenge for computational analysis, which has become an essential tool for microbial genomics. Indeed, computational methods for high-throughput genomic analysis have become the bottleneck of microbial genomics.

A number of computational methods have been recently proposed to solve the issue of species identification within microbial communities, most of them based on sequence similarity and phylogeny (Dröge and McHardy, 2012; Li *et al.*, 2012; Mande *et al.*, 2012). Nevertheless, with the growing number of sequenced samples, the improvement of sequencing technologies in terms of read-length, and the improvement of reference genome libraries, computational complexity of handling the metagenomic data lags behind the needs of current analyses. Hence, further reducing the computational complexity becomes a central challenge for metagenomic analysis methods.

In this article, we improve on the previously proposed TANGO algorithm (Alonso-Alemany *et al.*, 2011; Clemente *et al.*, 2011). TANGO is among the most accurate of all the recently proposed tools for the assignment of reads to organisms based on the computation of the lowest common ancestor (LCA); see (Ribeca and Valiente, 2011) for a recent survey. More precisely, the TANGO algorithm starts from a reference taxonomy and a set of sequence read alignments and looks for the assignment of sequence reads at the best possible taxonomic rank. This procedure allows one to better understand the composition of a metagenomic sample. TANGO is particularly suited to the taxonomic assignment of ambiguous sequence reads, that is, sequence reads with more than one candidate match to organisms; for instance, with the same E-value as the top BLAST hit. Starting with a set of sequence reads, it assigns each of them to ancestral taxons by computing the best suited of the least common taxonomic ancestors for all possible subsets of the set of sequence reads. It relies on an efficient evaluation of the number of mismatches between the sequence read and the reference taxonomy to balance the relevance of precision and recall in the assignment.

Assignment of reads to reference sequences is a necessary step in 16S ribosomal RNA gene-based identification. The 16S gene is crucial in bacterial species identification because it is conserved among organisms within a species while diverging across species. The obtained matches are further used for microbial identification by relying on a specific taxonomic scheme, as taxonomic read assignment provides more accurate identification than comparing reads from 16S ribosomal RNA with known databases by using clustering methods (Ribeca and Valiente, 2011).

Anyway, as different taxonomies can differ in both the topology and the adopted names (used to label the nodes), the taxonomic assignment process could produce divergent results when using different reference taxonomies. Moreover, to our knowledge, there is no tool that allows one to perform taxonomic assignments on different reference taxonomies.

*To whom correspondence should be addressed.

For this reason, we have designed a novel pipeline of the TANGO software that addresses specific issues in sequence read assignment because of characteristics of the input data, namely, taxonomies coming from different database sources. Contracting heterogeneous taxonomies to a common set of taxonomic ranks and then efficiently switching the sequence read assignment from one taxonomy (equalizing phase) to another one are two main novel features of TANGO. More precisely, we have realized a pre-processing of the input taxonomies, which makes the assignment of TANGO more flexible to process data.

In this article, we address the aforementioned issues that are relevant for speeding up and improving the quality of sequence read assignment by two novel phases in the TANGO pipeline that aim to pre-process the input data: *contraction* and *equalizing* (Fig. 1). Contracting taxonomies means to reduce each tree from a collection of heterogeneous taxonomies to a standard set of common taxonomic ranks; that is, the number of levels of the tree are the same for all trees in the collection. As trees from the previous phase are still different in the number of nodes for each level, the equalizing phase produces a correspondence between leaves and nodes of the contracted taxonomies: the mapping will allow an automatic translation of the sequence read alignments (or matches) from one taxonomy to the other.

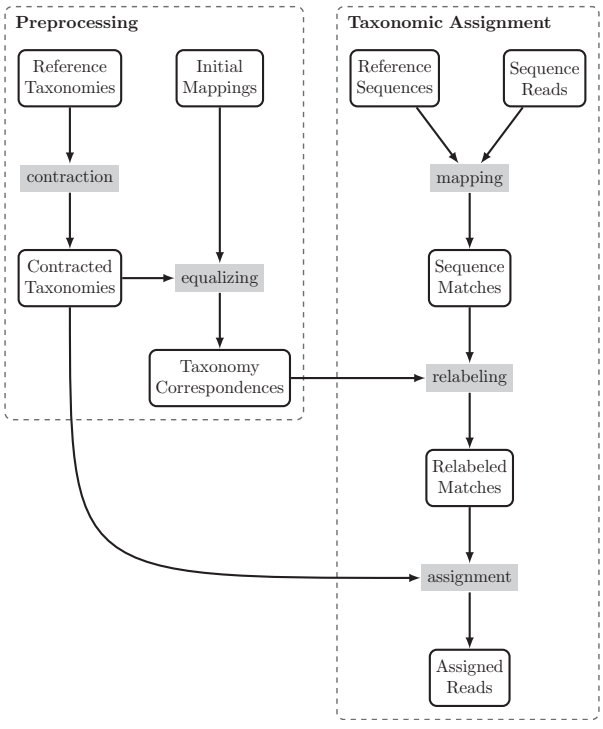


Fig. 1. Taxonomic assignment pipeline. In the pre-processing step, reference taxonomies are contracted to seven taxonomic ranks, and, then, starting from the obtained contracted taxonomies and the initial mappings provided with the original taxonomies, equivalences are computed among the contracted taxonomies (left part). In the taxonomic assignment step, once the sequence reads have been mapped to reference sequences, the best stratum of candidate sequences for each read (sequence matches) are obtained and taxonomic assignment proceeds by re-labeling the sequence matches using the taxonomy correspondences obtained in the equalization. After this operation, the assignment, at the best possible taxonomic rank, of the re-labeled matches to candidate organisms of the contracted taxonomies is performed (right part)

One of the advantages of this new pipeline, and in particular of the pre-processing step, is that it is preformed only once, independently from the set of input reads. In this work, we have considered the most used reference taxonomies, that are NCBI (Federhen, 2012), RDP (Cole *et al.*, 2009) and Greengenes (McDonald *et al.*, 2012). As anticipated before, the differences among taxonomies have a direct impact on the bacterial identification results. For example, there are 43 phyla in the RDP database, 94 in NCBI and 110 in Greengenes, and a number of attempts at taxonomy reconciliation have been undertaken, such as by McDonald *et al.* (2012). However, no consensus taxonomy exists at this day. Consequently, being able to compare assignments based on different taxonomies is particularly interesting, as well as being able to switch between taxonomies.

As a final remark, we want to observe that the first available implementation of the TANGO method was straightforward and did not aim at any particular data structure or computational optimizations. Therefore, the new version of the TANGO software, in addition to the novel functions mentioned before, provides an improvement in computational efficiency that is achieved thanks to the adoption of compact data structures and sound programmatic practices. In the rest of the article, we will also illustrate the contributions of these improvements on real and synthetic data on the aforementioned reference taxonomies.

2 METHODS

2.1 Preliminaries

The effective assignment of next-generation sequence reads to microbial species at the best possible taxonomic rank relies on a well-curated microbial taxonomy. Taxonomies are usually represented as general, n-ary trees that classify organisms at seven taxonomic ranks: kingdom, phylum, class, order, family, genus and species (which are usually referred to as KPCOFGS). Additional ranks such as *subfamily* or *superclass* are sometimes introduced to refine the classification. Species are usually the leaves of these trees, unless their full classification is still incomplete. More formally, let T be a taxonomic tree, that is, an n-ary tree rooted in r, in which every node is labeled (depending on its taxonomic rank). Moreover, given a taxonomic tree T, we will denote as $N(T)$ the set of nodes of T and $L(T)$ the set of leaves of T. Also, $I(T) = N(T) \setminus (L(T) \cup \{r\})$ will denote the set of internal nodes of T. Finally, the *taxonomic ranks* of T will be referred to as $R(T)$.

2.2 Pre-processing of taxonomies

Contracting reference taxonomies The particular taxonomy used as a reference for the classification of next-generation sequence reads, and the way it is modeled, is an important factor that poses constraints on the ability of an algorithmic method to discriminate among related species. The reference taxonomies (Santamaria *et al.*, 2012) most widely used for metagenomic analysis—NCBI, RDP and Greengenes—differ, among other aspects, in the number of taxonomic ranks that are used to classify organisms, in the completeness of their classification and in the tree structure.

To deal with such heterogeneous taxonomies, we first contract them to the aforementioned seven taxonomic ranks. The contraction of a taxonomic tree to a given set of taxonomic ranks is the subtree induced by these taxonomic ranks: the nodes of the contracted tree are the nodes of the original tree at these taxonomic ranks, and the branches of the contracted tree correspond to non-trivial paths in the original tree.

However, as leaves are the nodes in the tree that are associated with genomic sequences, we have to retain the leaves to be able to use them when assigning sequence reads. Consequently, even leaves located at ranks that are not part of the desired set are retained.

In general, let us define the contraction of a taxonomic tree, with respect to a subset of its taxonomic ranks.

Definition 1. *Let T be a taxonomic tree in which every node x is labeled with its taxonomic rank, that is, $rank(x) \in R(T)$, and let $R'(T) \subseteq R(T)$ be a set of valid ranks. The contracted tree T', derived from T with respect to $R'(T)$, is the tree such that $L(T') = L(T)$ and the two trees T' and T have the same root. Moreover, for each $x \in I(T)$, if $rank(x) \in R'(T)$, then $x \in I(T')$, and for each edge $(x, y) \in T'$, y is a proper descendant of x in T.*

Observe that the aforementioned notion of contracted tree assumes that both the root node and the leaves are kept in the contracted tree T'. As anticipated before, the reason why we have decided to keep all the leaves (also if they do not have a *valid rank*) is that these nodes of the tree are usually the ones with an associated genomic sequence. This means that, in the input set, the valid matches of the reads usually refer to nodes that are leaves of the tree. In fact, genomic sequences of these nodes are used for the alignment of the reads to the taxonomic tree. Contracting such nodes would cause invalid matches in the input set.

Starting from Definition 1, we have formulated the *Taxonomic Tree Contraction* problem as follows:

- **Input**: a taxonomic tree T in which every node x is labeled with its taxonomic rank $rank(x) \in R(T)$, and a set $R'(T) \subseteq R(T)$ of valid ranks.
- **Output**: the *contracted* tree T', derived from T with respect to $R'(T)$.

This problem can be solved by performing a *post-order traversal* of the tree T, which guarantees that, when visiting a node, all its children are already visited. During the visit, if a node has a label that is not among the valid ranks, it is contracted by assigning all its children to the parent node. Otherwise the node is maintained as it is.

Observe that the set $R'(T)$ in Taxonomic Tree Contraction problem consists of exactly the previously mentioned seven taxonomic ranks, meaning that the contracted taxonomic trees have all the same rank levels (and also the maximum depth). Solving the Taxonomic Tree Contraction problem is a necessary preliminary step of the TANGO pipeline, which is performed on all the considered taxonomic trees before doing the taxonomic assignments of the reads.

This method allows us to contract different trees to the same set of taxonomic ranks. Once we have contracted the reference taxonomies, we resolve any incomplete classification by first extracting the lineages of the organisms and then, inferring the taxonomic tree from the lineages.

Equalizing reference taxonomies Contracted taxonomic trees, although having the same depth (number of levels), may still differ in structure and number of nodes for each level. To be able to assign sequence reads using contracted heterogeneous taxonomies, we designed a procedure to establish correspondences between the nodes of different taxonomies, called *equalizing*. Such correspondence will only be used to re-label sequence matches, as detailed in the next section. The basic idea of the procedure is to process two input taxonomies T_1 and T_2 to build a correspondence from T_1 to T_2, that is, the *equalizing* mapping ϕ, that maps each node x of T_1 to a node $y = \phi(x)$ of T_2.

The mapping ϕ is built recursively, by a post-order traversal of the tree that guarantees that when defining the mapping $\phi(x)$ for a node x, such a mapping for the children of x, has been already computed. In fact, for each node x of taxonomy T_1 having children the nodes $x_1, x_2, \ldots, x_n$, $\phi(x)$ is the LCA of nodes $\phi(x_1), \phi(x_2), \ldots, \phi(x_n)$ in T_2. Observe that for a node $x \in L(T_1)$, $\phi(x)$ is provided within the annotated taxonomies. It must be pointed out that the equalizing mapping ϕ is not bijective; thus, to compare all contracted taxonomies for each pair of considered taxonomies T_1 and T_2, we have to compute the two possible equalizing mappings.

These correspondences between taxonomies can then be used to translate the read matches, that is, to assign the sequence reads using a taxonomy that is different from the one labeled by the genomic sequences to which the sequence reads are mapped. Contraction and equalization of reference taxonomies are both done only once, in a pre-processing step, when new releases of the taxonomies become available.

Figure 2 shows the result of equalizing the NCBI taxonomy to the Greengenes taxonomy for the Aquificae and the Thermodesulfobacteria phyla. The Aquificae phylum, along with the Aquificae class, the Aquificales order, the Aquificaceae family and the Thermocrinis and Hydrogenobacter genera, are all taken to correspond to the Bacteria kingdom because their descendent sequences in the NCBI taxonomy are split among different clades in the Greengenes taxonomy; the same holds for the Thermodesulfobacteria phylum, the Thermodesulfobacteria class, the Thermodesulfobacteriales order and the Thermodesulfobacteriaceae family. Also, the Thermovibrio and Desulfurobacterium genera are taken to correspond to the Desulfurobacteriaceae family, the Sulfurihydrogenibium, Persephonella and Hydrogenothermus genera to the Hydrogenothermaceae family and the Aquifex genus to the Aquificaceae family, and the Thermodesulfobacterium genus is taken to correspond to the Thermodesulfobacteriales order. The Phorcysia genus does not correspond to any rank in the Greengenes taxonomy; though, because none of its descendant sequences are properly annotated.

The example reported in Figure 2 highlights some 'anomalies' that could be present in the equalizing mapping. More specifically, because of the presence of nodes (of the first taxonomy) having their descendants split among different clades (in the second taxonomy), the resulting equalizing mappings of those nodes and their descendants are referred to a unique (higher) node. In other words, all the nodes (which usually correspond to not *monophyletic* groups) are collapsed to a unique node by the equalizing mapping.

Fig. 2. Equalizing reference taxonomies. When equalizing the contracted NCBI taxonomy (left) to the contracted Greengenes taxonomy (right), the Aquificales order, the Aquificaceae and Thermodesulfobacteriaceae families and the Thermocrinis and Hydrogenobacter genera are moved up to the Bacteria kingdom; the Thermovibrio and Desulfurobacterium genera to the Desulfurobacteriaceae family; the Sulfurihydrogenibium, Persephonella and Hydrogenothermus genera to the Hydrogenothermaceae family; the Aquifex genus to the Aquificaceae family; and the Thermodesulfobacterium genus to the Thermodesulfobacteriales order

Anyway, note that the main goal of the equalizing mapping from T_1 to T_2 is to obtain from the alignment of reads for taxonomy T_1, a corresponding alignment of reads for taxonomy T_2. In other words, the equalization from T_1 to T_2 will be used to only translate the input referred to taxonomy T_1 to an input referred to taxonomy T_2. Observe that the process of read assignment for T_2, which is done by the TANGO procedure, may be only partially influenced by the mapping ϕ, as it is mainly based on the read alignment of the leaves of the two taxonomies. More important, the fact that we have anomalies in the mapping of higher nodes of the taxonomies, as observed in Figure 2, does not highly affect the TANGO procedure, as we use the ϕ mapping only to translate the input between taxonomies, which is mainly referred to leaves whose mapping is assumed to be correct (as it is provided with the taxonomy).

Finally, we want to remark that the pre-processing step, that is, contraction and equalizing of taxonomies, does not depend on the input set of reads, and it is performed only when there are changes (new releases) in the taxonomies.

2.3 Taxonomic assignment

Re-labeling sequence matches The assignment of sequence reads using heterogeneous taxonomies also requires re-labeling the sequence matches in a given taxonomy to another one. This is done for every input set of sequence reads, by applying the appropriate correspondences between taxonomies obtained in the pre-processing step with the equalizing mapping computation.

More precisely, let s be an input sequence read and let $M_s = \{x_1, x_2, \ldots, x_n\}$, where $x_i \in N(T_1), 1 \leq i \leq n$ be its set of alignments to the taxonomy T_1. In the re-labeling step, we translate the matches in M from T_1 to another taxonomy T_2 by using the mapping ϕ from T_1 to T_2, computed in the equalizing step. The result of the re-labeling procedure is the set $M'_s = \{\phi(x_1), \phi(x_2), \ldots, \phi(x_n)\}$, where $\phi(x_i) \in N(T_2), 1 \leq i \leq n$.

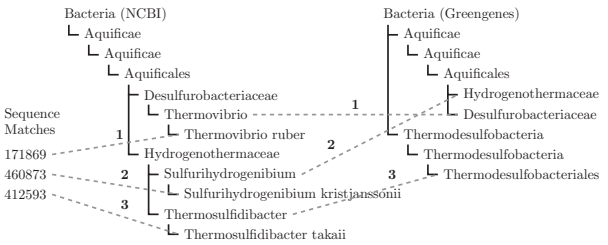


Fig. 3. Matches to sequences in the NCBI taxonomy are re-labeled to matches to the Greengenes taxonomy by using the taxonomy correspondences obtained in the pre-processing step. Numbers on the dashed mappings represent the three paths used to translate the matches to the NCBI taxonomy into matches to the Greengenes taxonomy

We also want to point out that because of the observations done in the equalizing section regarding the possible anomalies of the mapping ϕ, it is not guaranteed that corresponding nodes in the second taxonomy exist, and, moreover, that all the nodes are mapped to distinct nodes.

Sequence matches are always re-labeled, even when the original and the new taxonomies are the same, because sequence reads may have been matched to genomic sequences of organisms at some rank other than the seven taxonomic ranks in the contracted taxonomies.

Figure 3 shows the result of re-labeling matches to sequences in the NCBI taxonomy to matches to the Greengenes taxonomy. A read matching sequences of the Thermovibrio, the Thermosulfidibacter and the Sulfurihydrogenibium genera that are not present in the Greengenes taxonomy, is re-labeled to match sequences of the Desulfurobacteriaceae family, the Thermodesulfobacteriales order and the Sulfurihydrogenibium genus, respectively.

Assigning sequence reads The efficient assignment of next-generation sequence reads to heterogeneous taxonomies poses computational challenges because of the huge number of sequence reads that have to be processed in metagenomic analyses.

Once reference taxonomies have been contracted and equalized, sequence reads have been aligned to reference sequences, and sequence matches have been re-labeled, the final step of the taxonomic analysis pipeline involves resolving ambiguities by assigning those sequence reads with more than one candidate match to organisms at the closest possible taxonomic rank. The TANGO algorithm (Alonso-Alemany *et al.*, 2011; Clemente *et al.*, 2011) goes beyond LCA assignment methods (Huson *et al.*, 2007) by assigning each sequence read to a node in the reference taxonomy that provides for the best possible sensitivity and specificity.

The first release of the TANGO software (Clemente *et al.*, 2011) was only able to process small reference taxonomies in Newick format. We have replaced the object-based representation of phylogenetic trees in BioPerl (Stajich *et al.*, 2002) in this new release of the TANGO software with a hash-based, first-child next-sibling representation of general trees (Knuth, 1997). This allows for the effective taxonomic assignment of large metagenomic datasets using heterogeneous taxonomies, as discussed later in the text.

Taxonomic assignment is based on the penalty score that is computed at the LCA level for every node. Consequently, if for a given read the LCA is close to the root of the phylogenetic tree, then the penalty score computation can be invoked many times. An appropriate programmatic

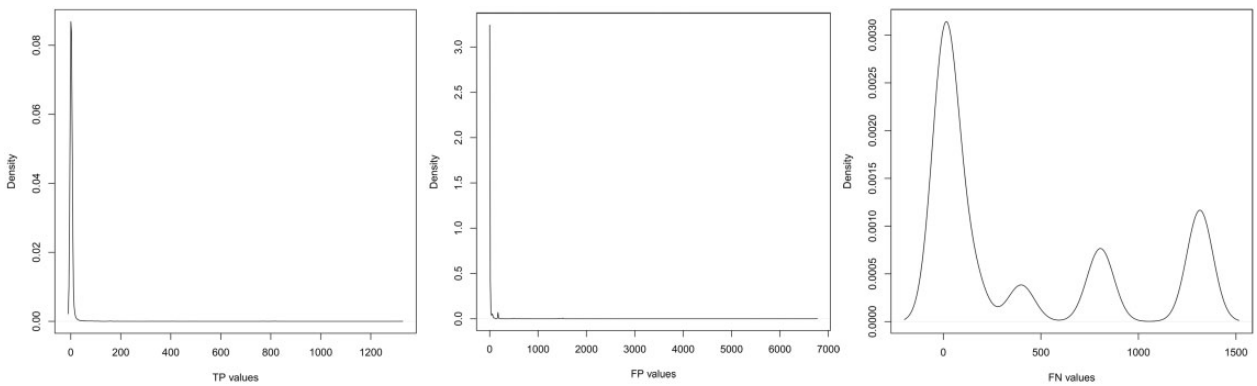


Fig. 4. Empirical distributions of TP, FP and FN. Values collected during penalty score computation for every read on a test case with 500 000 reads against Greengenes (available on http://sourceforge.net/projects/taxoassignment/)

solution is to pre-compute the result for the most frequent cases. The important question to answer here is the balance between the number of pre-computed scores (and consequently memory consumption) and the gain in execution time.

The penalty score is calculated based on the values of the parameter q (set by the user) and from the numbers of true positives (TP), false negatives (FN) and false positives (FP). We have examined on our test cases the distributions of values that these latter three parameters can take (Fig. 4). It appears that TP and FP take values often comprised within a tight interval (from 0 to 10), whereas FN follows a multi-modal distribution. We hypothesize that the four modes corresponding to four local maximums (see panel 'FN value' in Fig. 4) may represent FN values for different levels in the taxonomy. Imagine a read that has many matches. For such read, we will test many nodes situated at low levels of the taxonomy, thus generating high numbers of false negatives. The LCA for this read will most probably be rather high in the taxonomy. Going through the taxonomy from level i to level $i-1$ for this read, any given node will cover significantly more matches and thus produce less false negatives. Consequently, if the number of children was roughly the same at different levels of the taxonomy, the curve would be smooth. But in fact, at the bottom of the tree, the number of children of each node is very high and at the root, and at the intermediate levels, this number is very low. Thus, the multimodal form of the curve for FN. The most frequent values for FN fall within the mode having the highest peak and centered $\sim$1.

3 RESULTS

Metagenomic datasets

To evaluate the performance of our algorithm, we have analyzed four metagenomic datasets: marine environment samples (16S ribosomal RNA, V6 region, 222 291 reads) (Sogin *et al.*, 2006), mice gut samples (16S ribosomal RNA, V2 region, 1 119 519 reads; and V6 region, 817 942 reads) (Turnbaugh *et al.*, 2009) and rat gut samples (16S ribosomal RNA, V4 region, 515 112 reads) (Manichanh *et al.*, 2010).

We have analyzed these metagenomic datasets using three alternative taxonomies: the NCBI Taxonomy (Federhen, 2012) release December 21, 2012, with 833 216 reference sequences, of which we considered for read alignment only the 7543 sequences of the NCBI RefSeq Targeted Loci Project (Pruitt *et al.*, 2012) release December 12, 2012; the RDP Taxonomy (Cole *et al.*, 2009) release 10.31, with 2 639 157 reference sequences; and the Greengenes taxonomy (McDonald *et al.*, 2012) release May 9, 2011, with 406 997 reference sequences.

Mapping the metagenomic datasets with BLAST (Altschul *et al.*, 1990) to the NCBI RefSeq reference sequences, and taking as candidate alignments all those sequences with the same E-value as the top BLAST hit, with a 0.001 cut-off, resulted in 40 197 ambiguous reads (marine environment), 53 810 and 58 293 ambiguous reads (mice gut) and 226 637 ambiguous reads (rat gut); mapping them to the RDP reference sequences resulted in 195 030 ambiguous reads (marine environment), 899 291 and 797 513 ambiguous reads (mice gut) and 377 106 ambiguous reads (rat gut); and mapping them to the Greengenes reference sequences resulted in 153 423 ambiguous reads (marine environment), 937 294 and 792 697 ambiguous reads (mice gut) and 515 112 ambiguous reads (rat gut).

We have evaluated the improved TANGO 3 pipeline, in which the BioPerl representation of taxonomies is replaced by a hash-based, first-child next-sibling representation of general trees, along with parameter partial evaluation. Performance evaluation on the four metagenomic datasets is shown in Table 1.

Table 1. Performance evaluation on the four considered metagenomic datasets

Reference taxonomy	Pre-processing	Dataset			
		(a)	(b)	(c)	(d)
NCBI	63.41	12.62	64.29	33.08	48.74
RDP	203.82	890.86	4090.30	5526.83	662.79
Greengenes	27.34	20.85	156.32	210.38	39.77

Note: Time in seconds on an Intel Xeon X5670 with 32 GB memory running at 2.93 GHz for the pre-processing of the three reference taxonomies and the taxonomic assignment of the four metagenomic datasets using TANGO 3 (BioPerl representation of taxonomies replaced by a hash-based representation of general trees, and parameter partial evaluation). (a) marine environmental samples; (b) mice gut samples, V2 region; (c) mice gut samples, V6 region; (d) rat gut samples.

Synthetic datasets

Empirical evaluation on our test set resulted in the choice of pre-computing the scores in the form of a matrix M of size $10 \times 10 \times 10$, choice conservative in terms of memory and efficient in execution time speed-up. However, this matrix alone is not sufficient. Indeed, the q parameter still remains free. Consequently, we have pre-computed M_q for values of q between 0 and 1 with a step of 0.1. If the value of q provided by the user corresponds to a pre-computed M_q, then it is used; otherwise, M_q is generated on demand. See Table 2.

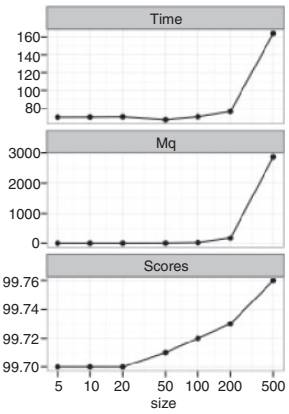
Taxonomic assignment is evaluated using the NCBI, RDP and Greengenes phylogenies. To further evaluate the performance of our algorithm and, in particular, the influence of the graph representing the phylogeny, we have generated random taxonomies to measure the variation in terms of time consumption. This generation was done by *re-sampling* to resemble real taxonomic structures.

A taxonomic tree has a root and seven underlying levels (KPCOFGS), its leaves corresponding to the operational taxonomic units. Phylogenetic classification of certain operational taxonomic units is not precisely known; consequently, some leaves can be children of nodes at depth smaller than seven. We emulate this topology starting from the root up to level S (species) by re-sampling the number of children at each level, which is equivalent to randomly permuting branches at each level [re-sampling has been chosen over a parametric model of number of children for two reasons: (i) the latter does not allow for setting the size of the resulting tree (its number of nodes) and (ii) no theoretical distribution fitted well enough to real data]. Once all the internal nodes are generated, leaves have to be attached to terminal nodes. Again, we use random sampling without replacement to follow the distribution of number of leaves per terminal node.

We have evaluated the contribution of different improvements to the TANGO pipeline as well as the influence of taxonomic structure (namely, NCBI, RDP and Greengenes) on time performance. Three versions of TANGO were tested: (i) TANGO 1,

Table 2. Influence of pre-computing the penalty score matrix on execution time and memory use

Size	Time	$\lvert M_q \rvert$	%Scores
0	80.2		
5^3	70.2	3 Kb	99.70
10^3	70.3	19 Kb	99.70
20^3	70.6	157 Kb	99.70
50^3	67.4	2,5 Mb	99.71
100^3	70.7	20 Mb	99.72
200^3	76.5	171 Mb	99.73
500^3	164.0	2,8 Gb	99.76



Note: Column *size* shows the number of entries of the M_q matrix, column time shows user running time in seconds on an AMD opteron 6134 running at 2.3 GHz, column $|M_q|$ shows the memory used by the matrix, and column %Scores shows the per cent of calls that correspond to a pre-computed M_q and do not require its generation at execution time. Same values are traced on the figure to the right. Evaluation done using Greengenes taxonomy and a test case with 500 000 reads.

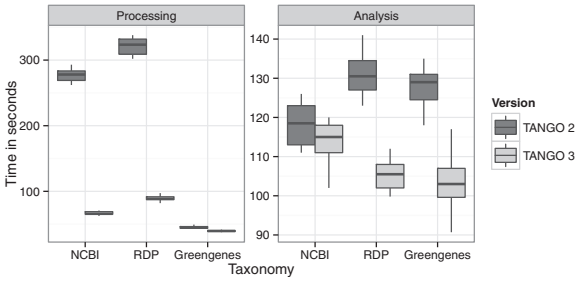


Fig. 5. Performance evaluation on synthetic datasets. Performance evaluation (execution time on an AMD operton 6134 running at 2.3 GHz) measured for 100 random instances. Random instances are generated by re-sampling to be Greengenes-, NCBI- and RDP-like. Time was measured separately for the pre-processing and taxonomic assignment steps of the pipeline. Reported times for TANGO 2 (BioPerl representation of taxonomies replaced by a hash-based representation of general trees) and TANGO 3 (final version with parameter partial evaluation)

the original version as published in Clemente *et al.* (2011), (ii) TANGO 2, the version where the BioPerl representation of taxonomies is replaced by a hash-based, first-child next-sibling representation of general trees and (iii) TANGO 3, the final version where parameter partial evaluation is added to the TANGO 2 solution. All three versions of TANGO were evaluated on 100 random instances generated as explained here earlier in the text, emulating the three reference taxonomies (NCBI, RDP and Greengenes). Figure 5 shows performance evaluation results (TANGO 1 results are not shown, as execution time exceeded 1 h for all cases).

4 DISCUSSION

Analysis of microbial communities and the possibility of their characterization by the NGS metagenomic approaches depend on the availability of reference databases and taxonomies. Up to date, major efforts have been deployed for the development of algorithmic approaches for taxonomic assignment. The taxonomic information for this assignment can be obtained from various and often discordant sources. *Further steps in TANGO* method address this limitation by providing means for contracting heterogeneous taxonomies to a common set of taxonomic ranks and then efficiently switching the sequence read assignment from one taxonomy (equalizing phase) to another.

Indeed, despite the tremendous effort spent in classifying bacteria and archaea, even today not enough is known about evolutionary relationships to establish clearly defined taxonomic classes and orders for many of them. Databases, such as the NCBI, RDP and Greengenes, provide access to sets of ribosomal RNA sequence databases necessary for the identification of microbes in a culture-independent analysis of microbial communities. However, one of the hurdles is the fact that these taxonomies do not contain all of the taxonomic levels attached to the published names of the bacterial and archaeal sequences, and, moreover, that there are major topological differences among them. Consequently, taxonomic assignment will produce divergent results when using different reference taxonomies, and there was no tool up to now to freely move between taxonomic assignments that rely on different reference taxonomies.

In this article, we improve on our TANGO method for taxonomic classification of microbial communities. The pre-requisite to using TANGO—as well as other taxonomic assignment tools—is to perform an alignment of the input sequence reads against reference 16S ribosomal RNA sequences and collecting positive hits. A metagenomic sequence read is then classified by computing the LCA of the species in the set of hits for this sequence read. TANGO improves classification accuracy by balancing the true and false positive assignments depending on different taxonomic levels. Furthermore, TANGO now allows freely moving between different taxonomies and enables easy comparison of taxonomic assignments that rely on competing classifications.

The general problem of reconciling different reference taxonomies is still open, and we have only provided in this article a partial answer as LCA preserving mapping, which does not provide much information when the taxonomies have conflicting classifications. Future work also includes providing an efficient web service for taxonomic assignment of microbial communities with TANGO using heterogeneous reference taxonomies.

The new version of TANGO including implementation improvements and taxonomy contraction and equalization, as well as the accompanying software, are freely available. TANGO can be downloaded from http://sourceforge.net/projects/taxoassignment/. Random taxonomy generator can be downloaded from http://sourceforge.net/projects/randomtaxonomy/.

ACKNOWLEDGEMENT

G.V. wrote the original TANGO software. P.B., M.N. and G.V. designed the new TANGO pipeline. D.A., A.B. and S.B. wrote the

new TANGO software. All authors prepared the manuscript, contributed to the discussion and have approved the final manuscript.

Funding: S.B. and P.B. are supported by the MIUR PRIN 2010–2011 grant 'Automi e Linguaggi Formali: Aspetti Matematici e Applicativi', code H41J12000190001 and FAR grant 'Algorithmic methods and combinatorial structures in Bioinformatics' (Università Degli Studi di Milano-Bicocca).

Conflict of Interest: none declared.

REFERENCES

Alonso-Alemany,D. *et al.* (2011) Taxonomic assignment in metagenomics with TANGO. *EMBnet. J.*, **17**, 46–50.

Altschul,S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Clemente,J.C. *et al.* (2011) Flexible taxonomic assignment of ambiguous sequencing reads. *BMC Bioinformatics*, **12**, 8.

Cole,J.R. *et al.* (2009) The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res.*, **37**, D141–D145.

Dröge,J. and McHardy,A.C. (2012) Taxonomic binning of metagenome samples generated by next-generation sequencing technologies. *Brief. Bioinform.*, **13**, 646–655.

Federhen,S. (2012) The NCBI taxonomy database. *Nucleic Acids Res.*, **40**, D136–D143.

Huson,D.H. *et al.* (2007) MEGAN analysis of metagenomic data. *Genome Res.*, **17**, 377–386.

Knuth,D.E. (1997) Binary Tree Representation of Trees. In: *The Art of Computer Programming*. Vol. 1, 3rd edn. Fundamental Algorithms, Reading, Massachusetts: Addison-Wesley, pp. 334–348.

Li,W. *et al.* (2012) Ultrafast clustering algorithms for metagenomic sequence analysis. *Brief. Bioinform.*, **13**, 656–668.

Mande,S.S. *et al.* (2012) Classification of metagenomic sequences: methods and challenges. *Brief. Bioinform.*, **13**, 669–681.

Manichanh,C. *et al.* (2010) Reshaping the gut microbiome with bacterial transplantation and antibiotic intake. *Genome Res.*, **20**, 1411–1419.

McDonald,D. *et al.* (2012) An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J.*, **6**, 610–618.

Petrosino,J.F. *et al.* (2009) Metagenomic pyrosequencing and microbial identification. *Clin. Chem.*, **55**, 856–866.

Pruitt,K.D. *et al.* (2012) NCBI reference sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res.*, **40**, D130–D135.

Ribeca,P. and Valiente,G. (2011) Computational challenges of sequence classification in microbiomic data. *Brief. Bioinform.*, **12**, 614–625.

Santamaria,M.B. *et al.* (2012) Reference databases for taxonomic assignment in metagenomics. *Brief. Bioinform.*, **13**, 682–695.

Sogin,M.L. *et al.* (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proc. Natl Acad. Sci. USA*, **103**, 12115–12120.

Stajich,J.E. *et al.* (2002) The BioPerl toolkit: Perl modules for the life sciences. *Genome Res.*, **12**, 1611–1618.

Turnbaugh,P.J. *et al.* (2009) A core gut microbiome in obese and lean twins. *Nature*, **457**, 480–484.