Unbiased Taxonomic Annotation of Metagenomic Samples

Bruno Fosso[1], Graziano Pesole[1], Francesc Rosselló[2], and Gabriel Valiente[3](✉)

[1] Institute of Biomembranes and Bioenergetics,
Consiglio Nazionale delle Ricerche, 70126 Bari, Italy
[2] Department of Mathematics and Computer Science,
Research Institute of Health Science,
University of the Balearic Islands, 07122 Palma de Mallorca, Spain
[3] Algorithms, Bioinformatics, Complexity and Formal Methods Research Group,
Technical University of Catalonia, 08034 Barcelona, Spain
valiente@cs.upc.edu

Abstract. The classification of reads from a metagenomic sample using a reference taxonomy is usually based on first mapping the reads to the reference sequences and then, classifying each read at a node under the lowest common ancestor of the candidate sequences in the reference taxonomy with the least classification error. However, this taxonomic annotation can be biased by an imbalanced taxonomy and also by the presence of multiple nodes in the taxonomy with the least classification error for a given read. In this paper, we show that the Rand index is a better indicator of classification error than the often used area under the ROC curve and F-measure for both balanced and imbalanced reference taxonomies, and we also address the second source of bias by reducing the taxonomic annotation problem for a whole metagenomic sample to a set cover problem, for which a logarithmic approximation can be obtained in linear time.

Keywords: Metagenomics · Classification · Taxonomic annotation · Correlation · Set cover

1 Introduction

Next generation sequencing technologies have moved forward the development of metagenomics, a new field of science devoted to the study of microbial communities by the analysis of their genomic content, directly sequenced from the environment [15,20,21]. A sequenced metagenomic sample consists of a large number of relatively short DNA or RNA fragments, called reads, and one of the first steps in the computational analysis of a metagenomic sample is the identification of the organisms present in the sequenced environment and their relative abundance, that is, the classification of the metagenomic sample.

In this paper, we focus on the taxonomic annotation problem, that is, the classification of the reads from a metagenomic sample using a reference taxonomy, for which we adapt some basic notions from statistical classification in

	Positive prediction	Negative prediction
Positive class	True Positive (TP)	False Negative (FN)
Negative class	False Positive (FP)	True Negative (TN)

Fig. 1. Confusion matrix for a binary classification problem

machine learning. We abstract away from the computational problem of mapping reads to reference sequences, and assume that a set of candidate sequences in a reference taxonomy is given for each read in the metagenomic sample to be classified. These candidate sequences are usually obtained either by sequence composition methods (those reference sequences with oligonucleotide frequencies within a given distance threshold to the oligonucleotide frequencies of the read) or by sequence similarity methods (those reference sequences that the read can be aligned to within a given threshold of sequence similarity, or those reference sequences that the read can be mapped to with at most a given number of mismatches).

In a statistical binary classification problem, the confusion matrix (Fig. 1) shows the number of correctly and incorrectly classified instances of each class. True positives (TP) are the correctly classified positive instances, true negatives (TN) are the correctly classified negative instances, false positives (FP) are the misclassified negative instances, and false negatives (FN) are the misclassified positive instances. The *true positive rate*, *sensitivity*, or *recall* R of a classification is the ratio $TPR = TP/(TP + FN)$ of true positives to the total number of positive instances, the *false positive rate* is the ratio $FPR = FP/(FP + TN)$ of false positives to the total number of negative instances, the *true negative rate* or *specificity* is the ratio $TNR = TN/(FP + TN)$ of true negatives to the total number of negative instances, and the *false negative rate* is the ratio $FNR = FN/(TP + FN)$ of false negatives to the total number of positive instances. Further, the *precision* of a classification is the ratio $P = TP/(TP + FP)$ of true positives to the total number of positive predictions. They are usually combined into a single indicator of classification error as either the area under the ROC curve $AUC = (TPR - FPR + 1)/2$ or the F-measure, which is the harmonic mean $F = 2/(1/P + 1/R)$ of precision and recall [18].

In a metagenomic classification problem, the annotation of a read as coming from a particular sequence in a reference taxonomy often involves solving the ambiguity of multiple candidate sequences, caused among other factors by reads being not long enough to ensure a unique identification of the reference sequences they come from. Reference taxonomies are rooted trees, with the leaves labeled by sequences at the taxonomic rank of species or strain, and these ambiguities are solved by annotating reads as coming from internal nodes, at higher taxonomic ranks in the reference taxonomy. When classifying a read as coming from an internal node in a reference taxonomy (Fig. 2), the leaves under the internal node are true positives if they are labeled by candidate sequences, otherwise they are false positives, and the remaining leaves under the lowest common ancestor (LCA) of the candidate sequences are false negatives if they are labeled

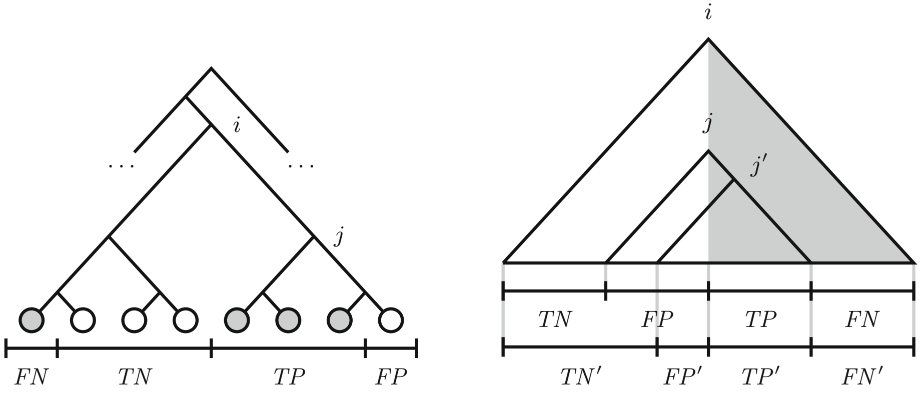

Fig. 2. Classifying a read using a reference taxonomy. The grayed leaves are the candidate sequences for the classification of the read, and node i is their LCA in the reference taxonomy. The taxonomic annotation of the read at node i implies the absence of true negatives and false negatives. With a taxonomic annotation of the read at node j, which is the LCA in the reference taxonomy of the true positives, however, the remaining grayed leaves are the false negatives, the remaining leaves under node j are the false positives, and the still remaining leaves under node i are the true negatives of the metagenomic classification problem

by candidate sequences, otherwise they are true negatives. Annotating a read as coming from the LCA of the candidate sequences in a reference taxonomy [12] maximizes precision, as in that case there are no true negatives and no false negatives, but at the expense of specificity, because the number of false positives in a reference taxonomy can be very large. Annotating a read as coming from an internal node with the largest F-measure value [1,3,8,9] minimizes the classification error as a combination of precision and sensitivity.

However, there are at least two sources of bias in the taxonomic annotation of a metagenomic sample. One the one hand, reference taxonomies are imbalanced, that is, the instances of one class significantly outnumber the instances of the other classes, and this can be observed at any taxonomic rank. For example, the NCBI Taxonomy [5,6], which is the most comprehensive taxonomic reference to date, includes as of 13 March 2017 an imbalanced number of sequences for Bacteria (1,412,065), Eukaryota (685,380), and Archaea (27,322). Within the Bacteria, for example, there is also an imbalanced number of sequences for the Actinobacteria (593,837), Proteobacteria (440,315), Firmicutes (245,632), Bacteroidetes (77,866), Planctomycetes (8,899), Fusobacteria (7,789), and others (37,727). In a statistical binary classification problem, imbalanced datasets result in a good coverage of the positive instances and a frequent misclassification of the negative instances, since most of the standard machine learning algorithms consider a balanced training set [16]. In a metagenomic classification problem, an imbalanced reference taxonomy may also yield an imbalance between the positive and negative classes, because the larger the clade of the LCA in a reference taxonomy of the candidate sequences for a read, the larger the negative class for the classification of the read. In Sect. 2, we show that this is in general not the case, and we also show that the Rand index is a better indicator of classification error than the often used area under the ROC curve and F-measure, when the reference taxonomy is imbalanced and also for balanced reference taxonomies.

Another source of bias in the taxonomic annotation of a metagenomic sample lies in the existence of multiple candidate nodes in a reference taxonomy with the least classification error for a given read, one of which is usually chosen arbitrarily for the taxonomic annotation of the read [1,3]. Instead of breaking ties independently for each read in a metagenomic sample, we show in Sect. 3 that the shift from a one-sequence-read-at-a-time view to a whole-set-of-sequence-reads view yields a better resolution of any remaining ambiguities in the taxonomic annotation of a metagenomic sample.

2 Taxonomic Annotation Using Imbalanced Reference Taxonomies

Recall from Sect. 1 that in a metagenomic classification problem, an imbalanced reference taxonomy yields an imbalance between the positive and negative classes. Let us define the *balance ratio* of a classification problem as the ratio of the size of the positive class to the size of the negative class.

Definition 1. *Let TP, TN, FP, and FN be the number of true positives, false positives, true negatives, and false negatives in a binary classification problem. The balance ratio of the classification problem is $(TP + FN)/(FP + TN)$.*

Recall also from Sect. 1 that the reference taxonomies used in metagenomic classification are highly imbalanced. It turns out that balanced and imbalanced reference taxonomies yield exactly the same metagenomic classification problems, as long as they have the same number of internal nodes. Some evidence supporting this observation follows.

The topology of the most possible balanced binary reference taxonomy is a complete binary tree, as every internal node (and also the root) has two descendant clades of exactly the same size. On the other hand, the topology of the least possible balanced binary reference taxonomy is a degenerate binary tree, as every internal node (and also the root) has one big descendant clade and one small (with only one node) descendant clade.

Now, in a metagenomic classification problem, any subset of the leaves of a reference taxonomy may be labeled by the candidate sequences for the classification of a given read. For a given subset of the leaves of a reference taxonomy, each candidate internal node (at or under the LCA of the subset of the leaves) for the taxonomic annotation of the read yields a certain number of true positives, false positives, true negatives, and false negatives. For example, for the reference taxonomy in Fig. 2, the subset of grayes leaves yields, for the candidate internal node j, a metagenomic classification problem with $TP = 3$, $FP = 1$, $TN = 3$, $FN = 1$ and thus, balance ratio $(3 + 1)/(1 + 3) = 1$. Table 1 shows the distribution of the number of true positives, false positives, true negatives,

Table 1. Distribution of TP, FP, TN, FN values (left) and distribution of $TP + FN$ values (right) in metagenomic classification problems for different taxonomic reference topologies: complete (C) and degenerate (D) binary trees with 8 leaves

TP	FP	TN	FN	C	D
0	2	0	6	4	1
0	2	1	5	24	6
0	2	2	4	60	15
0	2	3	3	80	20
...	...	...	...	...	...
7	0	1	0	0	1
7	1	0	0	8	8
8	0	0	0	1	1

$TP + FN$	Count
1	56
2	196
3	392
4	490
5	392
6	196
7	56
8	7

and false negatives for all subsets of the leaves of a reference taxonomy and for every candidate internal node for the taxonomic annotation of a read having as candidate sequences the subset of the leaves, for both a complete binary tree and a degenerate binary tree with 8 leaves.

The resulting distribution of $TP + FN$ values (Table 1, right) is exactly the same in both cases and thus, a complete binary tree and a degenerate binary tree with the same number of leaves have the same balance ratio. In fact, any two reference taxonomies for the same taxa have the same balance ratio as long as they have the same number of internal nodes, because they yield a metagenomic classification problem for any subset of the leaves and for any candidate internal node, and $TP + FN$ equals the number of leaves in the subset.

Let us assume that the reads in a metagenomic sample to be classified come from known sequences in a reference taxonomy, as it is usually the case in the taxonomic annotation of metagenomic samples, whereas reads coming from novel sequences are annotated by using clustering methods instead. Given a read and a set of candidate sequences in a reference taxonomy, the taxonomic annotation of the read at a certain node in the clade of the LCA in the reference taxonomy of the set of candidate sequences can then be taken to be correct if, and only if, the candidate sequence that the read comes from lies in the clade of the node at which it is annotated.

Based on this observation, we have studied the performance of some of the most often used indicators of classification error: the Yule ϕ [23], also known as Matthews correlation coefficient [17], the area under the ROC curve, the Youden J [22], the F-measure [18], the Jaccard similarity coefficient [13], and the Rand index [19], in the taxonomic annotation of metagenomic samples.

Definition 2. *Let TP, TN, FP, and FN be the number of true positives, false positives, true negatives, and false negatives in a binary classification problem.*

- *The Yule ϕ is given by*

$$\phi = \frac{TP\,TN - FP\,FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

- *The Youden J is given by*

$$J = \frac{TP\,TN - FP\,FN}{(TP + FN)(FP + TN)}$$

- *The area under the ROC curve is given by*

$$AUC = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{FP + TN}\right)$$

- *The F-measure is given by*

$$F = \frac{2\,TP}{2\,TP + FP + FN}$$

- *The Jaccard similarity coefficient is given by*

$$C = \frac{TP}{TP + FP + FN}$$

- *The Rand index is given by*

$$R = \frac{TP + TN}{TP + FP + TN + FN}$$

If the denominator in any of these formulas is zero, the value of the indicator is arbitrarily set to zero.

We have computed the value of all these indicators of classification error for each possible set of candidate sequences in a reference taxonomy and for each possible candidate node for the taxonomic annotation of a read coming from each of the candidate sequences, for different taxonomic reference topologies: complete binary trees, that have the largest possible balance but yield the least balanced metagenomic classification problems, and degenerate binary trees, that have the smallest possible balance but yield the most balanced metagenomic classification problems. For these classification problems, we have counted the number of times the taxonomic annotation is correct, that is, the number of times a read is annotated to a node in the reference taxonomy whose clade includes the reference sequence that the read comes from.

The results (Table 2) show that the worst indicator of classification error is the Yule ϕ, followed by AUC and the Youden J (which are equivalent, as $J = 2\,AUC - 1$), the F-measure and the Jaccard similarity coefficient C (which are also equivalent, as $C = F/(2 - F)$), and that the Rand index R is the best indicator of classification error for the taxonomic annotation of metagenomic samples. This can be explained by the fact that in a metagenomic classification problem, we focus on the correct classification of a correct taxonomic annotation while in a statistical classification problem in machine learning, where both positive and negative instances are taken into account, correlation measures such as the Yule ϕ (which is equivalent to the Pearson correlation coefficient for binary classification problems) often are the best indicators of classification error.

Table 2. Total number of correct taxonomic annotations under the Yule (ϕ), the area under the ROC curve (A) or the Youden J, the F-measure (F) or the Jaccard similarity coefficient, and the Rand index (R) for reads coming from known sequences, for different taxonomic reference topologies (complete binary tree and degenerate binary tree) with n leaves

Complete binary tree															
n	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
ϕ	4	14	40	70	262	306	824	1,450	4,318	6,156	17,064	28,158	63,378	118,292	270,448
A	4	14	40	70	262	306	920	1,530	4,726	6,316	22,056	29,528	79,322	138,477	352,496
F	4	12	32	78	220	407	984	2,234	5,188	10,251	24,844	49,019	112,812	235,322	493,856
R	4	12	48	90	344	485	1,544	2,742	8,308	11,845	37,764	54,757	154,012	239,147	672,416
Degenerate binary tree															
n	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
ϕ	4	14	38	80	203	388	945	1,961	4,344	8,592	20,152	39,474	88,063	183,603	398,700
A	4	14	38	80	211	384	973	1,952	4,628	8,346	22,230	38,088	94,962	188,986	421,697
F	4	12	32	79	195	441	1,024	2,270	5,104	10,994	24,491	51,959	113,305	241,277	518,937
R	4	12	36	89	222	512	1,191	2,652	5,949	12,971	28,459	61,189	132,263	281,547	602,076

Now, the taxonomic annotation of a metagenomic sample involves obtaining the candidate nodes in a reference taxonomy with the least classification error (for a given indicator) for each of the reads in the metagenomic sample. We have proved in [3] that, when the F-measure is taken as indicator, it suffices to consider candidate nodes that are either candidate sequences themselves, or the LCA of two or more candidate sequences in the reference taxonomy. That is, it suffices to consider as candidate nodes the LCA skeleton tree [7] of the set of candidate sequences for a given read.

We prove below that it also suffices to consider the LCA skeleton tree when the Youden J, the area under the ROC curve, or the Jaccard similarity coefficient is taken as indicator of classification error. The proof for the Yule ϕ is left to the reader.

Let T be a reference taxonomy, let M_i be the set of candidate sequences for the classification of read i, and let T_i be the subtree of T rooted at the LCA of M_i. See Fig. 2 for a schematic view.

Definition 3. *A node j in T_i is called* relevant *if it is equal to a candidate sequence in M_i or equal to the LCA of two or more candidate sequences in M_i.*

Also, for every node j in T_i, let $T_{i,j}$ be the subtree of T_i rooted at j, let L_i be the set of all candidate sequences in T_i, and let N_i be the set of all candidate sequences in T_i that do not belong to M_i (hence, $L_i = M_i \cup N_i$). Similarly, let $M_{i,j}$ be the set of all candidate sequences in $T_{i,j}$ that belong to M_i, let $N_{i,j}$ be the set of all candidate sequences in $T_{i,j}$ that do not belong to $M_{i,j}$, and let $L_{i,j} = M_{i,j} \cup N_{i,j}$. Using this notation, for the taxonomic annotation at node j of a read i with candidate sequences M_i (see Fig. 2), the true positives are $TP_{i,j} = M_{i,j}$, the false positives are $FP_{i,j} = N_{i,j}$, the true negatives are $TN_{i,j} = N_i \setminus N_{i,j}$, and the false negatives are $FN_{i,j} = M_i \setminus M_{i,j}$. Let $C_{i,j}$ be the Jaccard correlation coefficient for node j in T_i, that is,

$C_{i,j} = TP_{i,j}/(TP_{i,j} + FP_{i,j} + FN_{i,j})$. Similarly, let $J_{i,j}$ and $A_{i,j}$, and $F_{i,j}$ be the Youden J and the area under the ROC curve for node j in T_i, respectively. We have:

Theorem 1. *For each node j in T_i, there exists a relevant node j' such that $J_{i,j'} \geqslant J_{i,j}$, $A_{i,j'} \geqslant A_{i,j}$, and $C_{i,j'} \geqslant C_{i,j}$.*

Proof. Suppose that j is a node in T_i that is not relevant. Let j' be the LCA of the candidate sequences in $M_{i,j}$. Clearly, j' is relevant and, furthermore, $|M_{i,j}| = |M_{i,j'}|$ while $|N_{i,j}| \geqslant |N_{i,j'}|$ since $T_{i,j'}$ is a subtree of $T_{i,j}$.

Let $TP = |M_{i,j}|$, $FP = |N_{i,j}|$, $FN = |M_i| - |M_{i,j}|$, $TN = |N_i| - |N_{i,j}|$ and, similarly, let $TP' = |M_{i,j'}|$, $FP' = |N_{i,j'}|$, $FN' = |M_i| - |M_{i,j'}|$, $TN' = |N_i| - |N_{i,j'}|$. We have that $TP' = TP$, $FP' \leqslant FP$, $FN' = FN$, $TN' \geqslant TN$, and $TN' + FP' = TN + FP$.

- Youden J: It has to be proved that

$$\frac{TP'\,TN' - FP'\,FN'}{(TP' + FN')(FP' + TN')} \geqslant \frac{TP\,TN - FP\,FN}{(TP + FN)(FP + TN)}$$

  We have that $(TP' + FN')(FP' + TN') = (TP + FN)(FP + TN)$. Then, it suffices to prove that $TP'\,TN' - FP'\,FN' \geqslant TP\,TN - FP\,FN$, that is, $TP(TN' - TN) \geqslant FN(FP' - FP)$. But $TP \geqslant 0$, $(TN' - TN) \geqslant 0$, $FN \geqslant 0$, $(FP' - FP) \leqslant 0$ and thus, the inequality follows.
- Area under the ROC curve: It has to be proved that

$$\frac{TP'(FP' + TN') + TN'(TP' + FN')}{(TP' + FN')(FP' + TN')} \geqslant \frac{TP(FP + TN) + TN(TP + FN)}{(TP + FN)(FP + TN)}$$

  We have that $(TP' + FN')(FP' + TN') = (TP + FN)(FP + TN)$ and $TP'(FP' + TN') = TP(FP + TN)$. Then, it suffices to prove that $TN'(TP' + FN') \geqslant TN(TP + FN)$. But $TP' = TP$, $FN' = FN$, $TN' \geqslant TN$ and thus, the inequality follows.
- Jaccard similarity coefficient: It has to be proved that

$$\frac{TP'}{TP' + FP' + FN'} \geqslant \frac{TP}{TP + FP + FN}$$

  We have that $TP' = TP$, $FP' \leqslant FP$, $FN' = FN$ and thus, the inequality follows.
- Rand index: It has to be proved that

$$\frac{TP' + TN'}{TP' + FP' + TN' + FN'} \geqslant \frac{TP + TN}{TP + FP + TN + FN}$$

  We have that $TP' = TP$, $FN' = FN$, $TN' \geqslant TN$, $FP' + TN' = FP + TN$ and thus, the inequality follows.

□

Corollary 1. *The Youden $J_{i,j}$, the area under the ROC curve $A_{i,j}$, the Jaccard correlation coefficient $C_{i,j}$ and the Rand index $R_{i,j}$ only need to be computed for nodes j in T_i that are relevant.*

3 A Set Cover Approach to Taxonomic Annotation

Let us recall from [10] that an instance of the set cover problem is a collection C of subsets of a finite set X whose union is X, and a solution to the set cover problem is a subset $C' \subseteq C$ such that every element in X belongs to at least one member of C'. The set cover problem is NP-complete, but a logarithmic approximation can be computed in linear time [2,14].

Recall also that in a metagenomic classification problem, the are often multiple candidate nodes in a reference taxonomy with the least classification error for a given read. As a set cover problem, the set of elements X is the set of candidate nodes in a reference taxonomy with the least classification error for the reads in a metagenomic sample, and the collection C of subsets of X is the collection of sets of candidate nodes in the reference taxonomy with the least classification error for each read.

The following example is adapted from [4, Sect. 35.3]; see Fig. 3.

Example 1. Consider a metagenomic sample with reads $x_1, \ldots, x_{12}$ and candidate nodes in a reference taxonomy with the least classification error as follows: $\{y_1, y_3\}$ for x_1, $\{y_1, y_4\}$ for x_2, $\{y_1, y_5\}$ for x_3, $\{y_1, y_3\}$ for x_4, $\{y_1, y_2, y_4\}$ for x_5, $\{y_1, y_2, y_5\}$ for x_6, $\{y_3, y_4\}$ for x_7, $\{y_2, y_4\}$ for x_8, $\{y_2, y_5\}$ for x_9, $\{y_3, y_6\}$ for x_{10}, $\{y_4, y_6\}$ for x_{11}, and $\{y_5\}$ for x_{12}. Then, as an instance of the set cover problem, $X = \{x_1, \ldots, x_{12}\}$ and $C = \{y_1 \ldots, y_6\}$, where $y_1 = \{x_1, x_2, x_3, x_4, x_5, x_6\}$, $y_2 = \{x_5, x_6, x_8, x_9\}$, $y_3 = \{x_1, x_4, x_7, x_{10}\}$, $y_4 = \{x_2, x_5, x_7, x_8, x_{11}\}$, $y_5 = \{x_3, x_6, x_9, x_{12}\}$, and $y_6 = \{x_{10}, x_{11}\}$.

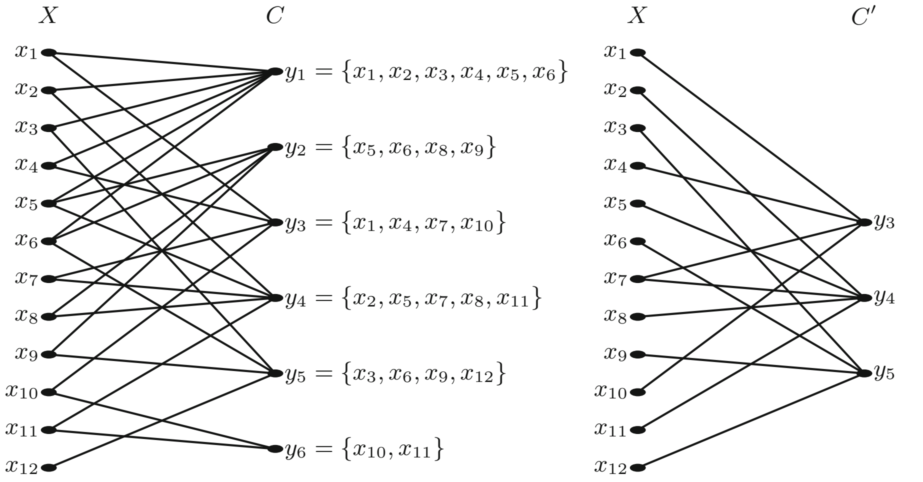

Fig. 3. (left) A metagenomic classification problem viewed as a set cover problem. X is the set of reads from a metagenomic sample, and C is the collection of candidate nodes in the reference taxonomy with the least classification error for some read from the metagenomic sample. (right) The smallest solution to the set cover problem instance.

In a solution C' to a metagenomic classification problem viewed as a set cover problem (X, C), each read in X is annotated to a node in $C' \subseteq C$. Such a taxonomic annotation is not necessarily unique, and there may still be ambiguities in the classification of the metagenomic sample. For the problem instance from Example 1, the smallest solution is $\{y_3, y_4, y_5\}$, which implies the taxonomic annotation of reads x_1, x_4 and x_{10} to node y_3, reads x_2, x_5, x_8 and x_{11} to node y_4, reads x_3, x_6, x_9 and x_{12} to node y_5, and read x_7 to either node y_3 or node y_4 in the reference taxonomy. The greedy algorithm of [14] yields the approximate solutions $\{y_1, y_4, y_5, y_3\}$ and $\{y_1, y_4, y_5, y_6\}$.

The taxonomic annotation of a metagenomic sample can thus be seen as the reduction, and ideally the removal, of ambiguity in the identification of the reads in the metagenomic sample, where a read is ambiguous if it is annotated to more than one node in a reference taxonomy. Viewing the metagenomic classification problem as a set cover problem, an element of X is ambiguous if it belongs to more than one subset of the collection $C' \subseteq C$. The subsets of a set cover overlap on ambiguous elements.

Definition 4. *Let X be a finite set and let C be a collection of subsets of X whose union is X. The overlap of a set cover $C' \subseteq C$ is the total size of the subsets minus the size of X.*

Let the *size* of a set cover be the number of subsets of X that it contains, and let the *total size* of a set cover be the total size of the subsets of X that it contains. This corresponds to set cover problems I and II in [14]. It turns out that a set cover of smallest size does not necessarily have the least overlap, while a set cover of smallest total size always has the least overlap.

Proposition 1. *A set cover with the least number of subsets does not necessarily have the least overlap.*

Proof. Let $X = \{1, \ldots, n\}$ and assume, without loss of generality, that $n = 2k$ for $k \geqslant 3$. Let S be the following collection of subsets of X:

$$\{1, 2\}, \{3, 4\}, \ldots, \{n-1, n\}, \{1, \ldots, n-1\}, \{2, \ldots, n\}$$

The set cover $\{1, \ldots, n-1\}, \{2, \ldots, n\}$ has size 2, which is the smallest possible for S and X, and overlap n. The set cover $\{1, \ldots, n-1\}, \{n-1, n\}$ also has size 2, but it has overlap 1. Same for the set cover $\{1, 2\}, \{2, \ldots, n\}$, and S and X have no other set cover of size 2. However, the set cover $\{1, 2\}, \{3, 4\}, \ldots, \{n-1, n\}$ has size $n/2$ and overlap 0, which is the least possible overlap.

The following result follows directly from Definition 4.

Corollary 2. *A set cover with the least total size of subsets has the least overlap.*

Based on the solution of a set cover problem with the least total size of subsets, the abundance profile of a metagenomic sample is given by the proportion of reads mapped to each node in the set cover, adjusted by a uniform distribution of any still ambiguous reads among all the nodes in the set cover which they are mapped to.

We have implemented the set cover approach to taxonomic annotation in a next release of the TANGO software [1,3], which belongs in the BioMaS [9] and MetaShot [8] pipelines. The new implementation of TANGO consists of a Python script for taxonomic annotation using the NCBI Taxonomy [5,6], based on the ETE Toolkit [11], along with another Python script for resolving any remaining ambiguities by finding an approximate solution to a set cover problem with the least total size of subsets. While the first script processes the input metagenomic sample one-sequence-read-at-a-time, the second script processes the output of the first script for the whole set of reads, and produces both a taxonomic annotation of the reads and an abundance profile of the metagenomic sample.

4 Conclusion

We have addressed two potential sources of bias in the taxonomic annotation of metagenomic samples, which is usually done by first mapping the reads to the reference sequences and then, classifying each read at a node in the clade of the LCA of the candidate sequences in the reference taxonomy with the least classification error. On the one hand, we have shown that the reference taxonomy being balanced or imbalanced does not affect the balance of the metagenomic classification problem, and we also shown that the Rand index is a better indicator of classification error for metagenomic classification problems than the often used area under the ROC curve and F-measure. On the other hand, we have reduced the taxonomic annotation problem for a whole metagenomic sample to a set cover problem, for which a logarithmic approximation can be obtained in linear time, and we have shown that a solution to the set cover problem with the least total size of subsets minimizes the ambiguity in the taxonomic annotation of the reads in a metagenomic sample.

Future work includes extending the computation of balance ratio and total number of correct taxonomic annotations from Sect. 2 to the NCBI Taxonomy, taking ancestry relationships among the nodes in the reference taxonomy into account in the set cover formulation of the taxonomic annotation problem from Sect. 3 and last, but not least, extending the set cover problem formulation of the taxonomic annotation problem to a non-taxonomic metagenomic classification problem, with reference sequences but without a reference taxonomy.

Acknowledgements. Partially supported by Spanish Ministry of Economy and Competitiveness and European Regional Development Fund project DPI2015-67082-P (MINECO/FEDER).

References

1. Alonso, D., Barré, A., Beretta, S., Bonizzoni, P., Nikolski, M., Valiente, G.: Further steps in TANGO: improved taxonomic assignment in metagenomics. Bioinformatics **30**(1), 17–23 (2013)
2. Bar-Yehuda, R., Even, S.: A linear-time approximation algorithm for the weighted vertex cover problem. J. Algorithms **2**(2), 198–203 (1981)
3. Clemente, J.C., Jansson, J., Valiente, G.: Flexible taxonomic assignment of ambiguous sequencing reads. BMC Bioinform. **12**(1), 8 (2011)
4. Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C.: Introduction to Algorithms, 3rd edn. MIT Press, Cambridge (2009)
5. Federhen, S.: The NCBI taxonomy database. Nucleic Acids Res. **40**(D1), D136–D143 (2012)
6. Federhen, S.: Type material in the NCBI taxonomy database. Nucleic Acids Res. **43**(D1), D1086–D1098 (2015)
7. Fischer, J., Huson, D.H.: New common ancestor problems in trees and directed acyclic graphs. Inform. Process. Lett. **110**(8–9), 331–335 (2010)
8. Fosso, B., Santamaria, M., D'Antonio, M., Lovero, D., Corrado, G., Vizza, E., Passero, N., Garbuglia, A.R., Capobianchi, M.R., Crescenzi, M., Valiente, G., Pesole, G.: MetaShot: An accurate workflow for taxon classification of host-associated microbiome from shotgun metagenomic data. Bioinformatics (2017, in press)
9. Fosso, B., Santamaria, M., Marzano, M., Alonso, D., Valiente, G., Donvito, G., Monaco, A., Notarangelo, P., Pesole, G.: BioMaS: a modular pipeline for bioinformatic analysis of metagenomic amplicons. BMC Bioinform. **16**(1), 203 (2015)
10. Garey, M.R., Johnson, D.S.: Computers and Intractability: A Guide to NP-Completeness. Freeman, Dallas (1979)
11. Huerta-Cepas, J., Serra, F., Bork, P.: ETE 3: reconstruction, analysis and visualization of phylogenomic data. Mol. Biol. Evol. **33**(6), 1635–1638 (2016)
12. Huson, D.H., Auch, A., Qi, J., Schuster, S.C.: MEGAN analysis of metagenomic data. Genome Res. **17**(3), 377–386 (2007)
13. Jaccard, P.: Étude comparative de la distribution florale dans une portion des Alpes et du Jura. Bull. Soc. Vaud. Sc. Nat. **37**(142), 547–579 (1901)
14. Johnson, D.S.: Approximation algorithms for combinatorial problems. J. Comput. Syst. Sci. **9**(3), 256–278 (1974)
15. Kunin, V., Copeland, A., Lapidus, A., Mavromatis, K., Hugenholtz, P.: A bioinformatician's guide to metagenomics. Microbiol. Mol. Biol. Rev. **72**(4), 557–578 (2008)
16. López, V., Fernández, A., García, S., Palade, V., Herrera, F.: An insight into classification with imbalanced data: empirical results and current trends on using data intrinsic characteristics. Inform. Sci. **250**(1), 113–141 (2013)
17. Matthews, B.W.: Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim. Biophys. Acta **405**(2), 442–451 (1975)
18. Powers, D.M.W.: Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. J. Mach. Learn. Tech. **2**(1), 37–63 (2011)
19. Rand, W.M.: Objective criteria for the evaluation of clustering methods. J. Am. Stat. Assoc. **66**(336), 846–850 (1971)
20. Thomas, T., Gilbert, J., Meyer, F.: Metagenomics: a guide from sampling to data analysis. Microb. Inform. Exp. **2**(1), 3 (2012)
21. Wooley, J.C., Godzik, A., Friedberg, I.: A primer on metagenomics. PLoS Comput. Biol. **6**(2), e1000667 (2010)
22. Youden, W.J.: Index for rating diagnostic tests. Cancer **3**(1), 32–35 (1950)
23. Yule, G.U.: On the methods of measuring association between two attributes. J. R. Statist. Soc. **75**(6), 579–642 (1912)