A Set Cover Approach to Taxonomic Annotation

Francesc Rosselló Gabriel Valiente

Department of Mathematics and Computer Science
Research Institute of Health Science, University of the Balearic Islands
Palma de Mallorca, Spain

Algorithms, Bioinformatics, Complexity and Formal Methods Research Group
Technical University of Catalonia
Barcelona, Spain

LSD & LAW 2018, London, UK, 8–9 February 2018

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 the presence of multiple nodes in the taxonomy with the least classification error for a given read. In this talk, we reduce 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 an exact solution can be obtained by integer linear programming.

Metagenomic Samples

Annotation of Metagenomic Samples

Taxonomic Annotation of Metagenomic Samples

Taxonomic Annotation of Metagenomic Samples via LCA

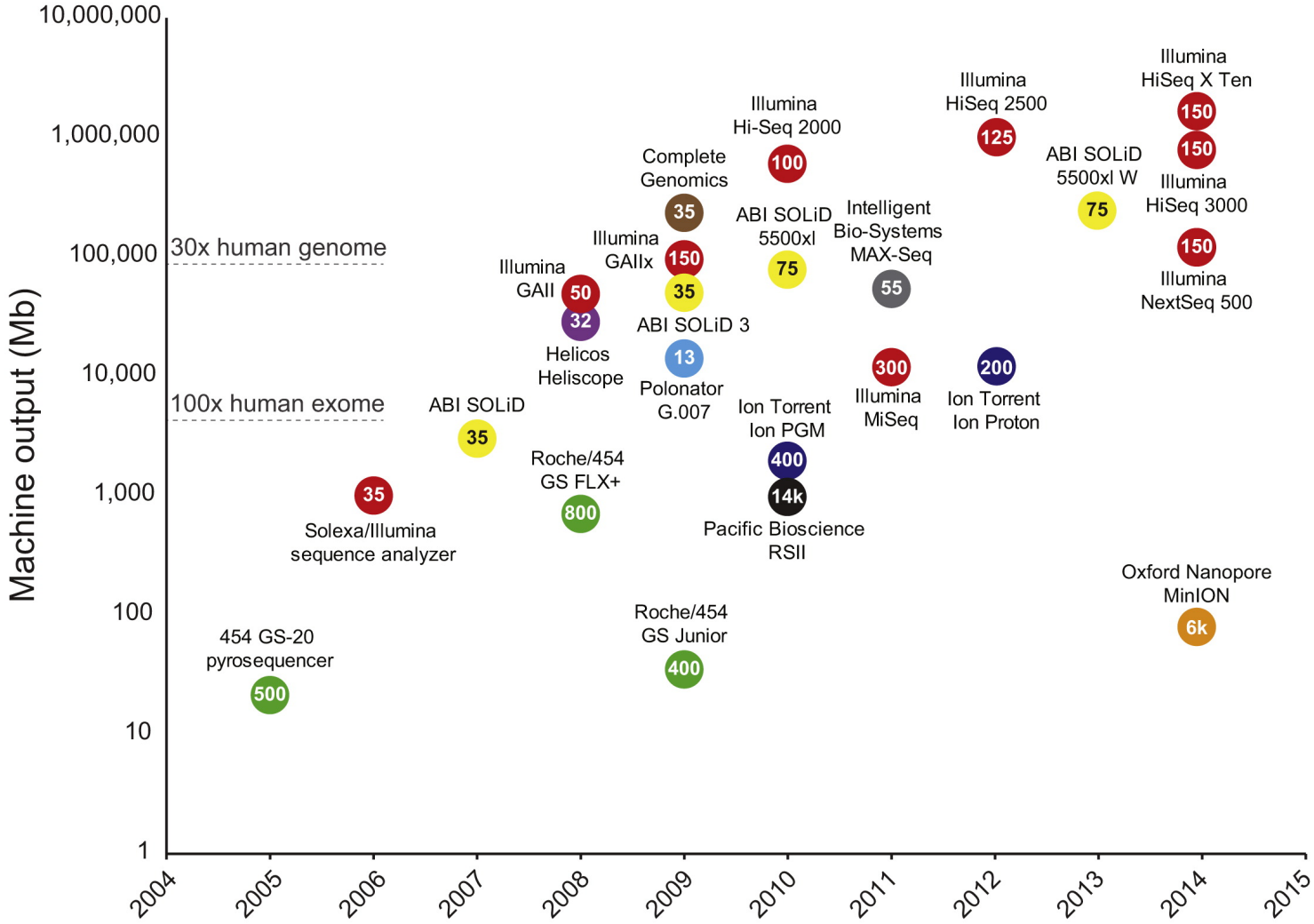
Taxonomic Annotation of Metagenomic Samples via Set Cover

Annotation of Metagenomic Samples via Set Cover

LP Formulation of the Set Cover Approach

Experimental Results


gaccctttcg
agccgcctgaaatg
tcccaaagcacaaaatt
gaatcgaatatagggctta
ggagacgacaccgagatcg
aatccatgctgataagga
tatgaaagccagcggag
tcgccaacatcgaaa
ccatcgtggtggc
ttcaaaaagtt

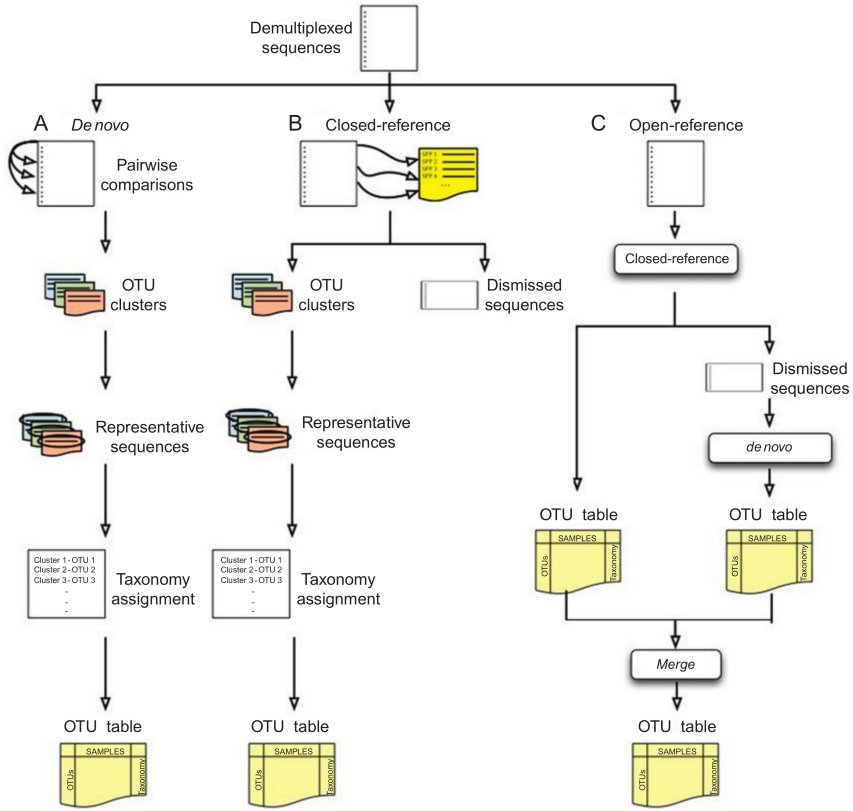


- J. A. Reuter, D. V. Spacek, and M. P. Snyder. High-throughput sequencing technologies. *Mol. Cell*, 58(4):586–597, 2015

- 16S ribosomal RNA sequencing is a common amplicon sequencing method used to identify and compare bacteria present in a given metagenomic sample
- Shotgun metagenomic sequencing allows sampling all genes in all organisms present in a given metagenomic sample

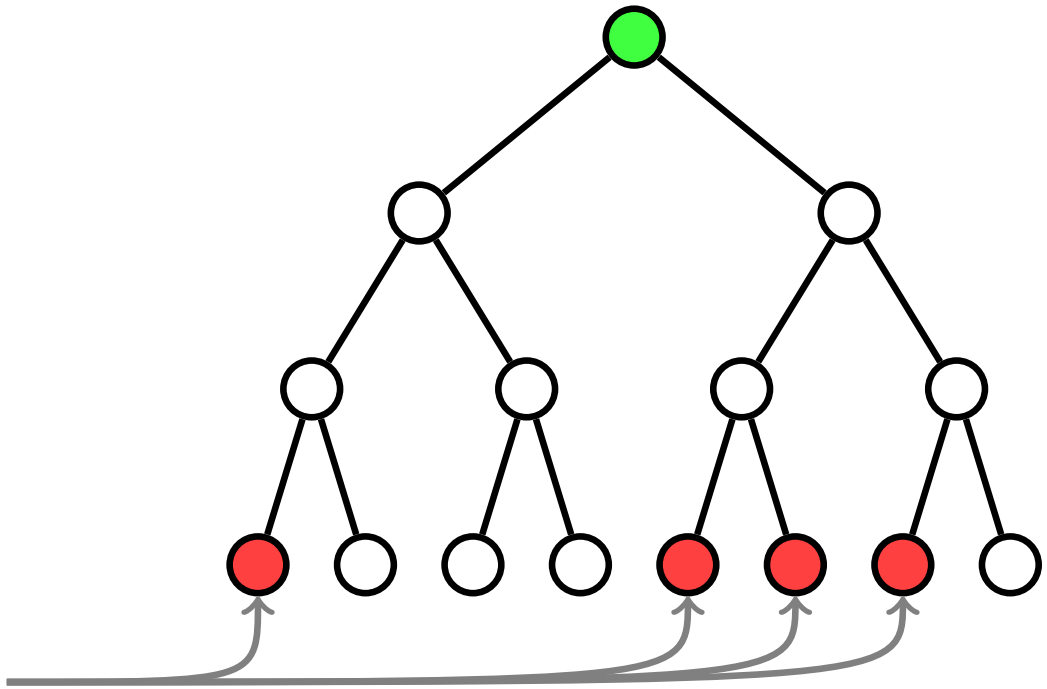
- Pattern matching problem: Map reads to reference genome
- Metagenomics: Multiple reference genomes

- The combined length of the reads can be much larger than the length of the reference genome



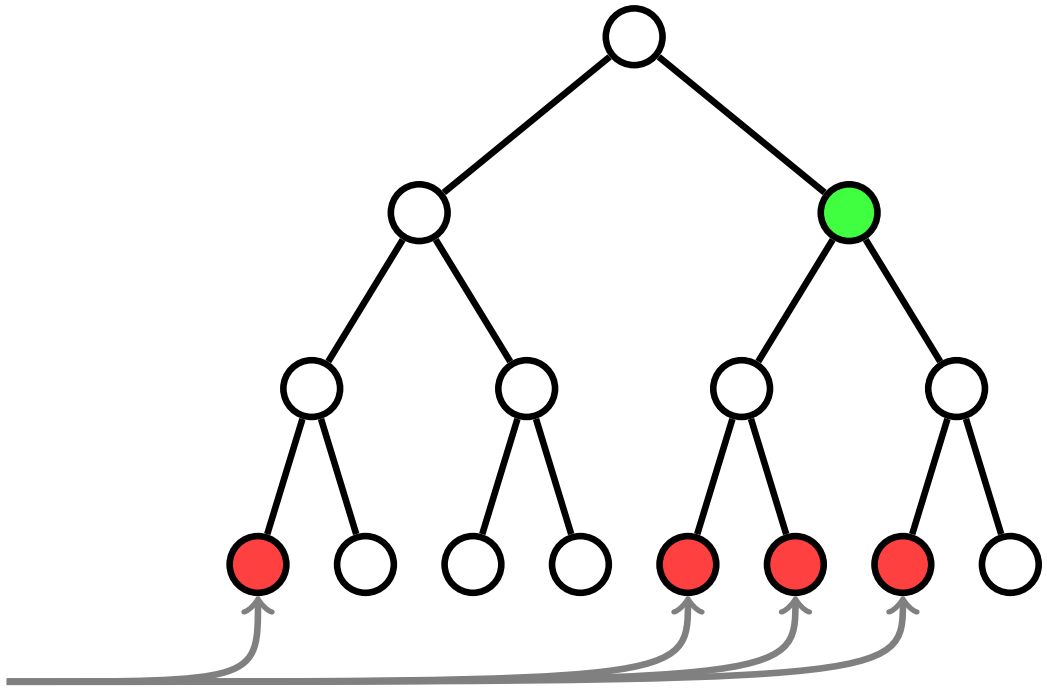
- J. A. Navas-Molina, J. M. Peralta-Sánchez, A. González, P. J. McMurdie, Y. Vázquez-Baeza, Z. Xu, L. K. Ursell, C. Lauber, H. Zhou, S. J. Song, J. Huntley, G. L. Ackermann, D. Berg-Lyons, S. Holmes, J. G. Caporaso, and R. Knight. Advancing our understanding of the human microbiome using QIIME. In E. F. Delong, editor, *Methods in Enzymology*, volume 531, chapter 19, pages 371–444. Elsevier, 2013

ACGTACGT ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGT ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGT ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGT ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGT ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT



- D. Huson and N. Weber. Microbial community analysis using MEGAN. In E. F. Delong, editor, *Methods in Enzymology*, volume 531, chapter 21, pages 465–485. Elsevier, 2013

ACGTACGT ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGT ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGT ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGT ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGT ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT



- J. C. Clemente, J. Jansson, and G. Valiente. Flexible taxonomic assignment of ambiguous sequencing reads. *BMC Bioinformatics*, 12:8, 2011

Metagenomic Samples

Annotation of Metagenomic Samples

Taxonomic Annotation of Metagenomic Samples

Taxonomic Annotation of Metagenomic Samples via LCA

Taxonomic Annotation of Metagenomic Samples via Set Cover

Annotation of Metagenomic Samples via Set Cover

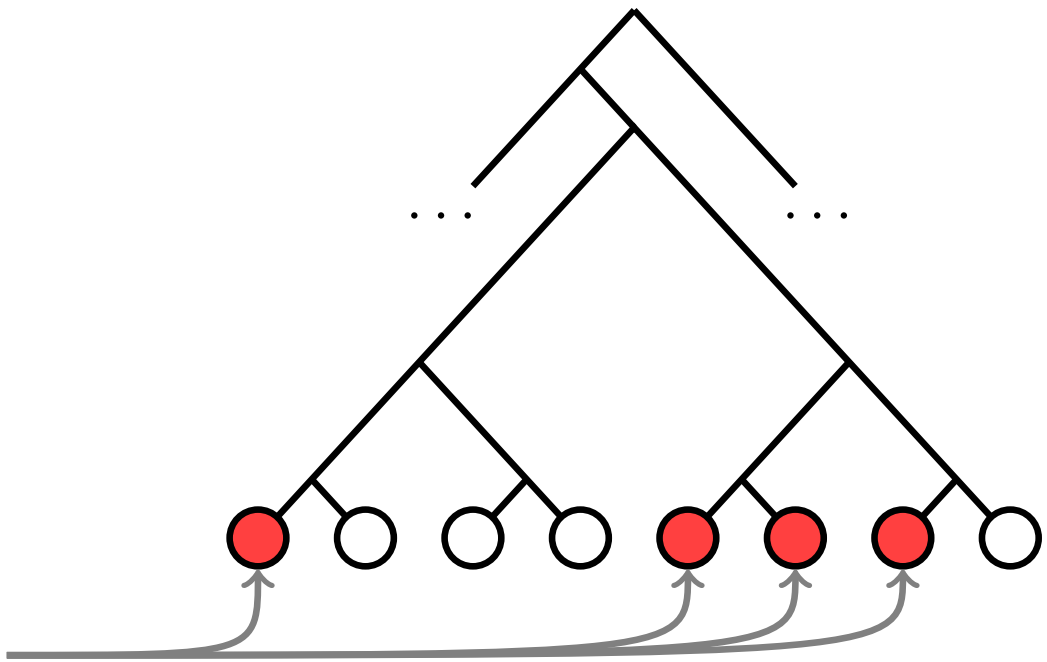
LP Formulation of the Set Cover Approach

Experimental Results

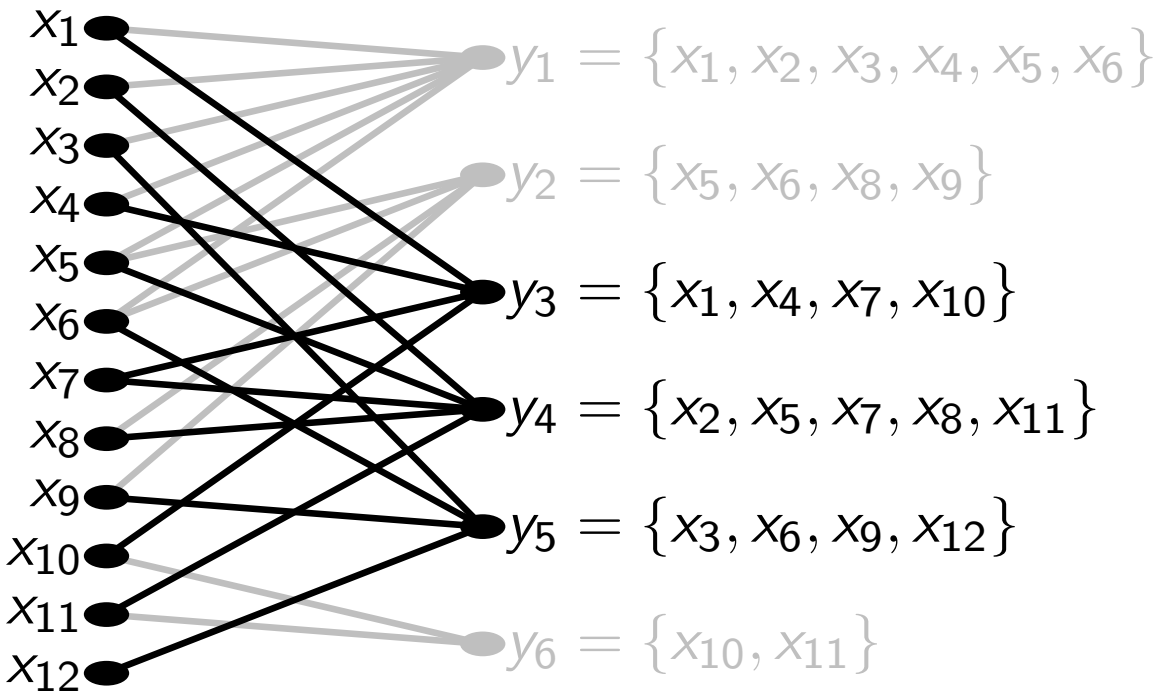
- An instance of the set cover problem is a collection C of subsets of a finite set X whose union is X
- A solution to the set cover problem is a smallest subset $C' \subseteq C$ such that every element in X belongs to at least one member of C'

- The set of elements X is the set of reads in the metagenomic sample
- The collection C of subsets of X is the set of candidate nodes in the reference taxonomy with the least classification error for the reads
- Each read in X is annotated to a candidate node in a solution $C' \subseteq C$

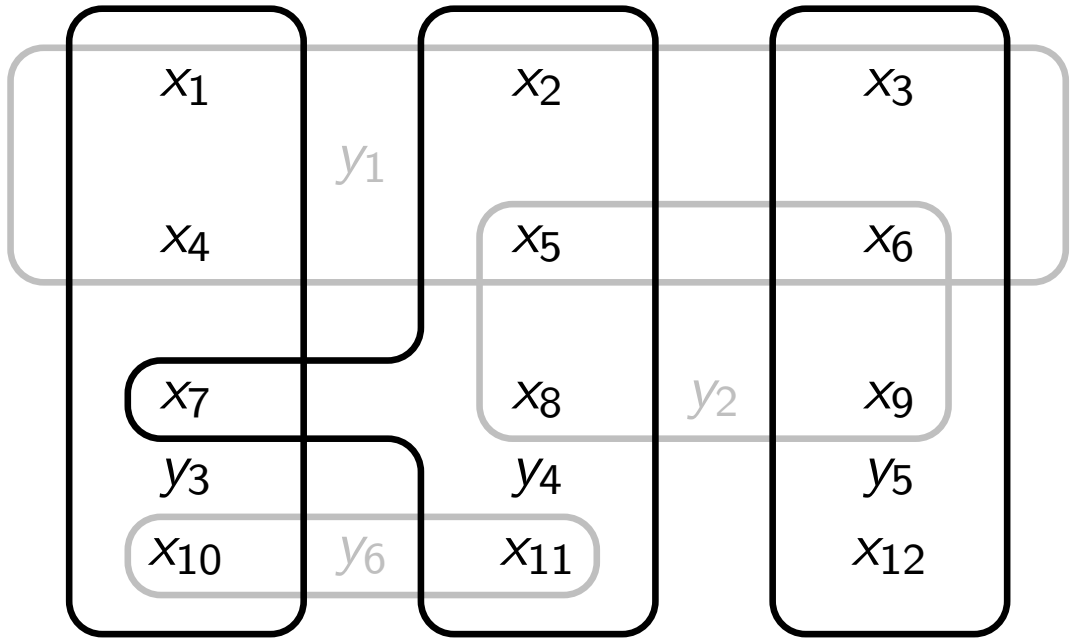
ACGTACGT ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGT ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGT ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGT ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGT ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT



ACGTACGT ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGT ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGT ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGT ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGT ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT



ACGTACGT ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGT ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGT ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGT ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGT ACGTACGTACGTACGTACGTACGTACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT
ACGTACGT



- An instance of the set cover problem is a collection C of subsets of a finite set X whose union is X
- A solution to the set cover problem is a smallest subset $C' \subseteq C$ such that every element in X belongs to at least one member of C'

- The set of elements X is the set of reads in the metagenomic sample
- The collection C of subsets of X is the set of candidate sequences for the reads
- Each read in X is annotated to a candidate sequence in a solution $C' \subseteq C$

- 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

- A set cover with the least number of subsets does not necessarily have the least overlap
- A set cover with the least total size of subsets has the least overlap

	y_1	y_2	y_3	y_4	y_5	y_6
x_1	✓		✓			
x_2	✓			✓		
x_3	✓				✓	
x_4	✓		✓			
x_5	✓	✓		✓		
x_6	✓	✓			✓	
x_7			✓	✓		
x_8		✓		✓		
x_9		✓			✓	
x_{10}			✓			✓
x_{11}				✓		✓
x_{12}					✓	
	22.2%	13.9%	16.7%	19.4%	19.4%	8.3%

	y_1	y_2	y_3	y_4	y_5	y_6
x_1	✓		✓			
x_2	✓			✓		
x_3	✓				✓	
x_4	✓		✓			
x_5	✓	✓		✓		
x_6	✓	✓			✓	
x_7			✓	✓		
x_8		✓		✓		
x_9		✓			✓	
x_{10}			✓			✓
x_{11}				✓		✓
x_{12}					✓	
	25.0%		20.8%	29.2%	25.0%	

	y_1	y_2	y_3	y_4	y_5	y_6
x_1	✓		✓			
x_2	✓			✓		
x_3	✓				✓	
x_4	✓		✓			
x_5	✓	✓		✓		
x_6	✓	✓			✓	
x_7			✓	✓		
x_8		✓		✓		
x_9		✓			✓	
x_{10}			✓			✓
x_{11}				✓		✓
x_{12}					✓	
	33.3%			29.2%	25.0%	12.5%

	y_1	y_2	y_3	y_4	y_5	y_6
x_1	✓		✓			
x_2	✓			✓		
x_3	✓				✓	
x_4	✓		✓			
x_5	✓	✓		✓		
x_6	✓	✓			✓	
x_7			✓	✓		
x_8		✓		✓		
x_9		✓			✓	
x_{10}			✓			✓
x_{11}				✓		✓
x_{12}					✓	
			29.2%	37.5%	33.3%	

- $X = \{x_1, x_2, \ldots, x_{12}\}$ (reads)
- $Y = \{y_1, y_2, \ldots, y_6\}$ (candidate nodes or sequences) where
  - $y_1 = \{x_1, x_2, x_3, x_4, x_5, x_6\}$
  - $y_2 = \{x_5, x_6, x_8, x_9\}$
  - $y_3 = \{x_1, x_4, x_7, x_{10}\}$
  - $y_4 = \{x_2, x_5, x_7, x_8, x_{11}\}$
  - $y_5 = \{x_3, x_6, x_9, x_{12}\}$
  - $y_6 = \{x_{10}, x_{11}\}$
- Minimize $\sum_j n_j y_j$
- Subject to $\sum_j a_{ij} y_j \geqslant 1$ for all i

  and $y_j \geqslant 0$ for all j

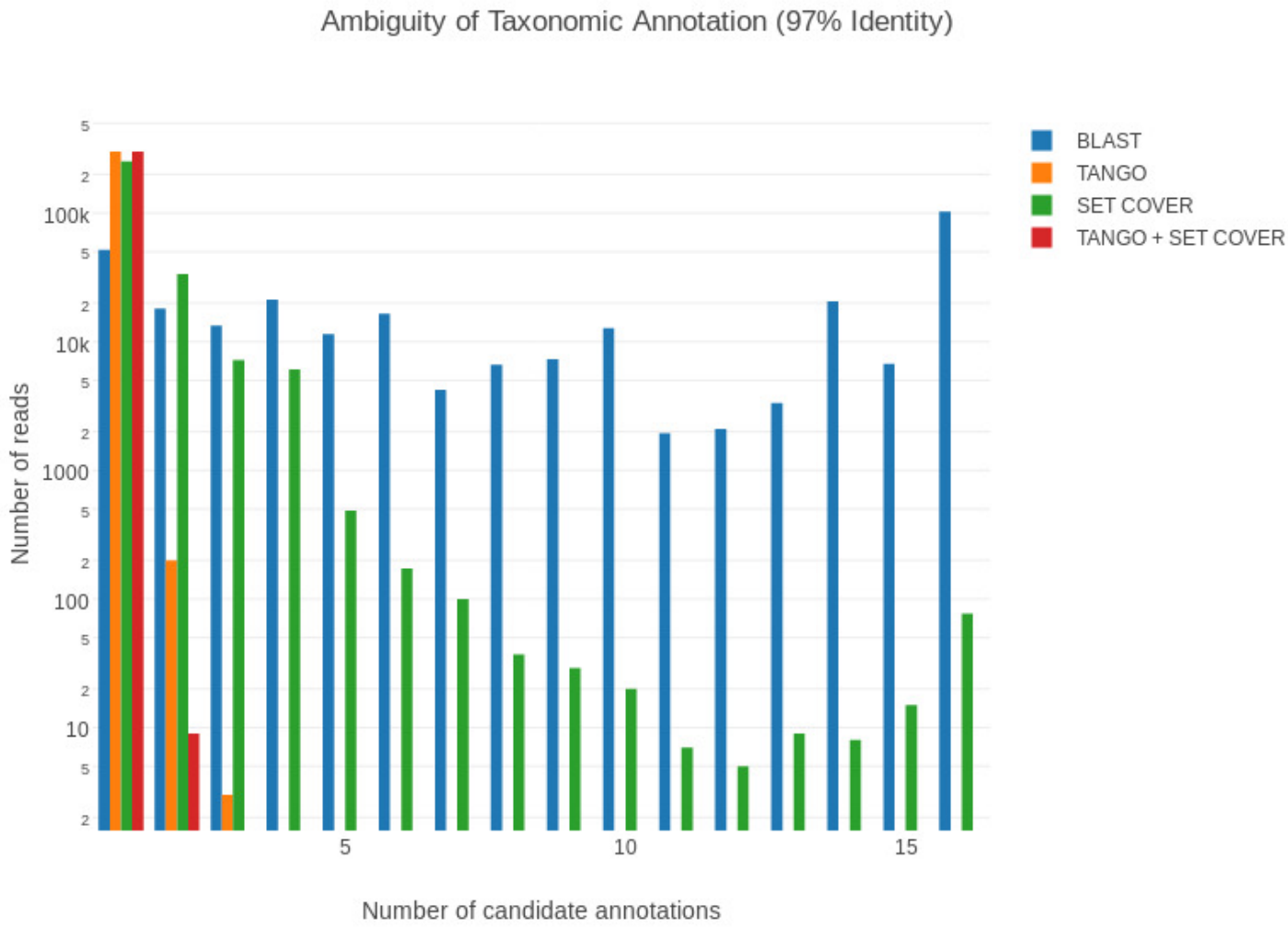
  and $y_j \leqslant 1$ for all j

a_{ij}	y_1	y_2	y_3	y_4	y_5	y_6	m_i
x_1	1	0	1	0	0	0	2
x_2	1	0	0	1	0	0	2
x_3	1	0	0	0	1	0	2
x_4	1	0	1	0	0	0	2
x_5	1	1	0	1	0	0	3
x_6	1	1	0	0	1	0	3
x_7	0	0	1	1	0	0	2
x_8	0	1	0	1	0	0	2
x_9	0	1	0	0	1	0	2
x_{10}	0	0	1	0	0	1	2
x_{11}	0	0	0	1	0	1	2
x_{12}	0	0	0	0	1	0	1
n_j	6	4	4	5	4	2	25

a_{ij}	y_1	y_2	y_3	y_4	y_5	y_6	m_i
x_1	1	0	1	0	0	0	2
x_2	1	0	0	1	0	0	2
x_3	1	0	0	0	1	0	2
x_4	1	0	1	0	0	0	2
x_5	1	1	0	1	0	0	3
x_6	1	1	0	0	1	0	3
x_7	0	0	1	1	0	0	2
x_8	0	1	0	1	0	0	2
x_9	0	1	0	0	1	0	2
x_{10}	0	0	1	0	0	1	2
x_{11}	0	0	0	1	0	1	2
x_{12}	0	0	0	0	1	0	1
n_j	6	4	4	5	4	2	25

- Subset of 302,581 reads of length 152 bp
- Aligned with BLAST to the 99,322 reference sequences of mean length 1,432 bp from Greengenes release 13.5 clustered at 97% identity
- The candidate annotations for a read are those reference sequences with the same E-value as the top hit
- Taxonomic annotation with TANGO
- Annotation with the set cover approach
- Taxonomic annotation with TANGO refined with the set cover approach

- J. G. Caporaso, C. L. Lauber, E. K. Costello, D. Berg-Lyons, A. Gonzalez, J. Stombaugh, D. Knights, P. Gajer, J. Ravel, N. Fierer, J. I. Gordon, and R. Knight. Moving pictures of the human microbiome. *Genome Biol.*, 12(5):R50, 2011



- B. Fosso, G. Pesole, F. Rosselló, and G. Valiente. Unbiased taxonomic annotation of metagenomic samples. *J. Comput. Biol.*, 2018. In press

Taxonomic rank	*BLAST*	*TANGO*	*Set cover*	*TANGO+*	*QIIME*
Archaea	0.013284	0.013284	0.013284	0.013284	0.015510
Crenarchaeota	0.013284	0.013284	0.013284	0.013284	0.015510
Bacteria	99.986716	99.986716	99.986716	99.986716	99.957297
Acidobacteria	0.074225	0.074391	0.074391	0.074391	0.036466
Actinobacteria	10.982357	10.982365	10.982365	10.982365	8.929160
Armatimonadetes	0.006642	0.006642	0.006642	0.006642	0.002070
Bacteroidetes	26.141444	26.141609	26.141443	26.141609	27.918210
Chloroflexi	0.091996	0.091993	0.091993	0.091993	0.018201
Cyanobacteria	2.564576	2.564843	2.564511	2.564843	1.989813
Deferribacteres	0.001328	0.001328	0.001328	0.001328	0.001742
Firmicutes	32.500312	32.463552	32.539437	32.463552	29.932524
Fusobacteria	3.802929	3.802929	3.802929	3.802929	4.529422
Gemmatimonadetes	0.029723	0.029889	0.029557	0.029889	0.001994
Planctomycetes	0.034207	0.034207	0.034207	0.034207	0.008272
Proteobacteria	21.029588	21.025871	21.028528	21.025871	25.774641
Spirochaetes	0.064096	0.064096	0.064096	0.064096	0.048609
Synergistetes	0.082141	0.119558	0.044834	0.119558	0.035557
Tenericutes	0.052810	0.052805	0.052805	0.052805	0.047571
Verrucomicrobia	2.395138	2.395138	2.395138	2.395138	0.601991
[Thermi]	0.085683	0.085683	0.085683	0.085683	0.054147
Other	0.047521	0.049817	0.046829	0.049817	0.026906
Unassigned	0.000000	0.000000	0.000000	0.000000	0.027193
Other	0.000000	0.000000	0.000000	0.000000	0.027193

All numbers are percentages.

- J. C. Clemente, J. Jansson, and G. Valiente. Flexible taxonomic assignment of ambiguous sequencing reads. *BMC Bioinformatics*, 12:8, 2011
- D. Alonso, A. Barré, S. Beretta, P. Bonizzoni, M. Nikolski, and G. Valiente. Further steps in TANGO: Improved taxonomic assignment in metagenomics. *Bioinformatics*, 30(1):17–23, 2014
- B. Fosso, G. Pesolo, F. Rosselló, and G. Valiente. Unbiased taxonomic annotation of metagenomic samples. In Z. Cai, O. Daescu, and M. Li, editors, *Proc. 13th Int. Symp. Bioinformatics Research and Applications*, volume 10330 of *Lecture Notes in Bioinformatics*, pages 162–173. Springer, 2017
- B. Fosso, G. Pesole, F. Rosselló, and G. Valiente. Unbiased taxonomic annotation of metagenomic samples. *J. Comput. Biol.*, 2018. In press