Microbial Ecology and Biogeography

— OF THE —

Southern Ocean

David Wilkins

Submitted in fulfillment of the requirements for the Degree of Doctor of Philosophy.

School of Biotechnology and Biomolecular Sciences
University of New South Wales, Sydney

March 2013

Contents

List of Figures

List of Tables

List of Acronyms

GAAS Genome relative Abundance and Average Size.

MEGAN Metagenome Analyzer.

IP Integer Programming.

ITS Internal Transcribed Spacer.

KEGG Kyoto Encyclopedia of Genes and Genomes.

LP Linear Programming.

OTU Operational Taxonomic Unit.

SO Southern Ocean.

Acknowledgements

MINSPEC, a bioinformatic tool for metagenomics

Sections of this chapter have been previously published in .

Summary

Introduction

Metagenomic analysis of microbial assemblages

The identification of the species or Operational Taxonomic Units (OTUs) that compose a microbial community is a primary aim of metagenomics. Typically this is achieved using one of two methods.

The first method is the identification, using a search and alignment algorithm such as BLAST, of specific marker genes or other sequences which are diagnostic for a particular OTU. Common targets in microbial ecology are the 16S or other ribosomal subunit rDNA sequences, and the Internal Transcribed Spacer (ITS) regions between 16S–23S rDNA sequences (e.g. **?**). This method provides several advantages. The selected regions are usually highly conserved, and through cultivation and full-genome sequencing have been reliably associated with a particular OTU, allowing very accurate identification and analysis of diversity down to the ecotype level (e.g. **?**). If the copy number of the gene or region is well known, this method also allows for accurate estimations of cell abundance from metagenomes. However, a disadvantage of this method is that the large majority of metagenomic reads will not cover the region of interest, and will contribute nothing to the analysis. Low-abundance OTUs will therefore be missed, as the region of interest is unlikely to have been sequenced.

The second method is to compare assembled or unassembled metagenomic reads to a reference database, using an algorithm such as BLAST, then use probabilistic methods to assign identifications and abundances with varying degrees of confidence. Most commonly, the reads are compared to a database of full genomes (e.g. **??**). This method makes more efficient use of metagenomic data compared to the first, as any read can potentially yield a BLAST match and thus contribute to the identification of an OTU. However, interpretation of the results, and particularly calculation of abundances, is more complex. For example, the software tool Genome relative Abundance and Average Size (GAAS) makes use of BLAST match quality, number of matches and estimated genome size to estimate the relative abundances of OTUs in a sample (**?**).

Such relative abundance estimates are confounded by the presence of multiple OTUs which can generate high-quality BLAST matches (“hits”) to a given read. Multiple high-quality hits to a single read are the norm, rather than the exception, in metagenomic studies for several reasons. A microbial assemblage will often include a number of closely-related OTUs (e.g. congeners) which share large sections of highly similar or identical genomic sequence. If several such OTUs are present in the reference database, a metagenomic read from one will yield high-quality BLAST hits to them all. Further, even distantly related OTUs are likely to share large regions of identity, and the selection of hit quality thresholds to discriminate between them (for example, a minimum bit score or maximum expectation value) is effectively arbitrary. Thus, while metagenomic studies using whole-genome comparisons almost always use such thresholds as the sole discriminators between OTUs, this method (hereafter the “naïve” method, after **?**) will almost inevitably result in the identification of OTUs which are not

Table 1: Selected examples of OTUs identified in a marine metagenome using the naïve method. These OTUs were identified in a single sample from the SO (sample 346; see "**??**"). The sample was compared to the RefSeq database of full genomes using TBLASTX with an E-value maximum of 1.0×10^{-3}, i.e. only high-quality hits were included. Relative abundances were calculated using GAAS (**?**).

Species	**Relative Abundance (%)**	**Notes**
Encephalomyocarditis virus	1.98	Human pathogen.
Marek's disease virus type 1	1.49	Chicken pathogen.
Marek's disease virus type 2	0.85	Chicken pathogen.
Francisella philomiragia	0.041	Human and animal pathogen.
Agrobacterium vitis	0.040	Plant and opportunistic human pathogen.
Brucella suis	0.011	Human and swine pathogen (causes brucellosis).
Enterobacter sp. 638	0.0085	Animal commensal/pathogen.
Bordetella parapertussis	0.0075	Mammalian pathogen (causes mild form of whooping cough).
Neisseria meningitidis	0.0074	Human pathogen.
Yersinia pestis	0.0060	Human/animal pathogen (causes bubonic plague).

present in the assemblage, skewing the relative abundance estimates of those which are truly present.

This problem is compounded by a systematic overrepresentation within full genome databases of of taxa of particular interest to humans, such as human and agricultural pathogens. Environmental OTUs are comparatively underrepresented. For example, Table 1 gives examples of terrestrial plant and animal pathogens, *a priori* unlikely to be truly present, which were identified in an open ocean metagenome with the naïve method.

One commonly used software tool to address this problem, Metagenome Analyzer (MEGAN), aggregates reads with hits to many OTUs to the most recent common ancestor of those OTUs, represented by a higher taxonomic rank e.g. family (**?**). This approach increases the fidelity of the results, but comes at the cost of reduced taxonomic resolution. Particularly in marine assemblages where even fine genomic differences can represent distinct ecological functions (e.g. **?**), a tool which reduces spurious identifications without compromising taxonomic resolution would clearly be valuable.

The maximum parsimony approach

? identified an analogous problem in the annotation of biochemical pathways in genomes and metagenomes. They noted that a common method is to annotate a pathway as present if a single protein within that pathway attracts at least one high-quality BLAST hit. However, because many proteins are shared by multiple pathways, and databases of orthologous genes are often incomplete, this method has resulted in many clearly spurious annotations, such as an ascorbic acid synthesis pathway in the human genome (humans require dietary vitamin C) and a mitochondrial pathway in *Escherichia coli* (annotated in the Kyoto Encyclopedia of Genes and Genomes (KEGG) PATHWAY database).

The authors developed a software tool, MinPath, to combat this problem and increase the accuracy and fidelity of pathway annotations. MinPath computes the smallest possible set of pathways ("maximum parsimony") sufficient to explain a set of annotated proteins. As a simple example, if a genome is annotated with all the proteins that belong to pathway A, and one of those proteins also happens to belong to pathway B — that is, it is shared by both pathways — the naïve approach would annotate both pathways as present. However, the most parsimonious explanation is that pathway A is present, and B is not.

MinPath was implemented by framing the construction of a maximum parsimony pathway set as an Integer Programming (IP) problem. IP is a subset of algorithms for solving Linear Programming (LP) problems, which seek to maximise the value of a linear function (the objective function) within a set of constraints. In this case, the objective function was maximised by decreasing the number of annotated biochemical pathways, while the constraint was that every high-quality protein annotation had to be represented at least once in the annotated pathways. Validation and testing of MinPath showed it was successful in eliminating spuriously annotated pathways while retaining those genuinely present.

It was noted that this as this problem is isomorphic with that of spurious annotations in microbial metagenomes, the "maximum parsimony" method would also be likely to work in the latter domain. The aim of the project described in this chapter was thus to develop and test a software tool, MINSPEC, which would find the most parsimonious set of OTUs necessary to explain a set of observed BLAST hits generated by a metagenome, using the approach of (**?**) as a model.

Methods

Implementation of MINSPEC

A computational method to minimise false OTU identifications and increase the accuracy of OTU abundance estimates (MINSPEC) was developed and implemented in PERL[1]. Following the approach of **?** to the parsimonious reconstruction of biochemical pathways (MINPATH), MINSPEC computes the smallest set of OTUs sufficient to explain a set of observed high-quality hits against RefSeq (or any other sequence database). The minimal set computation was framed as a IP problem and solved with GLPSOL (The GNU Linear Programming/MIP solver) (Free Software Foundation, Boston).

The objective function for the IP problem was constructed as follows (adapted from **?**):

$$\min \sum_{j=1}^{s} A_j$$

where s is the number of OTUs in the assemblage, and $A_j = 1$ if OTU j is in the assemblage, 0 if not. In other words, the objective function is satisfied by minimising the number of OTUs in the assemblage. The constraint function was constructed as follows (adapted from **?**):

$$\sum_{j=1}^{s} M_{ij} A_j \geq 1 \quad \forall i \in [1, n]$$

where $M_{ij} = 1$ if read i has a mapping (i.e. a high-quality BLAST hit) to OTU j, 0 if not, and $[1, n]$ is the set of all reads. In other words, the constraint function fails if any read does not have at least one of its high-quality BLAST hits represented in the assemblage.

This approach eliminates many of the spurious OTU identifications which result from reads with strong identity to more than one OTU. The "minimal OTU set" is liable to exclude some low-abundance OTUs, but gives more faithful abundance estimates and eliminates many false positives.

It was noted that in some special cases, it may be desirable to include an OTU in the assemblage even if it is not part of the minimal set, if that OTU generated a very large number of BLAST hits. An example of such a situation might be if the sample was known with certainty to contain a two very closely related OTUs at roughly equal abundance. In such a case, it would be expected that almost all metagenomic reads generated by each of these OTUs would also attract BLAST hits to the other, and MINSPEC would thus probably eliminate whichever happened to generate slightly fewer hits. To allow for this, an option was added such that MINSPEC will not eliminate OTUs to which a specified threshold number of reads attract high-quality hits.

Validation of MINSPEC

To establish the usefulness of MINSPEC, a validation method was devised to experimentally determine its error rates and efficacy (i.e. number of spurious OTUs identified and removed).

A set of simulated microbial OTUs was generated. To simulate genomic sequence identity between OTUs, each simulated OTU went through up to fifty rounds in which another OTU was selected at random and marked as having sequence identity with the first. This process was terminated with a 10% probability at each round, simulating an exponential curve of interrelatedness between OTUs. A

[1] MINSPEC and the associated metagenomic simulation and validation scripts are open source and available at `https://github.com/wilkox/minspec`.

random subset of the simulated OTUs were then selected to form a simulated microbial assemblage. Because of the previously established simulated sequence identity between OTUs, some OTUs in the assemblage would be marked as having identity to other OTUs both within the assemblage and outside of it.

A simulated metagenomic sampling was then performed. In each round, an OTU was selected at random. To produce a natural rank-abundance curve of OTU abundance within the assemblage, the probability that the selected OTU yielded a read was

$$\frac{1}{ln(x) + 1}$$

where x is the OTU's rank. Simulated BLAST matches to the OTU were generated for the read. These matches would include accurate high-quality "genuine" hits to the OTU that produced the read, as well as to other randomly selected OTUs both within and out of the assemblage which had been previously marked as having sequence identity to the "genuine" OTU.

To fully explore the limits and reliability of MINSPEC, the simulated metagenomic experiment described above was performed with all possible permutations of the following parameters: number of simulated OTUs [100; 1,000; 10,000; 50,000; 100,000]; size of simulated assemblage [1; 10; 100; 300; 500; 1,000; 10,000]; number of simulated metagenomic reads [10; 100; 1,000; 10,000; 100,000; 200,000; 500,000]. Each permutation was repeated five times, except for those where the size of the assemblage would exceed the number of OTUs simulated.

The resulting simulated BLAST outputs were processed with MINSPEC, and the false positive (percentage of OTUs not in the assemblage which nevertheless survived MINSPEC filtering) and false negative (percentage of OTUs present in the assemblage which were not present after MINSPEC filtering) rates calculated. Because a high false negative rate can arise from undersampling, a problem in metagenomic studies both real and simulated, an additional "false negative (MINSPEC)" metric was calculated, which excluded OTUs which were present in the assemblage but through random chance did not generate any reads, the equivalent of "unsampled rare taxa". This rate thus represented only false negatives attributable to MINSPEC itself. Finally, as a measure of MINSPEC's usefulness, the proportion of "false OTUs" — OTUs that generated BLAST matches but were not part of the assemblage — successfully removed by MINSPEC was calculated.

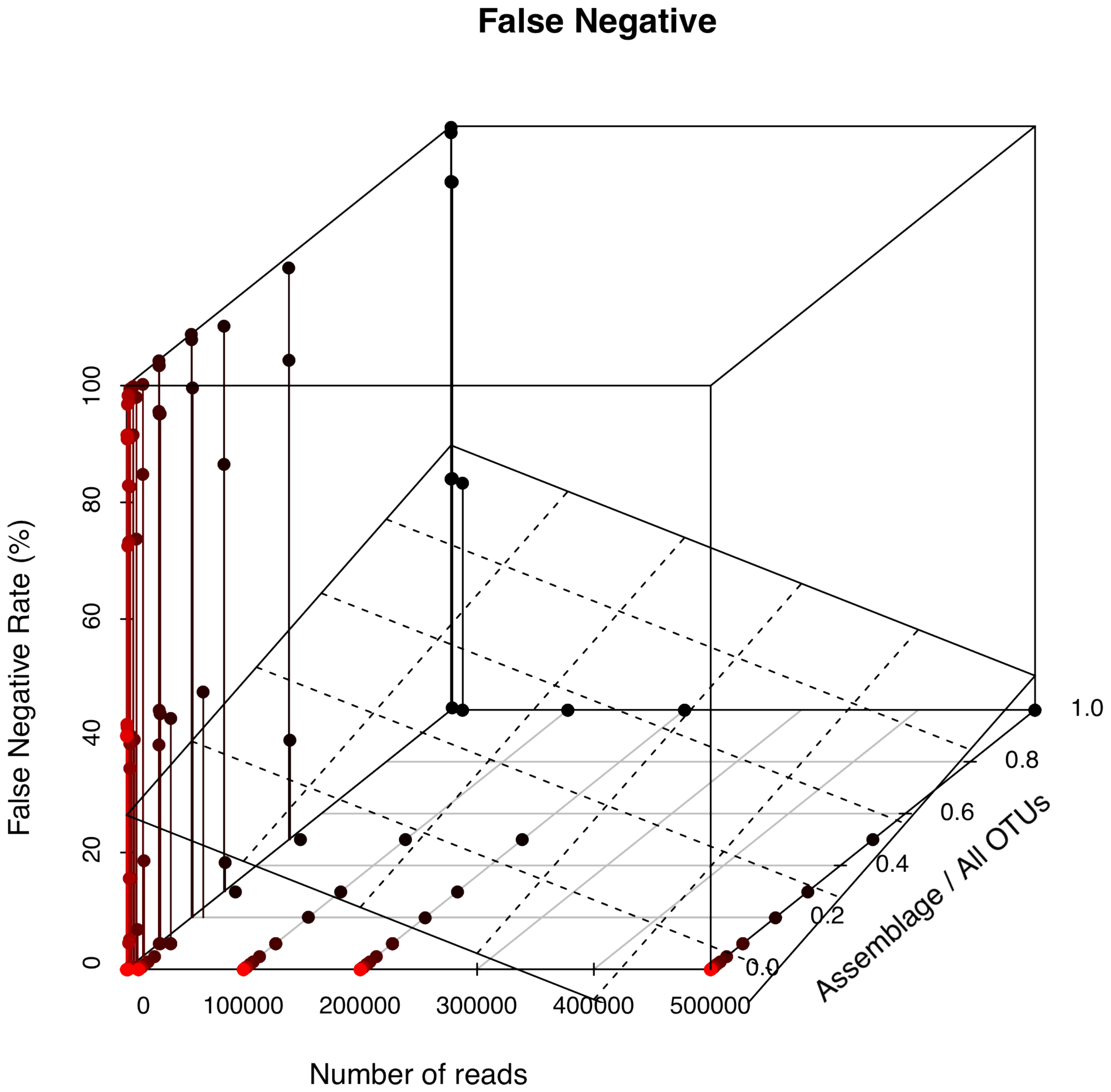
Results

Repeated simulated metagenomic experiments with a wide range of permutations of parameters showed that MINSPEC was reliable and able to substantially reduce the rate of false positive OTU identifications, although its effectiveness varied with the parameters of the assemblage and metagenomic experiment (**??**).

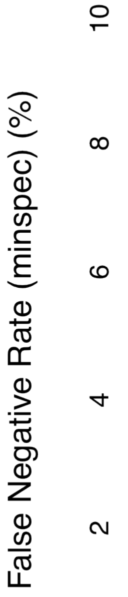
Discussion

The false negative rate, or percentage of OTUs in the simulated assemblage which were absent from the BLAST results following MINSPEC processing, was generally high, ranging from ∼20% under ideal conditions (a low assemblage / all OTUs ratio, and 500,000-read metagenomic sample) to ∼90% in the worst case (a high assemblage / all OTUs ratio and a small metagenomic sample) (**??**). The assemblage / all OTUs ratio (hereafter referred to as "assemblage ratio") indicates the proportion of simulated OTUs ("all OTUs") that were chosen to form the simulated assemblage. A higher ratio means that any OTU is more likely on average to be part of the assemblage, and thus that any individual failure to detect a OTU is an error. This problem is mitigated with increasing the number of reads, as this makes it less likely that a given OTU would go unsampled. The extreme false negative rates, in some cases 100%, represent extreme simulated scenarios (e.g. an assemblage of 1 OTU drawn from a pool of 100,000), and thus do not reflect real metagenomic studies.

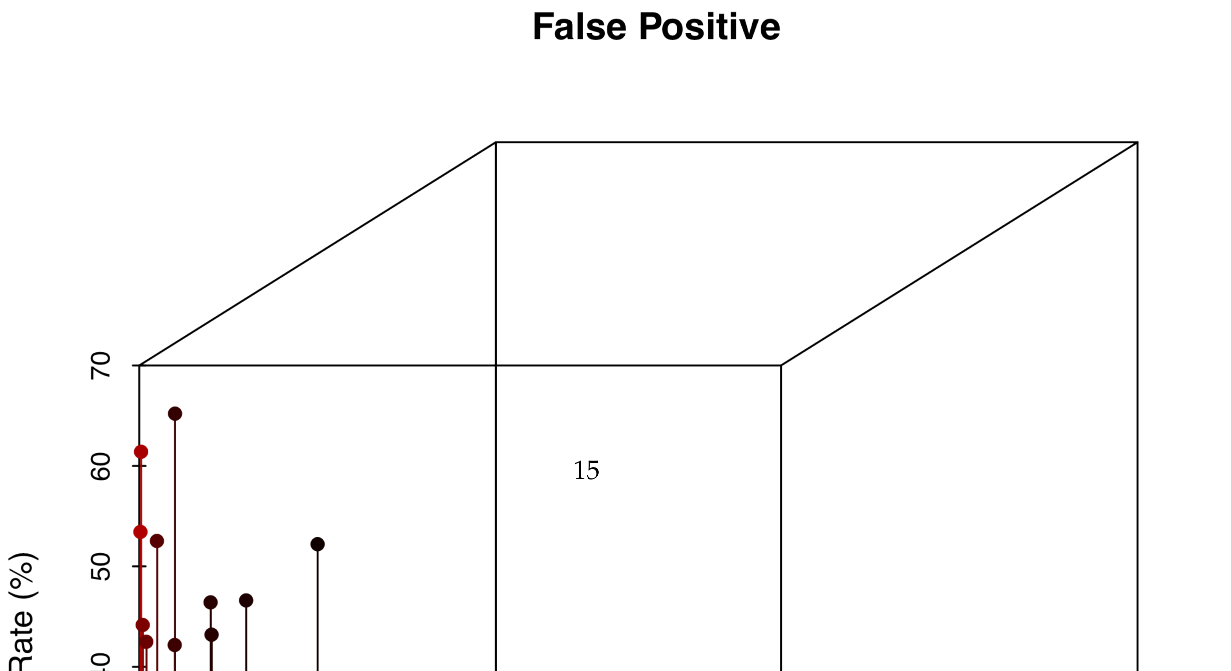
Because the majority of false negatives are attributable to undersampling and failure of OTUs to generate BLAST hits — properties the simulated metagenomic experiments share with real ones — a second metric, the false negative (MINSPEC) rate, was calculated (**??**). This is the proportion of OTUs

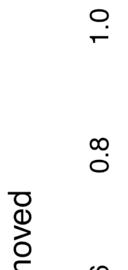


(a) False negative rate (%) — the percentage of OTUs in the assemblage that were absent from the BLAST results following MINSPEC processing.



(b) MINSPEC-a
were incorrec





in the assemblage that generated BLAST hits, but were incorrectly removed by MINSPEC. This rate thus represents error attributable only to MINSPEC. The false negative (MINSPEC) rate was generally low, ranging from ∼0–1% for low assemblage ratios, to ∼15–20% under high ratios. Surprisingly, increasing the number of reads only slightly decreased the rate, at both low and high assemblage ratios. This suggests MINSPEC is more affected by the degree of similarities between OTUs than by undersampling.

The false positive rate, or percentage of OTUs not in the assemblage which nevertheless generated high-quality BLAST matches that were not identified and removed by MINSPEC, was generally ∼0–5% except for extremely small read sets and low assemblage ratios, where it reached as high as 60% (**??**). These results reinforce the value of larger read sets, and show that once a modest metagenome size is reached (∼100,000 reads) very few false positives can be expected.

The proportion of false OTUs removed was calculated to measure MINSPEC's efficacy in identifying and eliminating OTU which are not part of the sampled assemblage yet generate high-quality BLAST matches. This rate varied from 0–1 depending on the parameters of the assemblage (**??**). For simulations with a low assemblage ratio, the proportion was generally high (> 0.6), although there were simulated experiments with a low ratio where the proportion was low or zero. However, in all simulations with an assemblage ratio of 1, the proportion was 0, and the regression indicated a generally inverse relationship between the ratio and the proportion of false OTUs removed. This is likely because in assemblages with a higher assemblage ratio, there are fewer false OTUs to remove; in assemblages with a ratio of 1, there are none. The high proportion of false OTUs correctly identified in simulations with a low assemblage ratio is thus a good indication that MINSPEC is effective at identifying and removing false OTUs, especially as this proportion far exceeds the false positive and false negative (MINSPEC) rates for comparable experiments. As expected, increasing the number of reads improved MINSPEC's accuracy.

Conclusions

Overall, the simulated experiments validated both the accuracy and usefulness of MINSPEC as a tool for reducing error in metagenomic studies. It is worth noting that the assemblage ratio is not an inherent property of an assemblage, although it is limited by the assemblage's OTU richness. Rather, it can be decreased, and thus the accuracy of the metagenomic experiment improved, by performing BLAST searches against larger databases with finer taxonomic resolution. These results thus reinforce the value of both large read sets and comprehensive reference databases in obtaining high-quality metagenomic results.

At the time of writing, MINSPEC has been used in two published projects: **?** and **?**.