Bioinformatics, 2016, 1–7
doi: 10.1093/bioinformatics/btw357
Advance Access Publication Date: 16 June 2016
Original Paper

OXFORD

Data and text mining

@MInter: automated text-mining of microbial interactions

Kun Ming Kenneth Lim[1,2], Chenhao Li[1,3], Kern Rei Chng[1] and Niranjan Nagarajan[1,3,*]

[1]Computational and Systems Biology, Genome Institute of Singapore, Singapore 138672, Singapore, [2]Computational Biology Program, Faculty of Science and [3]Department of Computer Science, National University of Singapore, Singapore, Singapore

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on March 13, 2016; revised on May 10, 2016; accepted on May 31, 2016

Abstract

Motivation: Microbial consortia are frequently defined by numerous interactions within the community that are key to understanding their function. While microbial interactions have been extensively studied experimentally, information regarding them is dispersed in the scientific literature. As manual collation is an infeasible option, automated data processing tools are needed to make this information easily accessible.

Results: We present @MInter, an automated information extraction system based on Support Vector Machines to analyze paper abstracts and infer microbial interactions. @MInter was trained and tested on a manually curated gold standard dataset of 735 species interactions and 3917 annotated abstracts, constructed as part of this study. Cross-validation analysis showed that @MInter was able to detect abstracts pertaining to one or more microbial interactions with high specificity (specificity = 95%, AUC = 0.97). Despite challenges in identifying specific microbial interactions in an abstract (interaction level recall = 95%, precision = 25%), @MInter was shown to reduce annotator workload 13-fold compared to alternate approaches. Applying @MInter to 175 bacterial species abundant on human skin, we identified a network of 357 literature-reported microbial interactions, demonstrating its utility for the study of microbial communities.

Availability and implementation: @MInter is freely available at https://github.com/CSB5/atminter.

Contact: nagarajann@gis.a-star.edu.sg

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

In its natural state, microbial life often does not form a homogenous population, but is instead present as heterogeneous communities composed of distinct species i.e. a *Microbiome*. Microbiomes are virtually ubiquitous amongst telluric environments; soil, banknotes and even macroscopic organisms are all environmental niches that play host to one or more microbial communities (Donaldson *et al.*, 2015; Jalali *et al.*, 2015; Zeglin *et al.*, 2015).

In recent years, microbial consortia resident in the human body have become a topic of particular interest. The human microbiome is now known to be intricately linked to overall human health (Guinane and Cotter, 2013). Microbiome dysfunction has been implicated in a number of disease conditions (Morgan *et al.*, 2012; Turnbaugh *et al.*, 2006; Weyrich *et al.*, 2015), while bacteria in a 'healthy' microbiome are known to act in a mutualistic fashion towards their host (Bergonzelli *et al.*, 2006; Rossi *et al.*, 2011). Despite this, the ecological structure and functional interconnections of the human microbiome are still not well understood (Chen *et al.*, 2014).

These gaps in our knowledge can be attributed in part to the fundamental complexity of microbial communities. Microbiota functions as a system of multiple interdependent actors (Tyler *et al.*,

2014). The stability and dynamics of this system are thus not the direct result of any one organism but instead emerge from the sum total of the interactions contained within (Buffie *et al.*, 2014; Pepper and Rosenfeld, 2012). Understanding the functional behavior of a microbiome thus requires information on not only the composition of the microbiome, but also how its individual components interact (Stein *et al.*, 2013).

Data from large-scale studies of the human microbiome provide an avenue for a data-driven understanding of interactions within a community. These studies typically use high-throughput sequencing to quantify individual abundances within a microbial community. These abundances can then be used by a number of statistical techniques to predict microbial interactions (Ban *et al.*, 2015; Friedman and Alm, 2012). However, the general performance of such methods can be difficult to assess in the absence of a 'gold standard' database of experimentally validated interactions.

Microbial interactions have long been experimentally investigated. Coculture experiments and growth assays allow for identification of both direction and strength of bacterial interactions (Trosvik *et al.*, 2010). However, current information describing microbial interactions is spread across a large, constantly increasing set of papers; PubMed alone references upwards of 24 million articles at the time of writing, of which almost two million refer to bacteria. Consolidation of literature curated data into a single gold-standard database would be a valuable resource for refining microbial interaction models. However, the sheer quantity of information makes this an imposing proposition with brute-force manual collation being infeasible.

Microbial interactions are not the first field to experience this challenge. The implementation of automated systems for information extraction has been frequently used to mine large datasets such as the unstructured text compendiums of scientific literature. Text-mining techniques have shown promising results in biomedical information extraction tasks involving Protein–Protein Interactions (PPIs) (Donaldson *et al.*, 2003), disease–drug/drug–drug interactions (Chen *et al.*, 2008; Tari *et al.*, 2010) and literature annotation (Liu *et al.*, 2015) among others. For example, the database preBIND augmented traditional curation methods with a PPI identification system. By pre-filtering of the literature to remove irrelevant data, they achieved an estimated 60% reduction in annotation time (Donaldson *et al.*, 2003).

For microbial interactions, Freilich *et al.* (2010) performed the pioneering, and at the time of writing, only automated analysis of the literature surrounding microbial interactions. They analyzed the co-occurrence of bacterial species in the literature through a statistical testing approach, organizing hits into a co-occurrence network. However, other than this study, text-mining has not been applied for the identification of microbial interactions. In particular, the application of machine learning approaches has not been explored and is also not straightforward, as evidence for a microbial interaction can be embedded in a hard to parse fashion in paper abstracts. For example, a paper abstract could describe the production of a specific molecule M by species A in the first part of the abstract, while the second part highlights the reduction in growth of species B in the presence of M, without explicitly stating that A can inhibit the growth of B. While this is a simplified example, in practice, even more complex textual and logical analysis may be needed to mine this information. A more tractable approach may be to use combinations of keywords to first identify abstracts reporting microbial interactions as proposed here. An additional challenge for machine learning approaches is the absence of sufficiently large datasets for model training.

Here, we describe @MInter (acronym for *A*utomated *T*ext-mining of *M*icrobial *Inter*actions), a machine-learning-based text-mining system for automated identification and extraction of microbial interaction data. @MInter complements manual annotation and curation through the identification of potentially informative texts, to accelerate the pace of data curation. The @MInter system comprises two core components, a Bioconductor-based crawler for the acquisition of article data and a Support-Vector Machine (SVM)-based classifier for the identification of abstracts containing interaction information. For comparison, we also evaluated the performance of the co-occurrence method devised by Freilich *et al.*, as well as a rudimentary pattern-based classifier available as part of @MInter. @MInter's SVM-based approach was found to classify abstracts containing one or more microbial interactions with a specificity of 95% and notably reduce annotator workload for identifying specific microbial interactions (13-fold compared to competing approaches).

In addition, we present a large annotated dataset for microbial interactions consisting of 3917 abstracts (annotated to describe the presence and type of the interaction), as well as a collated set of 735 known species interactions, that can serve as a resource for the development of new interaction mining algorithms. In addition, as a proof-of-concept, @MInter was employed to systematically construct a curated database of interactions between 175 bacterial species commonly found on human skin, highlighting the feasibility of large-scale use and providing a resource for ongoing studies of the human microbiome.

2 Materials and methods

2.1 Creation of a training dataset for @MInter

2.1.1 Acquisition of paper abstracts

Supervised classifiers such as the SVM used in the @MInter system require an annotated set for training and testing purposes. To facilitate this, we first created an annotated corpus of PubMed abstracts. To acquire raw abstract data, 1747 query species associated with the human gastrointestinal tract were initially selected as reported in MEDUSA data (Karlsson *et al.*, 2014). Due to practicality concerns, species with low abundance (mean read count <50) were discarded, resulting in a total of 482 species. Queries following the pattern 'Species_1 AND Species_2' were then derived from all possible species pairs. These were queried against the PubMed database to produce a collection of all papers containing both species names. The PubMed ID, abstract text and title of each paper were retained and organized into their respective interaction pairs for later analysis. Only pairs possessing at least one article were retained. For the initial dataset, information for 11 546 interaction pairs was collected and sampled from.

2.1.2 Construction of a core dataset

Lactobacillus acidophilus is a probiotic bacterium well known for its antagonistic interactions with multiple other bacteria such as *Escherichia coli*. As such, the *L. acidophilus–E. coli* query pair was selected as a core dataset for the training of all SVMs. For this dataset, all abstracts pertaining to the pair ($n = 350$) were manually classified as follows (Supplementary File S1):

1. No suggested interaction (negatives)
2. Interaction between other species (positives)
3. Interaction between *L.acidophilus* and *E.coli* (positives)

2.1.3 Construction of an extended dataset

To construct an extended dataset for training and evaluation of classifiers, a subset of 1000 species pairs (735 after excluding ambiguous species names) was randomly sampled from the abstract dataset and annotated on two levels.

First, species level annotation considered whether any of the abstracts containing a species pair described a true interaction. To do this, abstracts were manually scanned by a single annotator (L.K.M.K.) for information pertaining to interaction between species pairs in the list. Species pairs where the first 200 abstracts containing both species did not report an interaction were considered to be not interacting (to reduce annotator load). Overall, 62 interacting pairs were identified (out of 735) out of which 56 were inhibitory and only 6 were non-inhibitory (Supplementary File S2).

Second, abstract level annotation identified abstracts that describe a true interaction. Here, brute force random sampling from all abstracts produced a heavily unbalanced set (3 positives; 197 negatives) that is not well suited for the training of a supervised classifier. To address this, an SVM trained on the core dataset (see Section 2.2) was applied to abstracts from the species level annotation to identify and enrich for true positives. Putative positives from the core dataset trained SVM were then manually annotated. By comparing to known interaction pairs from the species level annotation, false-negative species pairs reported by the core SVM were identified, i.e. species pairs that are known to be interacting but with no abstracts identified by the SVM. All abstracts for such pairs were further manually annotated. In total, 3917 (350 core, 3567 extended) individual abstracts (241 positives and 3676 negatives) were annotated at this level (Supplementary Files S1 and S3). Thus the abstract level annotation only includes abstracts from interacting species pairs. Because of the use of a SVM classifier to aid the generation of the dataset, it is expected to be biased and distinct from a random sampling of abstracts. However, the exclusion of abstracts from true-negative species pairs from the extended abstract dataset (likely to be easy to classify), and the inclusion of all false-positive and false-negative abstracts from the core SVM is likely to bias the training dataset to harder test cases for SVM training. Results from annotation of interactions between skin bacteria (see Section 3.4) further support the notion that the training datasets for the SVM were more challenging than real datasets.

2.2 Training of the @MInter classifier

A SVM is a supervised machine learning model based upon the principle of structural risk minimization (Joachims, 1998). SVMs are capable of learning patterns from text datasets, and have seen extensive use in a number of data categorization tasks (Joachims, 2001; Tong and Koller, 2001).

@MInter's classifier is based on the SVM module in the skLearn package (Pedregosa *et al.*, 2011). Training was done on the core dataset along with 5-fold cross-validation based on the extended dataset i.e. core dataset and four-fifths of extended dataset were used to train and one-fifth of the extended dataset was used to test. Abstracts were converted to a feature vector (count for each word as a feature) using the inbuilt skLearn vectorizer. Feature vectors were sparse as expected (0.4% non-zero elements), with an average abstract containing 183 words and feature space dimensionality of 25 783. Feature values were weighted using Term Frequency-Inverse Document Frequency (TF-IDF), a heuristic measure of term relevance commonly used in text analysis (Ramos *et al.*, 2003). Unlike raw term frequency counts, TF-IDF weighs document term frequencies against its frequency across the entire corpus, downscaling the relevance of frequently occurring terms. The implementation of TF-IDF used in this study is as follows:

$$\mathrm{TF-IDF}(t,d,D) = \mathrm{tf}(t,d) \times (\mathrm{IDF}(t,D) + 1),$$

where $\mathrm{tf}(t,d) = \mathrm{f}_{t,d}$, the frequency of term t in document(s) d and, $\mathrm{IDF}(t,D) = \ln(N/\mathrm{f}_{t,D})$, with N being the total number of documents in corpus D.

A linear kernel was chosen for the SVM. Radial basis function kernels were also evaluated, but did not provide significant performance improvement. The C parameter was kept at default value ($C = 1$; lower misclassification penalty) as optimizing it with a grid search and cross-validation did not result in significant performance improvements and we wished to conserve our limited training set. To account for class imbalance in the dataset, automated class weighting was used (Pedregosa *et al.*, 2011). Note that the SVM classifier in @MInter can only identify abstracts that are likely to report a microbial interaction. This is based on the ability of the SVM to identify combinations of features/terms that are more enriched in abstracts reporting interactions. Once an abstract is identified, all pairs of species reported in it then need to be considered as candidate microbial interactions, as abstracts frequently report multiple interactions.

2.3 An alternate pattern-based method in @MInter

A simple pattern-based system for interaction identification was implemented to complement and compare with the machine-learning approach used in @MInter. Source data consisted of abstracts from the core dataset, as well as the extended dataset. Permutations of bacterial names, as well as interaction-describing keywords were manually derived from the source data for later use as patterns. Due to the limited number of abstracts containing non-inhibitory interactions in the annotated dataset, keywords were restricted to those suggesting inhibitory interactions. Keywords included words such as 'inhibit', 'bacteriocin' and 'lysed'. To account for inflective variations in individual words, abstracts and keywords were stemmed using the Snowball algorithm (Porter, 2001). Derived rules were implemented using patterns of regular expression, where terms were species names from the evaluated bacterial pair, as well as selected keywords. The terms were then joined with the '.*' token (allowing for arbitrary intervening matches) to give the final pattern expressions. A sample pattern, 'Species_1 inhibits Species_2' would thus convert to the regex 'Species_1.*inhibit.*Species_2'. A full list of patterns used can be found in Supplementary File S4.

2.4 Reimplementation of a co-occurrence-based approach

The co-occurrence-based approach proposed by Freilich *et al.* (2010), and a straightforward extension, were implemented for comparison with @MInter. Specifically, for each species pair in the extended dataset, three queries were performed on PubMed to compute the number of papers containing each species (species name as query term), and the number containing both species (both species names joined by an 'AND' as query term). A straightforward extension also included the stemmed keywords from a manually curated term set (Supplementary File S4).

To compute statistical significance for the number of abstracts for any species pair (or pair and term), a Fisher's exact test was performed with a contingency table containing four categories: abstracts containing solely species A, abstracts containing solely species B, abstracts containing neither species A or B and abstracts containing both A and B. A Bonferroni corrected P-value threshold

of 0.01 was used to identify overrepresented and thus potentially interacting species pairs.

2.5 Evaluation metrics

Due to the presence of a skewed evaluation dataset (in terms of number of positives and negatives) accuracy-based measures were not used and instead, we report primarily the recall, specificity and precision of the methods, as well as the papers-per-hit (PPH) metric described below. All evaluations were done based on a held out test set in a cross-validation framework using the manually curated extended dataset.

2.5.1 Recall, specificity and precision

Given a set of classified entries, recall, specificity and precision are defined as follows:

$$\text{Recall} = \text{TP}/(\text{TP} + \text{FN})$$

$$\text{Specificity} = \text{TN}/(\text{TN} + \text{FP})$$

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP})$$

where TP, TN, FP and FN are true positives, true negatives, false positives and false negatives, respectively.

2.5.2 Papers-per-hit

In order to capture the value of the various approaches in reducing annotator effort, we computed PPH as an additional metric. Specifically, PPH is computed as the average number of abstracts reported by a method per true-positive interaction identified. For the co-occurrence-based approaches, all abstracts from false-positive and true-positive interaction pairs were considered as being reported.

2.6 Analysis of skin microbiome data

Skin metagenomic profiles from Oh *et al.* (2014) (291 samples in total) were used to infer interacting species based on correlations in abundances (Spearman correlation). In addition, CCREPE (Bielski and Weingart, 2014) was used to correct for the compositional effect (Faust *et al.*, 2012) arising from working with relative abundance profiles. Correlations with *P*-value less than 0.01 were compared with literature-curated interactions from @MInter analysis (see Section 3.4).

2.7 Software

@MInter and associated programs were written entirely in the Python 3 programming language. Source code is available at https://github.com/CSB5/atminter under the MIT license.

3 Results

3.1 @MInter has high specificity at the abstract level

The performance of @MInter was first evaluated at the level of abstracts (using the abstract level annotation). Cross-validation-based analysis indicated an 'area under curve of receiver operating characteristic' (AUC-ROC) of 0.97 for the SVM method (Fig. 1) . At a confidence threshold of 0.5, this provided 82% recall, 95% specificity and 49% precision. In comparison, the alternate 'pattern scanner' approach had lower recall (19%) with similar specificity (96%) and precision (44%). As multiple abstracts can potentially provide information about the same interaction pair, low recall at the abstract level could be compensated for at the species level. However, this may not be the case if recall is too low. Since the primary purpose of @MInter was to serve as a first-pass filter to enrich for abstracts that pertain to microbial interactions, high recall at the species level was set as the primary design goal (see Section 3.2). For applications where higher precision may be required, using a larger confidence value threshold for the SVM classifier might be more appropriate (e.g. threshold of 0.9 provided 72% precision, 16% recall and 99.5% specificity).

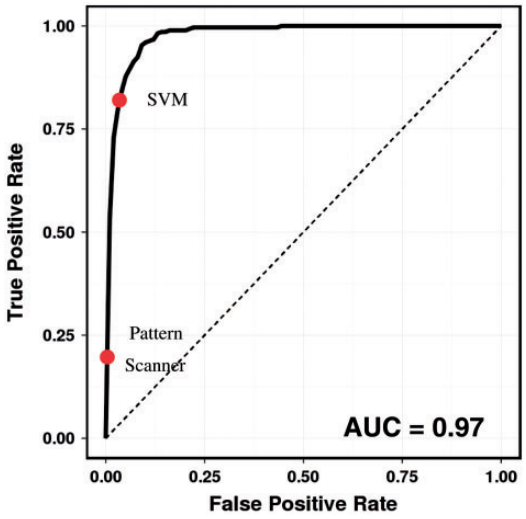

Fig. 1. ROC curve for the @MInter SVM classifier. The dots indicate the default performance of the SVM classifier (confidence threshold = 0.5) and the pattern scanner approach, respectively

3.2 @MInter improves performance at the species level

We next evaluated @MInter's SVM and pattern scanner approach at the species level and compared them to the Fisher's exact test used in Freilich *et al.* (as well as a modified version with inhibition terms: 'Fisher w/inhibition'). Note that the approaches based on statistical testing cannot be applied at the abstract level.

To measure enrichment over random selection, the ratio of recall to proportion of species pairs returned as true positives was computed and plotted for 5-fold cross-validation analysis (Fig. 2). Both the SVM and pattern scanner approaches provided an improvement (>2-fold) over random sampling where the ratio is expected to be 1. The pattern scanner approach was found to have better enrichment due to its high specificity (Table 1). In contrast, the approach proposed in Freilich *et al.* (Naïve Fisher) provided a more modest improvement that was further boosted by the inclusion of inhibition terms as proposed here.

Species-level results for @MInter and other methods are further summarized in Table 1. As expected, compared to the abstract level analysis, @MInter's SVM predictions have high recall at 95% (versus 82% at abstract level). This allows sensitive mining of abstracts, albeit at the expense of modest specificity (78%). Overall, @MInter's SVM performance was found to be a notable improvement compared to Freilich *et al.*, even compared to the improved approach using inhibition terms (Table 1). The pattern scanner approach on the other hand provided the best specificity (92%) but with low recall (24%). @MInter's species-level precision was found to be low despite being the best across all methods. This is partly because it identifies abstracts containing interactions and reports all pairs of species names in such abstracts. A more refined approach

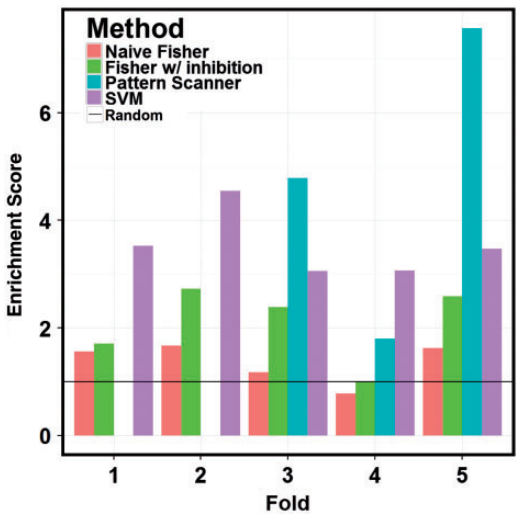

Fig. 2. Comparison of methods based on enrichment score. Enrichment was computed as the ratio of recall versus the proportion of abstracts returned as true positives by a method. Results shown are for 5-fold cross validation. @MInter's pattern scanner was not tested on folds 1 and 2 as patterns were derived from those folds

Table 1. Performance summary for different methods at the species level

Method	Recall (%)	Specificity (%)	Precision (%)
All Positives (AP)	100	0	7
@MInter: SVM	**95**	78	**25**
@MInter: Pattern Scanner	24	**92**	15
Naive Fisher (Freilich *et al.*)	51	63	10
Fisher w/Inhibition	50	77	15

The 'All Positives' approach reports all considered species pairs as being interacting ones and is provided here as a control method for comparison. Bold entries represent highest scores for each column.

that ranks species pairs in an abstract in terms of their likelihood of interacting could further improve precision.

3.3 Impact on annotator workload using @MInter

The primary utility of an information extraction system such as @MInter is the reduction of annotator workload, in this case when identifying microbial interactions from paper abstracts. As scientific abstracts can be complex to parse and often contain specialized vocabulary, it is unlikely that human intervention can be completely eliminated. We used the notion of PPH to measure the utility of various approaches from an annotator's perspective (see Section 2.5.2). Overall, both the SVM and pattern scanner approach in @MInter allow annotators to identify a reported microbial interaction by reading $\leq$ 6 abstracts on average (Table 2). This is a significant reduction compared to the statistical approach in Freilich *et al.* (>15-fold) as well as a naïve approach involving the annotation of all abstracts (AP; 13-fold). Notably, due to lower recall, the statistical approaches place a greater burden on the annotator than even a naïve approach (Table 2). The SVM-based approach was chosen as the default option in @MInter, due to its high species level recall (95%) compared to the pattern scanner approach.

Table 2. Comparison of different methods in terms of PPH

	AP	@MInter: SVM	@MInter: Pattern Scanner	Naïve Fisher (Freilich *et al.*)	Fisher w/ Inhibition
PPH	67	5	6	79	109

The 'All Positives' (AP) approach reports all considered species pairs as being interacting ones and is provided here as a control method for comparison.

3.4 Application of @MInter for large-scale annotation

Training of the @MInter SVM classifier was found to be quite efficient taking <5 min overall. Classification of abstracts can then be performed at the rate of >90 abstracts/s. Applying the @MInter SVM classifier to the PubMed database (>25 million abstracts) would, therefore, require slightly more than 3 days on a single compute core. The principal bottleneck for applying @MInter to the PubMed database is thus in the acquisition and storage of abstracts for further processing.

To investigate the feasibility of applying @MInter for large-scale annotation, we explored its use for constructing a microbial interaction database focused on skin bacteria. Specifically, we selected a set of 175 bacterial species known to be relatively abundant on human skin (Supplementary File S5) (Oh *et al.*, 2014). The PubMed database was then queried with @MInter for abstracts containing all pairs of species from this list. In total, 29 948 abstracts were obtained of which 3196 were marked as containing interactions by the @MInter SVM. Disjoint subsets of these were than manually annotated by a team of 11 annotators to obtain an interaction network involving >100 species and 357 literature-reported microbial interactions (Supplementary File S6). Reassuringly, observed annotator workload was lower than originally estimated in our test datasets (PPH = 3.3).

The curated network of interactions between bacteria commonly found on skin (Supplementary File S7; a subset of Supplementary File S6 as abstracts can contain other species interactions as well) highlighted several relevant interactions including the inhibition of *Staphylococcus aureus* (an important skin pathogen) by several *Lactobacillus* and *Bifidobacterium* species (known probiotics; Fig. 3). Interestingly, multiple *Streptococcus* species were also observed to exert inhibitory influence over *S. aureus*. To evaluate overlap with *in vivo* data, correlation analysis was used to identify bacterial interactions from skin microbiome data (see Section 2.6). From the significant edge overlap between the two networks (Fisher's exact test P-value = 0.00038), we obtained a high confidence sub-network that combines *in vivo* as well as *in vitro* evidence, and describes interactions involving two important skin pathogens (*S. aureus* and *Propionibacterium acnes*) (Fig. 4). Furthermore, the network provides information on the directionality of interactions (manually curated from abstracts) enabling the modeling of this core interaction network.

4. Discussion and conclusion

Interactions reported in the scientific literature are likely to have a culturability and 'interest' bias. In addition, this information can be sparse as most interactions have likely not been tested or have been tested under limited *in vitro* conditions. On the other hand, approaches that look for correlations in microbiome sequencing profiles are expected to be less biased and provide an *in vivo* perspective (Bielski and Weingart, 2014). Such predictive methods can,

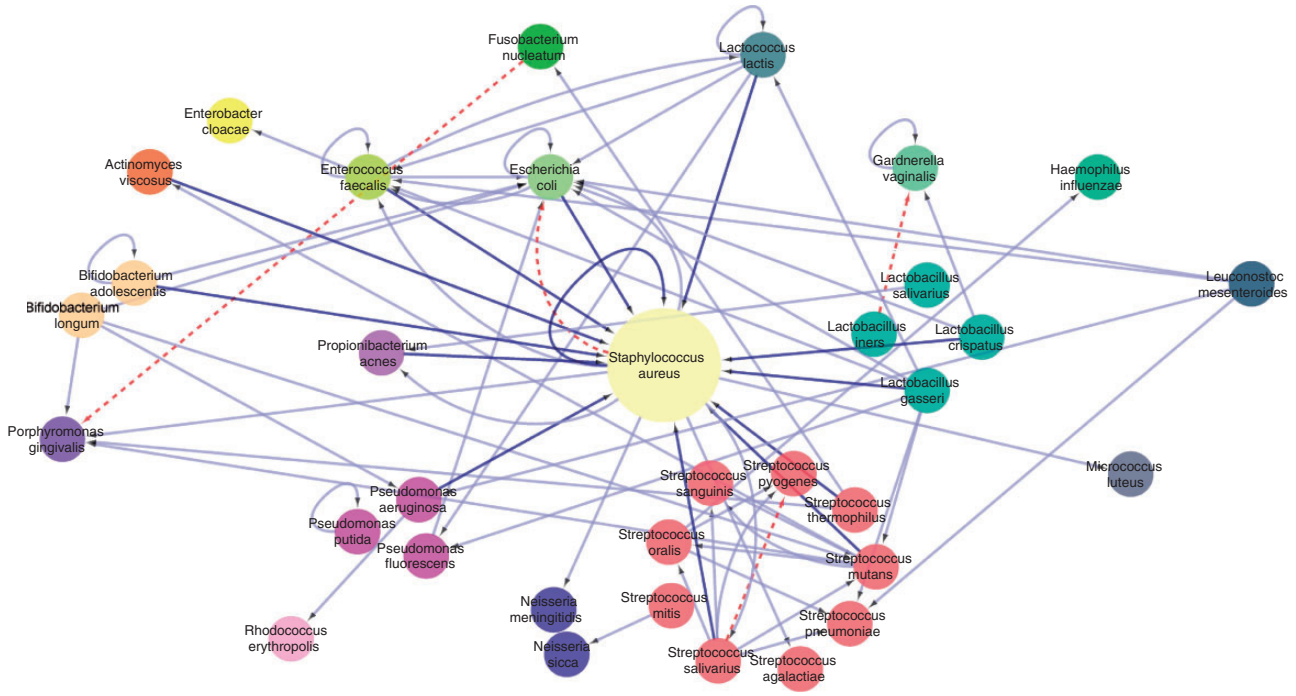

Fig. 3. A curated network for interactions between skin bacteria constructed using @MInter analysis of PubMed abstracts. Dashed lines represent non-inhibitory interactions while solid lines represent inhibitory interactions. Dark solid lines depict inhibitory interactions with *S. aureus,* a key skin pathogen

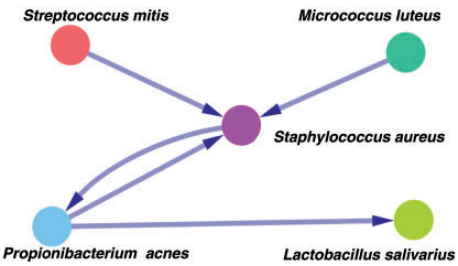

Fig. 4. A high-confidence core network of interactions between skin bacteria. The network was obtained based on overlap between literature-curated interactions (generally *in vitro;* Fig. 3) and correlation analysis based on *in vivo* skin microbiome data

however, be influenced by many sources of noise (sparse profiles, transitive interactions, etc.) that reduce their specificity and sensitivity. Collections of *in vitro* validated interactions curated from the literature can thus play a complementary role, facilitating the evaluation of interaction prediction algorithms and enabling a better understanding of microbial interaction networks.

The concept of microbial interactions is complex as it is context dependent and not restricted to two entities. The analysis in this manuscript was restricted to binary interactions, and interactions were collated at the species level as this is typically the taxonomic level for descriptions in abstracts. The text mining approach used in @MInter is however agnostic to the taxonomic level of analysis or mode of interaction and with appropriate training could be used to e.g. mine ternary interactions at the strain level.

@MInter improves significantly over existing approaches for text-mining of microbial interactions and it enables the large scale curation of interaction information. However, @MInter does not identify the mode/directionality of interactions or eliminate the need for manual curation. While further work in developing methods with higher precision (potentially by analyzing full manuscripts) could reduce manual effort, based on our own experience with the complexity of information in abstracts, we believe that it is unlikely that human input can be eliminated completely. We hope that the availability of curated datasets generated and used here would enable the further development of supervised machine learning approaches for this problem.

Acknowledgements

We would like to thank Ng Hui Qi Amanda, Boey Jia Hui Esther, Wu Guangxi, Denis Bertrand, Andreas Wilm, Davide Verzotto and Sun Miao for their assistance in manual annotation of abstracts.

Funding

This work was supported by the IMaGIN platform (project No. 102 101 0025), through a grant from the Science and Engineering Research Council as well as funding to the Genome Institute of Singapore from the Agency for Science, Technology and Research (A*STAR), Singapore.

Conflict of Interest: none declared.

References

Ban,Y. *et al.* (2015) Investigating microbial co-occurrence patterns based on metagenomic compositional data. *Bioinformatics (Oxford, England)*, btv364. doi: 10.1093/bioinformatics/btv364.

Bergonzelli,G.E. *et al.* (2006) GroEL of *Lactobacillus johnsonii* La1 (NCC 533) is cell surface associated: Potential role in interactions with the host and the gastric pathogen Helicobacter pylori. *Infect. Immunity*, **74**, 425–434.

Bielski,E. and Weingart,G. (2014) ccrepe: ccrepe_and_nc.score. *R package version 1.6.0.*

Buffie,C.G. *et al.* (2014) Precision microbiome reconstitution restores bile acid mediated resistance to Clostridium difficile. *Nature*, **517**, 205–208.

Chen,E.S. *et al*. (2008) Automated acquisition of disease-drug knowledge from biomedical and clinical documents: an initial study. *J. Am. Med. Inf. Assoc*., **15**, 87–98.

Chen,Y. *et al*. (2014) Functional gene arrays-based analysis of fecal microbiomes in patients with liver cirrhosis. *BMC Genomics*, **15**, 753.

Donaldson,G.P. *et al*. (2015) Gut biogeography of the bacterial microbiota. *Nature Reviews Microbiology, (October). Nature Publishing Group*, doi:10.1038/nrmicro3552

Donaldson,I. *et al*. (2003) PreBIND and Textomy – mining the biomedical literature for protein-protein interactions using a support vector machine. *BMC Bioinformatics*, **4**, 11.

Faust,K. *et al*. (2012) Microbial co-occurrence relationships in the Human Microbiome. *PLoS Comput. Biol*., **8**, doi:10.1371/journal.pcbi.1002606

Freilich,S. *et al*. (2010) The large-scale organization of the bacterial network of ecological co-occurrence interactions. *Nucleic Acids Res*., **38**, 3857–3868.

Friedman,J. and Alm,E.J. (2012) Inferring correlation networks from genomic survey data. *PLoS Comput. Biol*., **8**, 1–11. doi: 10.1371/journal.pcbi.1002687

Guinane,C.M. and Cotter,P.D. (2013) Role of the gut microbiota in health and chronic gastrointestinal disease: understanding a hidden metabolic organ. *Therap. Adv. Gastroenterol*., **6**, 295–308.

Jalali,S. *et al*. (2015) Screening currency notes for microbial pathogens and antibiotic resistance genes using a shotgun metagenomic approach. *Plos One*, **10**, e0128711.

Joachims,T. (1998) Text Categorization with Support Vector Machines: Learning with Many Relevant Features. Springer, Berlin, Heidelberg, pp. 137–142.

Joachims,T. (2001) A statistical learning learning model of text classification for support vector machines. In: *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 128–136.

Karlsson,F.H. *et al*. (2014) Metagenomic Data Utilization and Analysis (MEDUSA) and construction of a global gut microbial gene catalogue. *PLoS Comput. Biol*., **10**, doi:10.1371/journal.pcbi.1003706

Liu,W. *et al*. (2015) OntoMate: a text-mining tool aiding curation at the Rat Genome Database. *Database*, **2015**. doi:10.1093/database/bau129

Morgan,X.C. *et al*. (2012) Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biol*., **13**, R79.

Oh,J. *et al*. (2014) Biogeography and individuality shape function in the human skin metagenome. *Nature*, **514**, 59–64.

Pedregosa,F. *et al*. (2011) Scikit-learn: machine learning in Python. *J. Mach. Learn. Res*., **12**, 2825–2830.

Pepper,J.W. and Rosenfeld,S. (2012) The emerging medical ecology of the human gut microbiome. *Trends Ecol. Evol*., **27**, 381–384.

Porter,M.F. (2001) Snowball: A language for stemming algorithms.

Ramos,J. *et al*. (2003) Using TF-IDF to Determine Word Relevance in Document Queries. *Processing*. doi:10.1.1.121.1424

Rossi,M. *et al*. (2011) Folate production by probiotic bacteria. *Nutrients*, **3**, 118–134.

Stein,R.R. *et al*. (2013) Ecological modeling from time-series inference: insight into dynamics and stability of intestinal microbiota. *PLoS Comput. Biol*., **9**, 31–36.

Tari,L. *et al*. (2010) Discovering drug-drug interactions: a text-mining and reasoning approach based on properties of drug metabolism. *Bioinformatics*, **26**, 547–553.

Tong,S. and Koller,D. (2001) Support vector machine active learning with applications to text classification. *J. Mach. Learn. Res*., 45–66. doi:10.1162/153244302760185243

Trosvik,P. *et al*. (2010) Web of ecological interactions in an experimental gut microbiota. *Environ. Microbiol*., **12**, 2677–2687.

Turnbaugh,P.J. *et al*. (2006) An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature*, **444**, 1027–1031.

Tyler,A.D. *et al*. (2014) Analyzing the human microbiome: a "How To" guide for physicians. *Am. J. Gastroenterol*., **109**, 983–993.

Weyrich,L.S. *et al*. (2015) The skin microbiome: associations between altered microbial communities and disease. *Aust. J. Dermatol*., doi:10.1111/ajd.12253

Zeglin,L.H. *et al*. (2015) Organic matter quantity and source affects microbial community structure and function following volcanic eruption on Kasatochi Island, Alaska. *Environ. Microbiol*., doi:10.1111/1462-2920.12924