Observational Study

Medicine®

OPEN

A MeSH-based text mining method for identifying novel prebiotics

Guangyu Shan, MS, Yiming Lu, PhD, Bo Min, PhD, Wubin Qu, MS, Chenggang Zhang, PhD*

Abstract

Prebiotics contribute to the well-being of their host by altering the composition of the gut microbiota. Discovering new prebiotics is a challenging and arduous task due to strict inclusion criteria; thus, highly limited numbers of prebiotic candidates have been identified. Notably, the large numbers of published studies may contain substantial information attached to various features of known prebiotics that can be used to predict new candidates. In this paper, we propose a medical subject headings (MeSH)-based text mining method for identifying new prebiotics with structured texts obtained from PubMed. We defined an optimal feature set for prebiotics prediction using a systematic feature-ranking algorithm with which a variety of carbohydrates can be accurately classified into different clusters in accordance with their chemical and biological attributes. The optimal feature set was used to separate positive prebiotics from other carbohydrates, and a cross-validation procedure was employed to assess the prediction accuracy of the model. Our method achieved a specificity of 0.876 and a sensitivity of 0.838. Finally, we identified a high-confidence list of candidates of prebiotics that are strongly supported by the literature. Our study demonstrates that text mining from high-volume biomedical literature is a promising approach in searching for new prebiotics.

Abbreviations: AUC = area under the curve, MeSH = medical subject headings, NLM = National Library of Medicines, RF = random forest, ROC = receiver operating characteristic curve, XML = extensible markup language.

Keywords: Carbohydrates, MeSH-term, Prebiotics, Prebiotics prediction, Text mining

1. Introduction

The health benefits of prebiotics, such as cancer risk reduction, immune system enhancement, and constipation relief have been widely accepted. A food ingredient can be considered a prebiotic only when it satisfies 3 criteria: (1) resistant to gastric acidity and mammalian enzymes, (2) prone to fermentation by intestinal microbiota, and (3) selective to stimulation of the growth and/or activity of beneficial intestinal microbiota.[1] Identifying new prebiotics in accordance with these 3 criteria via the screening of various chemical compounds is a very laborious and challenging task. Scientists have been performing related work since 1995 when the criteria were first proposed. However, only two carbohydrates have been reported until 2007: Inulin and Fructooligosaccarides.[1]

Several researchers began to develop other approaches by reviewing published literature and searching for keywords in PubMed, and 3 carbohydrates were shown to alter the microbiota balance of the large bowel by increasing the number of *bifidobacteria* and *lactobacillus*. The success of these studies suggested the possibility of using a text mining-based method to identify prebiotics by transforming the inclusion criteria into a collection of literal features. Text mining efforts developed a variety of approaches to obtain information in structured biomedical text using techniques such as machine learning, natural language processing, biostatistics, information technology, and pattern recognition.[2]

In the rapidly growing fields of knowledge discovery and text mining, relevant literature can be used to obtain implicit and unrevealed information. Swanson[3] began to mine information from biomedical literature for Raynaud disease treatment in 1986. He found from a biomedical paper that Raynaud disease is a peripheral circulatory disorder associated with and exacerbated by high platelet aggregation, high blood viscosity, and vasoconstriction; in other biomedical literature, he found that fish oil could reduce these symptoms. Accordingly, he proposed the hypothesis that fish oil may be helpful for people suffering from Raynaud disease, which had not previously been reported. Three years later, this hypothesis was clinically confirmed by DiGiacomo et al.[4] Corresponding to this method, Ramadan et al[5] traced 11 indirect connections between migraines and magnesium using summaries of published papers, and the effect

Editor: Giovanni Tarantino.

GS and YL have contributed equally to this work.

Author Contributions: Conceived and designed the experiments: GS, YL, and BM. Performed the experiments: GS. Analyzed the data: GS, YL, and BM. Contributed reagents/materials/analysis tools: GS. Wrote the paper: GS, LY, WQ, and CZ.

Funding provided by the National Basic Research Project (973 program) (2012CB518200), the General Program (31401141, 81573251, 30900830) of the Natural Science Foundation of China, the State Key Laboratory of Proteomics of China (SKLP-Y201303, SKLP-O201104, and SKLP-K201004), and the Special Key Programs for Science and Technology of China (2012ZX09102301–016).

The authors have no conflicts of interest to disclose.

Supplemental Digital Content is available for this article.

Beijing Institute of Radiation Medicine, State Key Laboratory of Proteomics, Cognitive and Mental Health Research Center, Beijing, PR China.

* Correspondence: Chenggang Zhang, Academy of Military Medical Sciences, Beijing, PR China (e-mail: zhangcg@bmi.ac.cn).

Medicine (2016) 95:49(e5585)

Received: 7 August 2016 / Received in final form: 2 November 2016 / Accepted: 7 November 2016

http://dx.doi.org/10.1097/MD.0000000000005585

of magnesium was later experimentally validated.[6] Thus far, text mining has become an indispensable tool for extracting knowledge from biomedical literature.

Feature selection is a critical procedure for text mining to tease out valuable features from large amounts of data.[7] Many techniques, such as support vector machine (SVM),[8] genetic programming (GP),[9,10] logistic regression (LR),[11] and probabilistic neural network (PNN),[12] can perform this process only in a general and cursory manner. MedMeSH summarizer can assess very large amounts of biomedical data in a short period and is generally used for genome-wide expression profiles.[13] MedMeSH summarizer can achieve decent performance in specific as opposed to general assessments.

Inspired by MedMeSH and the philosophy of mining tacit knowledge from biomedical literature, we herein developed a novel medical subject headings (MeSH)-based text mining method for identifying new prebiotics utilizing the PubMed database. PubMed comprises more than 24 million citations for biomedical literature from MEDLINE, life science journals, and online books.[14] MeSH is the National Library of Medicines (NLM)-controlled vocabulary thesaurus specified for indexing articles from PubMed. We extracted from MeSH because it is easily available through the PubMed service of the National Library of MEDLINE, whereas full texts of research studies are often only accessible by subscription.[15] Additionally, utilizing MeSH rather than the full text not only reduces computation time but also enables higher dataset throughput.[16] Bhattacharya et al demonstrated that MeSH terms could represent the whole text accurately if screened appropriately, that is, we can extract representative features from massive amounts of literature using these high-quality widgets.[16]

We hypothesized that carbohydrates with the properties of prebiotics share similar literal features. To better extract the features of known prebiotics, we first used an exhaustive text mining approach to mine prebiotic-related topical MeSH terms from structured documents downloaded from PubMed. And then selected a list of optimal MeSH terms that are closely related to known prebiotics[17] and ranked a large set of carbohydrates according to the scores calculated from their MeSH frequency profiles. At last, we used a cross-validation technique to assess the prediction accuracy of our model.

2. Methods

2.1. Data preparation

Firstly, 2 kinds of data were being prepared: positive prebiotics set and carbohydrates set. We used a list of positive prebiotics summarized by Al-Sheraji et al.[14] The list is in Table 1 which contains 15 prebiotics that we denoted as positive prebiotics set. Nearly all positive prebiotics are non-digestible carbohydrates. Thus, we constructed carbohydrates set using the official names of all available carbohydrates from the NLM MeSH tree structures. To ensure the specificity of the prediction, only carbohydrates that belong to the lowest level of the tree were selected, with the exception of the lowest-level carbohydrates that could not cover the carbohydrates represented by their parent node (in this case, the parent node was also included). Positive prebiotics were also removed from the carbohydrates set. The final carbohydrates set contains 112 carbohydrates (Supporting Information, S1 Table. The official names of carbohydrates set. (XLSX), http://links.lww.com/MD/B447; S2 Table. The official names of 50 positives for method validation. (XLSX), http://links.lww.com/MD/B448). Each of the names of 15 positive prebiotics and 112 carbohydrates were used as a query to search relevant literature in PubMed, and the hit documents were downloaded in extensible markup language (XML) format, respectively. MeSH terms in the XML documents are extracted using the ElementTree Python package. Therefore, each substance contains a MeSH term list extracted from its relevant literature. Each list contains thousands of features, which will enable us a robust foundation for the final model. This study did not require the ethical approval and informed consent due to all analyses were carried out based on the data extracted from previous published literature.

Table 1

Types and sources of known prebiotics.

Type of prebiotic	Sources of prebiotics	References
Inulin	Wheat, onion, bananas	[1]
Fructooligosaccharides	Asparagus, sugar beet, garlic, etc.	[18]
Isomaltulose	Honey, sugarcane juice	[19]
Xylooligosaccharides	Bamboo shoots, fruits, vegetables, etc.	[20]
Galactooligosaccharides	Human's milk and cow's milk	[21]
Cyclodextrins	Water-soluble glucans	[22]
Raffinose oligosaccharides	Seeds of legumes, lentils, peas, etc.	[23]
Soybean oligosaccharides	Soybean	[24]
Lactulose	Lactose (milk)	[25]
Lactosucrose	Lactose	[26]
Palatinose	Sucrose	[19]
Maltooligosaccharides	Starch	[27]
Isomaltooligosaccharides	Starch	[27]
Arabinoxylooligosaccharides	Wheat bran	[28]
Enzyme-resistant dextrin	Potato starch	[29]

2.2. Stop words filtering

Stop words, which can undermine the efficacy and effectiveness of the mining task due to high frequency, usually need to be removed first. MeSH curators removed traditional stop words such as "a," "the," and "for"; however, some MeSH terms with extremely high-frequency remain, which significantly reduces model performance. These MeSH terms were filtered according to Zipf law. Zipf law states that the rank-proportional frequency of a word is inversely proportional to its frequency rank among all words in a given natural language corpus. Thus, the purity of the corpus can be optimized by removing MeSH terms with particularly high frequency under the following filter procedure.

1. Initiate a query list containing all carbohydrates in positive prebiotics set and carbohydrates set;
2. Rank their MeSH terms in descending order according to their total frequency. We considered the first region (top 20 terms with high frequency) of Zipf curve. Four colleagues in our lab majoring in prebiotics helped to examine the candidates list and remove those that are biologically important;
3. The remaining MeSH terms from this region constituted the MeSH stop words list.

2.3. Data normalization

The normalization of MeSH terms frequency is necessary because of well-studied prebiotics can retrieve much more literature than other prebiotics and will introduce bias into the ultimate feature set of the cluster. To avoid this situation, the frequency matrix is normalized according to Eq. (1), where α $(0 \leq \alpha \leq 1)$ is a

normalization parameter controlling the correlation degree with the corpus volume. $\alpha=0$ implies no normalization and $\alpha=1$ implies complete normalization. We first build a positive prebiotics MeSH frequency matrix f_{ij} with numerical value, where each row represents a prebiotic and each column refers to a MeSH term occurring in the positive prebiotics set. M denotes prebiotics (rows). Thus, F_{ij} is the absolute MeSH term frequency while f_{ij} is the relative MeSH term frequency of each positive prebiotics.

$$f_{ij} = \frac{F_{ij}}{\left(\sum_{i=1}^{M} F_{ij}\right)^{\alpha}} \quad (0 \le \alpha \le 1) \tag{1}$$

2.3.1. Feature selection. To select features from the matrix we mentioned above, we utilized the MedMeSH summarizers algorithm, which has been applied to assign pertinent MeSH terms to describe the functionality of a group of genes.[30] MedMeSH summarizer summarizes a group of genes by filtering biomedical literature and assigning relevant keywords describing the functionality of the genes. This system constructed a P*Q co-occurrence matrix where P denotes the genes in the cluster and Q reflects the MeSH terms that were extracted from the retrieved literature. The cell value of the matrix is the frequency of each MeSH term. With this matrix, an overall score of each MeSH term can be calculated and the most influential terms will be screened to describe the functionality of this cluster. Here, we utilized this matrix to classify all the MeSH terms into two fields: Major topics and Particular topics.

2.3.2. Major Topics. Terms occurring in most prebiotics with high frequency. N denotes MeSH terms (columns). Criterion R_1: rank the MeSH terms by decreasing order of the means μ_i.

$$\mu_i = \frac{\sum_{j=1}^{N} f_{ij}}{N} \quad (i = 1, \ldots, M) \tag{2}$$

2.3.3. Particular Topics. Terms occurring in a subset of prebiotics with high frequency. σ in Eq. (3) is the ratio of the mean/standard deviation of their MeSH feature vectors. Criterion R_2: rank the MeSH terms by decreasing order of the ratios σ_i^2/μ_i.

$$\sigma_i = \sqrt{\frac{\sum_{j=1}^{N} (f_{ij} - \mu_i)^2}{N}} \quad (i = 1, \ldots, M) \tag{3}$$

All MeSH terms in the matrix are ranked in accordance with the 2 criteria described previously and assigned to an overall rank R in Eq. (4). The weight parameter w aimed at providing a summary of the cluster by balancing the major and particular topics. MeSH terms are arranged by their overall relevance ranks R in ascending order. Truncated top k MeSH terms as prebiotics summary feature set to construct normalization matrix for subsequent prediction.

$$R = wR_1 + (1-w)R_2 \quad (0 \le w \le 1) \tag{4}$$

2.4. Parameter optimization

Three key parameters, including α, w, and k, were screened for feature selection. α ranges from 0 to 1; 0 implies no normalization and 1 implies complete normalization. w also ranges from 0 to 1; 1 implies that the major topic terms dominated the feature set and 0 implies that the particular topics dominated the model. The last parameter k is the number of features we saved for the final feature set.

An exhaustive global grid search is implemented for screening the optimal parameter set. All possible combinations of the parameter values are evaluated, and the best combination is retained. Each parameter is designated with a suitable variation scope: $\alpha \in [0,1]$, step $=0.2$; $w \in [0,1]$, step $=0.1$; $k \in [200,1000]$, step $=200$ for optimal parameter screening. To evaluate the performance of the parameter sets, we employed a 5-fold cross-validation method. After repeating the simulation 100 times, the average rank of 3 positive prebiotics is used to assess the performance of each parameter set. A more accurate model is expected to rank positive prebiotics at the top of the predicted list; thus, a smaller average rank value means higher rank positions for them, which indicates a better parameter set.

2.5. Feature enrichment analysis

In the XML document, each MeSH term has two attributes that were curated by an expert: "Descriptor Name" and "Qualifier Name". "Descriptor Name" refers to the official name of the MeSH terms, and "Qualifier Name" refers to the specific related fields. For example, MeSH term *Inositol* possesses a Descriptor Name—*Inositol* and 2 Qualifier Names—*Chemistry & Pharmacology*. Thus, to perform the enrichment analysis is to extract all "Qualifier Name" under each MeSH—"Descriptor Name" for frequency calculation. Principal groups in frequency distribution bar plot can denote the property of MeSH group.

2.6. Random forest model training for comparison

Random forest is an outstanding machine learning algorithm, which can handle sparse matrix and large amount of variables. Using the MeSH term frequency of positive and negative carbohydrates as features, the Random forest models were trained and tested with 100 times repeats of 5-fold cross-validation, and the averaged areas under the receiver operating characteristic curve (ROC) (area under the curve [AUC]) were used for performance comparison in different datasets. The training and testing procedures of random forest model were implemented using "randomForest" package in R programming language.

2.7. Model evaluation and predicting novel prebiotics

We build carbohydrate prediction matrix f_{ij} according to Eq. (1) with numerical value, where each row represents a carbohydrate and each column refers to a feature. This matrix can be used to predict novel prebiotics by Eq. (5). Each carbohydrate obtained R^B as their own score denotes the ability to be potential prebiotics.

$$R^B = \sum_{i=1}^{M} \frac{f_{ij}}{R_i} \tag{5}$$

Then, we carried out 5-fold cross-validation to evaluate the predictive performance of the model. In each round, 4 randomly generated folds were used for feature selection, and the fifth fold was reserved for prediction with carbohydrates set. That is to say. There will yield 2 columns with respect to prediction set in each round: R^B score column and binary state column (1 denotes prebiotics, 0 denotes not prebiotics). Two columns yielded by this

step can produce one AUC score and after the prediction procedure was repeated 100 times. The average AUC was deployed as a measure to evaluate the prediction performance.

A model returns a vector of scores between 0 and 1 for a combined prediction profile. These scores are then mapped to a binary state indicating "prebiotics" or "non-prebiotics" by choosing a cut-off. For each combination of profiles, the existence of a prebiotic is considered positive (P) or negative (N). True (T) means that the predicted and observed categories are identical, and false (F) implies otherwise. The notations TP, FP, TN, and FN combine these labels to return the number of data points (combined prediction profile) in each category. These values are consistent with a cut-off at which carbohydrates prediction ranks are mapped onto binary predictions. The predicted scores are transformed into binary predictions using sensitivity and specificity over the entire score range. The specificity is defined as TN/(FP + TN) and the sensitivity is TP/(TP + FN). Lastly, we calculate the average specificity and average sensitivity for each round (repeat 100 times). The best cut-off point for balancing the average sensitivity and average specificity of our model is the point on the curve closest to the (0, 1) point. We deploy the corresponding cut-off to indicate potential prebiotics, which is calculated via the R package named ROCR.[43]

3. Results

3.1. Text mining framework for novel prebiotics prediction

We developed a systemic MeSH-based text mining approach to robustly predict new prebiotics. The feature selection part of our method is inspired by the MedMeSH summarizer. It is a text mining algorithm to describe the functionality of a group of genes. But our method moves further from here, it not only summarizes a cluster by using MeSH terms, but also predicts novel concepts with the same property from the cluster. In addition, MedMeSH summarizer uses fixed parameter set for gene cluster summarizing. However, we found that a fixed parameter set usually introduce many unrelated terms emerged as topic terms in our dataset, which will undermine the subsequent prediction result. To overcome this problem, we developed an exhaustive global search method to determine the optimal parameter set for our dataset of prebiotics. High-profile features were screened out and were validated by feature enrichment analysis and the ROC plot.

The workflow of prebiotics prediction is shown in Fig. 1. We first collected known prebiotics from Table 1 and carbohydrates set from the NLM MeSH tree structure in our queries to retrieve MeSH-related documents from PubMed. To construct the profile

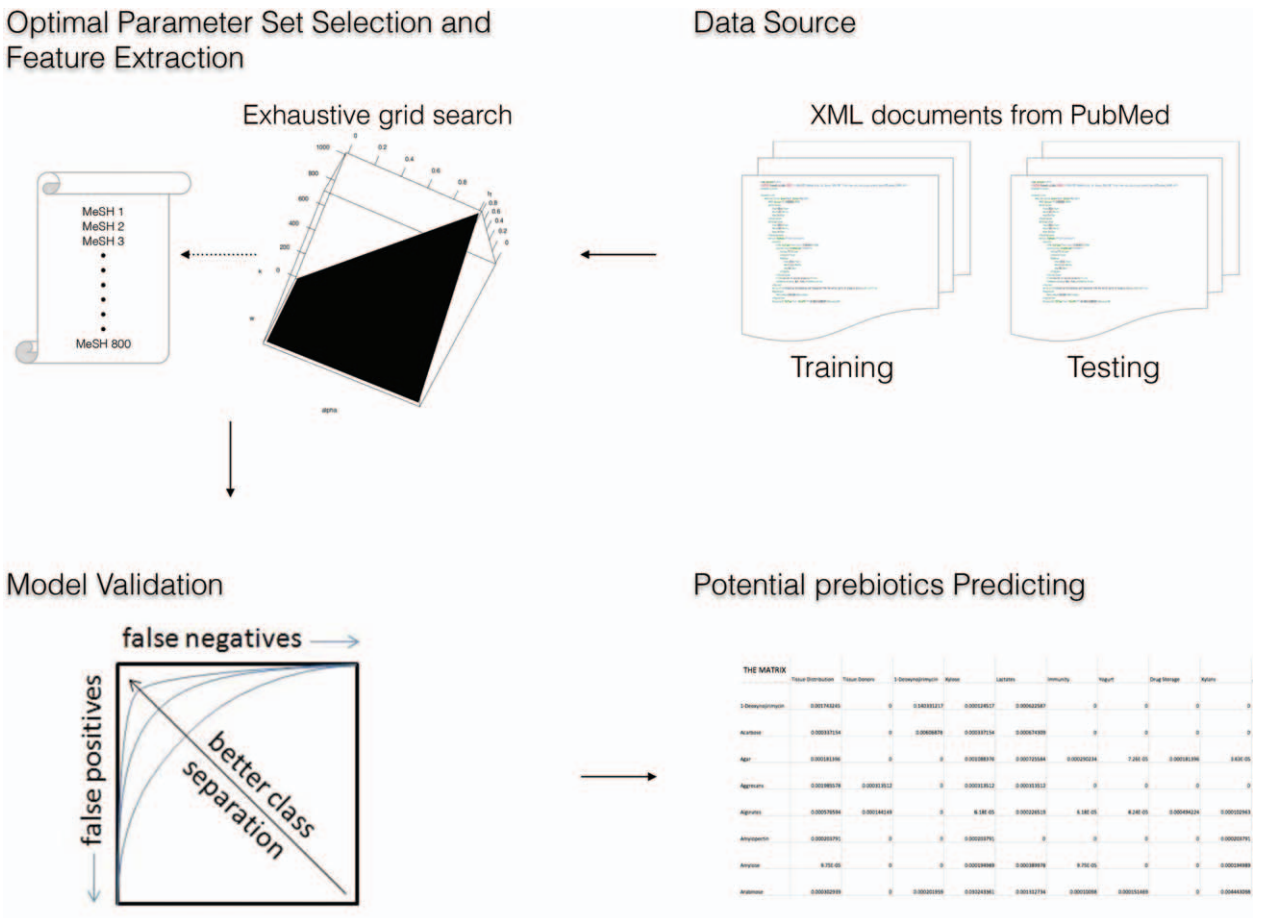

Figure 1. The framework of prediction. 1. Download PubMed XML documents of 127 carbohydrates, including 15 positive prebiotics and 112 carbohydrates. 2. Compute the optimal parameter set (α, w, and k) for the model by exhaustive grid search and assign top k features as model feature set. 3. Use ROC curve to evaluate the performance of the model. 4. Perform prediction procedure to mining novel prebiotics. ROC = receiver operating characteristic curve, XML = extensible markup language.

of each substance (prebiotics or carbohydrates), MeSH terms were extracted with respect to their retrieval literature and their frequency was calculated by Eq. (1). After that, we calculated 10 MeSH terms as stopwords, including Animals, Humans, Male, Female, Rats, Adult, Mice, Aged, Middle Aged, and Child. Those terms were removed from the corpus prior to the following analysis.

Our model primarily aims to predict new prebiotics on the basis of MeSH frequency by extracting highly representative features, which were originally employed by Kankar et al[30] in investigating the functionality of a gene group. We learn from his philosophy and adapted it to a more concrete task: novel prebiotics prediction. Unlike the previous one-fits-all solution for the gene set, we refined the feature discovery pattern by considering the unbalanced data across the feature selection procedure.

We calculated two parameters (R_1 and R_2) to identify different types of MeSH terms. R_1 is calculated by Eq. (2) that can take major topics into account whereas R_2 is produced by Eq. (3) which aims to consider particular topics. To improve feature selection step, we specified an exhaustive grid search method to determine an optimal parameter set with 5-fold cross-validation. Each parameter in the model is being traversed by certain step in the value range. Soon after that, we selected 800 features from 15 positive prebiotics that have been determined by the optimal parameter set ($\alpha=1$, $w=0.6$, $k=800$). Then, we deployed feature enrichment analysis and carbohydrates clustering to evaluate the performance of the feature set. The representative ability to prebiotics property of the feature set was very good, which also revealed the performance of the optimal parameter set on the other side. After that, we evaluated the final model and selected threshold which denote the boundary between carbohydrates with prebiotics property and without the property by ROC. According to the threshold, top 11 carbohydrates were identified as novel prebiotics. At last, we made a thorough literature investigation towards those new prebiotics.

3.2. Optimal parameter set for prebiotics prediction

Corpus volume that associated with a carbohydrate often substantially varies between positive prebiotics and carbohydrates. Well-studied prebiotics, such as inulin and fructooligosaccharides, are substantially more common in research than other carbohydrates, which introduce strong bias into the model. To balance the effect of the corpus volume, we introduced the parameter of α to control the extent of normalization of MeSH frequency. To balance the generic topics and particular topics, a weight parameter w is introduced to ensure that the final feature set could take these 2 diverse topics into full consideration. The last parameter k is the number of features we saved for the final feature set. An optimal set of parameters are crucial for precisely prediction of prebiotics, and we used an exhaustive global grid search method to determine the optimal parameter set (see Section 2).

Performance analyses of each parameter are shown in Fig. 2. $\alpha=1$ achieves best average rank regardless of the change in w, indicating that full normalization is necessary for the applied datasets, as shown in Fig. 2A. $w=0.6$ ($k=800$, $\alpha=1.0$) achieves the best average rank in Fig. 2B, suggesting that generic topics have been assigned more contribution for particular topics in

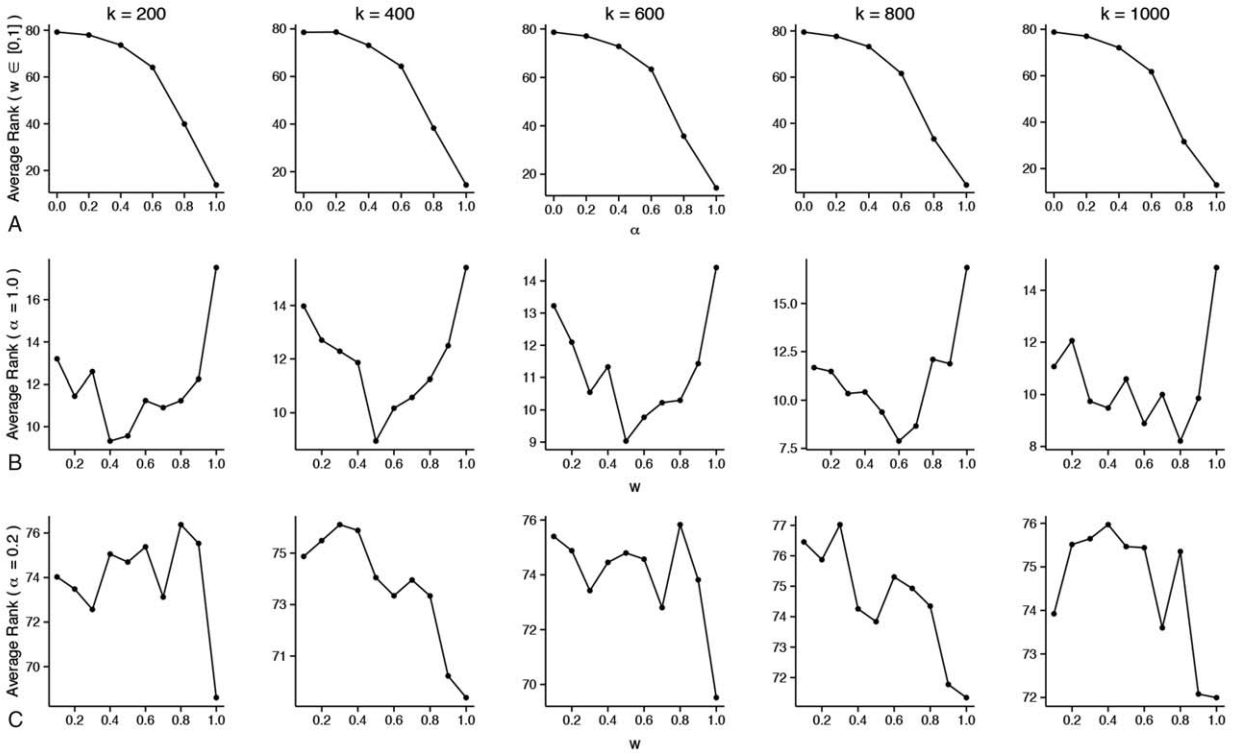

Figure 2. Exhaustive grid search for the optimal parameter set via 5-fold cross-validation. The figure describes the contribution of 3 parameters (α, w, and k) in the model. Each column adopts a fixed k. (A) described the optimal α was 1 while the optimal $w=0.6$, $k=800$ were screened in (B). After that weak normalization also has been investigated in (C).

known prebiotics summaries under full normalization circumstances. Beyond that, w under weak normalization ($\alpha = 0.2$) also has been investigated to further understand the impact of normalization (results shown in Fig. 2C). $w = 1.0$ achieves the best average rank regardless of the change in w under weak normalization, suggesting that generic topics are used to represent the entire known prebiotics summary, which indicates that full normalization is necessary when encountering unbalanced data (otherwise, the system will automatically abandon a particular instance to maintain performance). Notably, when screening the optimal parameter α, the average rank is represented by an integration of w. Finally, the optimal parameters of $\alpha = 1$, $w = 0.6$, and $k = 800$ are chosen for further analyses. After determining the optimal parameter set, two divergent topics (generic and particular) are balanced by parameter w to generate a feature summary of positive prebiotics.

3.3. Feature enrichment analysis and carbohydrates clustering

To investigate the major topics of selected features, an enrichment analysis was deployed (See Section 2). The result is shown in Fig. 3. Interestingly, >95%, >70%, and >70% correspond to metabolism, chemistry, and pharmacology, respectively, coinciding with our prior knowledge that those prebiotics usually play major roles in the metabolism of the human body due to their various chemical structure and pharmacology properties. In other words, these vital properties are concealed in the feature summary. We have innovated a method to excavate them out and effectuate them for prediction.

To examine the quality of the 800 selected features, we further conducted a hierarchical clustering method to determine if these features can excel in clustering the relevant carbohydrates adjacent to each other. Hierarchical clustering is a widely performed data analysis tool that provides dataset summaries by grouping similar observations into 1 cluster.[31] In the real-world case presented in Fig. 4, notably, the clustered carbohydrates shared a similar structure with the MeSH tree in NLM. For instance, cyclodextrins are cyclic oligosaccharides consisting of 6 α-cyclodextrins, 7 β-cyclodextrins, 8 γ-cyclodextrins, or more glucopyranose units linked by α-(1,4) bonds, which is the son node of dextrins in the MeSH tree (green block at 9 o'clock).[37] In addition to this dextrins branch, other branches, such as the Agar branch (red block at 8 o'clock), oligosaccharides branch (green block at 4 o'clock), and fructans branch (green block at 1 o'clock), etc., also achieve high similarity with the MeSH tree. These factors indicated that the features we selected may be effective in further prebiotics prediction.

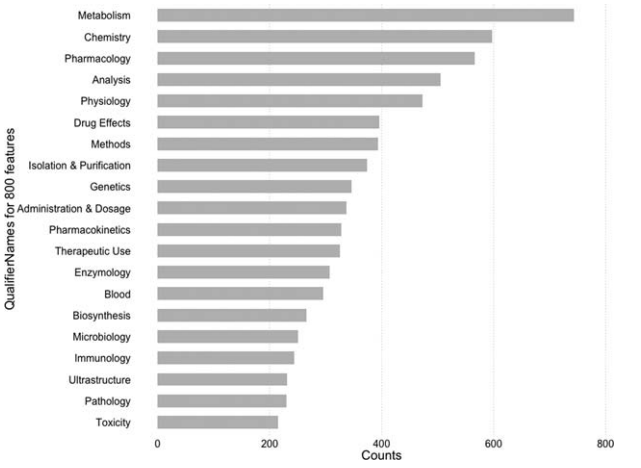

Figure 3. Feature enrichment analysis. Top 20 qualifier names were extracted from 800 features. The categories in the figure can roughly indicate the high-level concept of 800 features. Those concepts are highly correlated with real-world prebiotics chemical property.

3.4. Model evaluation and prebiotics prediction

The ROC curve is employed to model evaluation. Because of the limited number (only 15) of the positive set, we first enlarged the number of positive set to 50 to validate our method. Fifty positives contain previous 15 positive prebiotics and 35 carbohydrates which under polysaccharides node in NLM MeSH tree, their names are in S2 Table, http://links.lww.com/MD/B448. By using 50 positives and remaining 77 carbohydrates, we got our optimal parameters $\alpha = 1$, $w = 0.3$, and $k = 800$ with an average rank 11.905. The optimal parameters are utilized to deploy the model evaluation by 5-fold cross-validation ROC curve. In addition, we have performed a comparison of our method to machine learning method. The frequency matrix for machine learning is extremely sparse and there are more than 20,000 variables. Random forest algorithm can handle large amount of variables and overfitting very well. So, we decide to compare our method to random forest algorithm (see Section 2).

Figure 5A shows a 5-fold cross-validation ROC curve for the model with 50 positives. When we enlarged our positive set, our model can perform well with an AUC of 0.891. Also, the performance of our model is better than the random forest algorithm with an AUC of 0.846. After method validation step by 50 positives, we turned to 15 positive prebiotics and perform real-world ROC evaluation.

Figure 5B shows a 5-fold cross-validation ROC curve for the model with 15 positives. Surprisingly, the performance of our model is far better than random forest algorithm. It is, therefore, suggested that our method can be a good choice for the highly imbalanced data (112 negatives vs. 15 positives). We hit an AUC of 0.911 and a cut-off of 0.013 can maintain optimal balance between average specificity and average sensitivity. This cut-off helps select the corresponding rank 11, which may have prebiotics properties in the above prediction list. Those predicted novel prebiotics are presented in Table 2, and some of them have been investigated by prebiotics experts. The average specificity and sensitivity for samples were 0.876 and 0.838, respectively.

In addition to evaluating the model and predicting potential prebiotics, we also investigated related literature evidence for 11 potential prebiotics based on the original definition of prebiotics: "a prebiotic is a selectively fermented ingredient that allows specific changes, both in the composition and/or activity in the gastrointestinal microbiota, that confer benefits upon host well-being and health." Most of the predicted prebiotics are supported by the literature analysis for 2 of the 3 criteria of prebiotics (non-digestibility, fermentation, and selectivity), and there are no obvious conflicts with these criteria. Even for the most rigorous criterion (selectivity), these are also many considerable items with promising clues. For example, isomaltose has been shown to represent a prebiotic with digestion-resistant properties, raffinose is a complex 285 carbohydrate that can promote the growth of beneficial microorganisms, and acarbose is usually administered in diabetes treatment and has promising potential as a prebiotic.[40] Additionally, cyclodextrin is a saccharide that can reduce the digestion of carbohydrates and lipids. The

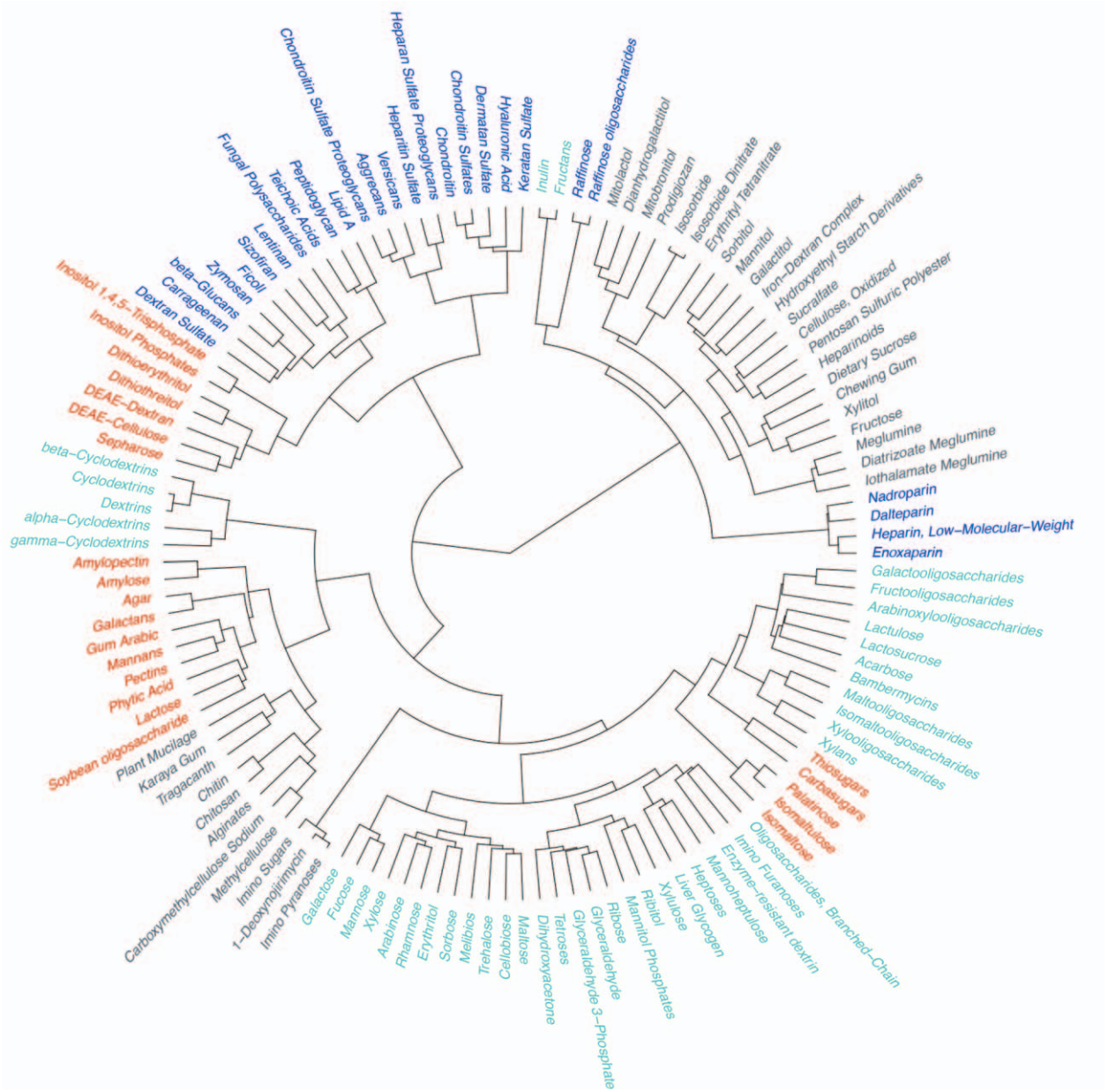

Figure 4. Hierarchical clustering of carbohydrates. If we observe a putative branch associated with the MeSH tree in NLM, we could, therefore, infer that features can be employed to predict potential prebiotics. Carbohydrates were clustered in hierarchical mode. Many branch structures are highly correlated with MeSH tree in NLM and we could therefore infer that the features have a large portion of prebiotics property, which can be employed to predict potential prebiotics in prediction step. MeSH = medical subject headings, NLM = National Library of Medicines.

derivative α-cyclodextrin is a soluble dietary fiber that possesses the ability to feed one of the *Lactococcus* sp. strains in the gastrointestinal tract,[42] whereas the other derivative (β-cyclodextrin) has been shown as an important component of low-fat foods.[43] In summary, this promising list not only shows prospective prebiotics but also demonstrated the efficacy of our model.

4. Discussion

It should be noted that our method depends on the MeSH terms. Curators typically summarize 10 to 12 MeSH terms to describe the most indexed papers from PubMed, but still there are a small portion of papers that have not been curated yet. For these overlooked papers, we suggest that keywords should be extracted manually from their abstracts and titles for information integrity. In addition, almost all text mining methods including ours are partly limited by the size and the type of the data set, and the predictive powers of our method in other data-intensive fields haven't been tested.

Prebiotics can supply vast health benefits to healthy or unhealthy people. Despite the significant demonstrated medical effect, the discovery and application of various prebiotics could not meet the growing needs of the prebiotic market simply by manually matching candidates to criteria. In an effort to improve prebiotics mining efficiency, we herein present a methodology utilizing text mining techniques to boost the variety of potential prebiotics from related literature.

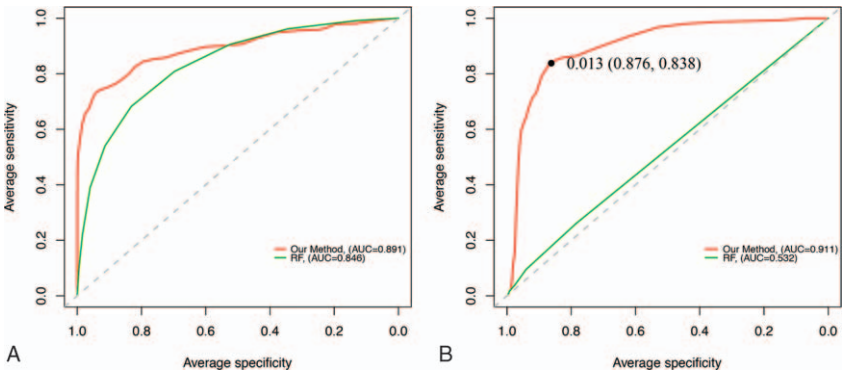

Figure 5. Cross-validation ROC analyses were used to evaluate model performance and determine the ranking threshold. (A) The ROC plot indicated our method (red) performs better than random forest (green) with 50 positives. That is to say, our method can discriminate well between known prebiotics and carbohydrates. The 45° diagonal line (dashed) indicates the theoretical plot of a test with no discrimination between known prebiotics and carbohydrates. (B) The ROC plot indicated our method (red) performs far better than random forest (green) with 15 positives. The cut-off means beyond which threshold can we deem carbohydrates possess prebiotics property. ROC = receiver operating characteristic curve.

We explored the optimal parameter set in an exhaustive grid search: each important parameter (α, w, and k) was evaluated according to a spectrum of potential values. In the parameter selection process, the parameter α is effective in corpus volume trade-off even if the volume of certain corpuses can reach a higher gulf (10^3–10^4). The parameters w and k also substantially impact the predictive performance. To more accurately determine the variation tendency for the corpus volume, we performed additional analyses to plot the average rank score against each w and k at a specific lower α ($\alpha=0.2$) after determining the optimum α (1.0). Corpus volumes in our experiment vary substantially; thus, α is intended to narrow the focus on yielding reasonable parameters. Likewise, our parameter selection process may provide a solution for other corpuses, especially those with volume-unbalanced data.

Notwithstanding inevitable practical constraints, we believe that our work is an important step in identifying more prebiotics, thereby yielding meaningful results and providing a basis for future development and experimentation. We identified critical factors affecting mining work and developed methods for characteristics selection of volume-unbalanced data to assess predictive performance. We also performed clustering measurements to evaluate the selected characteristics for known prebiotics. The ROC curve, which evaluates the model fit for an optimal parameter crew, showed that the possibility issues we identified are sufficiently consistent to create a list of potential prebiotics for further research. In a list of 11 potential prebiotics, apart from these promising specific carbohydrates, some relatively broad categories also found in it, such as xylans, fructans, and dextrins, indicate a promising field of potential prebiotics.

Overall, the MeSH-based text mining method provides a bridge between the availability of tens of thousands of studies with curated MeSH terms and the emerging functionality of prebiotics studies, which have found few prebiotics over many years. For the former, our algorithm dramatically enhances the power of discovering potential prebiotics underlying countless studies. For the latter, new candidates for potential prebiotics that are useful in prebiotics' research come to light. Regarding future directions: taken together, the thousands of studies at hand in an entire literature corpus (rather than individual studies) can assist us in other fields, such as finding bacteria that can perform certain functions or obtain food for soldiers, which may represent a niche need in future studies.

In this integrated analysis, we present new ideas and instructions that are helpful to researchers. Our results indicate that there are currently no universal parameters for the mining task and that the parameter set reported to work for a specific corpus may not be an appropriate choice for research. As we noted, an exhaustive grid search is recommended to customize

Table 2

Summary and conclusion on the prebiotic effect of 11 potential prebiotics.

Rank	Carbohydrates	Non-digestibility	Fermentation	Selectivity	References
1	Isomaltose	Yes	n.c.	Yes	[32]
2	Xylans	Yes	Yes	n.c.	[33]
3	Fructans	Yes	Yes	n.c.	[34]
4	β-Cyclodextrins	Yes	Yes	n.c.	[35]
5	Raffinose	Yes	n.c.	Yes	[36]
6	Dextrins	Yes	n.c.	n.c.	[37]
7	α-Cyclodextrins	Yes	Yes	n.c.	[38]
8	Mitobronitol	Probable	n.c.	n.c.	[39]
9	Oligosaccharides, branched-chain	Yes	Yes	n.c.	[32]
10	Acarbose	Yes	Yes	n.c.	[40]
11	Xylose	Yes	n.c.	Yes	[41]

n.c. = not clear.

the parameter set not only to determine the best parameter settings for given corpuses but also to assess their potential prediction performance. Taken together, algorithm development as a part of our study is meaningful in a wide range of biological scenarios, and the ultimate potential of the prebiotics set obtained in this study may provide novel text mining-based insights with clues in the prebiotics field. Follow-up studies are warranted to validate the findings herein; moreover, additional defined prebiotics substances and related documents will improve the model. Our text mining-based study lays the foundation for an efficient mining study for obtaining potential prebiotics, which may indicate a promising method in difficult field of prebiotics research.

Acknowledgments

We thank Miss Xin Song for critical discussion and suggestions. We would also like to acknowledge the generous funding provided by the National Basic Research Project (973 program) (2012CB518200), the General Program (31401141, 81573251, 30900830) of the Natural Science Foundation of China, the State Key Laboratory of Proteomics of China (SKLP-Y201303, SKLP-O201104 and SKLP-K201004), and the Special Key Programs for Science and Technology of China (2012ZX09102301–016).

References

[1] Roberfroid M. Prebiotics: the concept revisited. J Nutr 2007;137(Suppl 2):830S–7S.
[2] Gupta V, Lehal GS. A survey of text mining techniques and applications. J Emerg Technol Web Intell 2009;1:60–76.
[3] Swanson DR. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. Perspect Biol Med 1986;30:7–18.
[4] DiGiacomo RA, Kremer JM, Shah DM. Fish-oil dietary supplementation in patients with Raynaud's phenomenon: a double-blind, controlled, prospective study. Am J Med 1989;86:158–64.
[5] Ramadan NM, Halvorson H, Vande-Linde A, et al. Low brain magnesium in migraine. Headache 1989;29:416–9.
[6] Ferrari MD. Biochemistry of migraine. Pathol Biol 1992;40:287–92.
[7] Tsuruoka Y, Tateishi Y, Kim JD, et al. Developing a robust part-of-speech tagger for biomedical text. Lect Notes Comput Sci 2005;3746: 382–92.
[8] Tong S, Koller D. Support vector machine active learning with applications to text classification. J Mach Learn Res 2002;2:45–66.
[9] Escalante HJ, Garcia-Limon MA, Morales-Reyes A, et al. Term-weighting learning via genetic programming for text classification. Knowl-Based Syst 2015;83:176–89.
[10] Hirsch L, Saeedi M, Hirsch R. Evolving text classification rules with genetic programming. Appl Artif Intell 2005;19:659–76.
[11] Jurka TP. Maxent: an R package for low-memory multinomial logistic regression with support for semi-automated text classification. R J 2012;4:56–9.
[12] Ciarelli PM, Oliveira E. An Enhanced Probabilistic Neural Network Approach Applied to Text Classification. Prog Pattern Recog Image Anal Comput Vis Appl Proc 2009;5856:661–8.
[13] Lu ZY. PubMed and Beyond: A Survey of Web Tools for Searching Biomedical Literature. Oxford:Database; 2011.
[14] Al-Sheraji SH, Ismail A, Manap MY, et al. Prebiotics as functional foods: a review. J Funct Foods 2013;5:1542–53.
[15] Agarwala R, Barrett T, Beck J, et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 2015;43: D6–17.
[16] Bhattacharya S, Viet HT, Srinivasan P. MeSH: a window into full text for document summarization. Bioinformatics 2011;27:I120–8.
[17] Dhammi IK, Kumar S. Medical subject headings (MeSH) terms. Indian J Orthop 2014;48:443–4.
[18] Sangeetha PT, Ramesh MN, Prapulla SG. Recent trends in the microbial production, analysis and application of Fructooligosaccharides. Trends Food Sci Tech 2005;16:442–57.
[19] Lina BAR, Jonker D, Kozianowski G. Isomaltulose (Palatinose (R)): a review of biological and toxicological studies. Food Chem Toxicol 2002;40:1375–81.
[20] Vazquez MJ, Alonso JL, Dominguez H, et al. Xylooligosaccharides: manufacture and applications. Trends Food Sci Tech 2000;11: 387–93.
[21] Alander M, Matto J, Kneifel W, et al. Effect of galacto-oligosaccharide supplementation on human faecal microflora and on survival and persistence of Bifidobacterium lactis Bb-12 in the gastrointestinal tract. Int Dairy J 2001;11:817–25.
[22] Singh M, Sharma R, Banerjee UC. Biotechnological applications of cyclodextrins. Biotechnol Adv 2002;20:341–59.
[23] Johansen HN, Glitso V, Knudsen KEB. Influence of extraction solvent and temperature on the quantitative determination of oligosaccharides from plant materials by high-performance liquid chromatography. J Agr Food Chem 1996;44:1470–4.
[24] Mussatto SI, Mancilha IM. Non-digestible oligosaccharides: a review. Carbohyd Polym 2007;68:587–97.
[25] Villamiel M, Corzo N, Foda MI, et al. Lactulose formation catalysed by alkaline-substituted sepiolites in milk permeate. Food Chem 2002;76: 7–11.
[26] Kawase M, Pilgrim A, Araki T, et al. Lactosucrose production using a simulated moving bed reactor. Chem Eng Sci 2001;56:453–8.
[27] Kaneko T, Kohmoto T, Kikuchi H, et al. Effects of Isomaltooligosaccharides with different degrees of polymerization on human fecal bifidobactcria. Biosci Biotechnol Biochem 1994;58:2288–90.
[28] Eeckhaut V, Van Immerseel F, Dewulf J, et al. Arabinoxylooligosaccharides from wheat bran inhibit Salmonella colonization in broiler chickens. Poultry Sci 2008;87:2329–34.
[29] Barczynska R, Slizewska K, Jochym K, et al. The tartaric acid-modified enzyme-resistant dextrin from potato starch as potential prebiotic. J Funct Foods 2012;4:954–62.
[30] Kankar P, Adak S, Sarkar A, et al. MedMeSH summarizer: text mining for gene clusters. Siam Proc S 2002;548–565.
[31] Langfelder P, Zhang B, Horvath S. Defining clusters from a hierarchical cluster tree: the dynamic tree cut package for R. Bioinformatics 2008; 24:719–20.
[32] Gibson GR, Probert HM, Van Loo J, et al. Dietary modulation of the human colonic microbiota: updating the concept of prebiotics. Nutr Res Rev 2004;17:259–75.
[33] INTECH Open Access Publisher, da Silva AE, Oliveira EE, Egito EST, et al. Xylan, A Promising Hemicellulose for Pharmaceutical Use. 2012.
[34] Springer, Bosscher D. Fructan prebiotics derived from inulin. Prebiotics and Probiotics Science and Technology 2009;163–205.
[35] Slavin JL. Dietary fiber and body weight. Nutrition 2005;21:411–8.
[36] Su P, Henriksson A, Mitchell H. Selected prebiotics support the growth of probiotic mono-cultures in vitro. Anaerobe 2007;13:134–9.
[37] Binns N. Probiotics, prebiotics and the gut microbiota. Probiotics, Prebiotics Gut Microbiota 2013. 1–32.
[38] Delzenne NM, Cani PD. Nutritional modulation of gut microbiota in the context of obesity and insulin resistance: Potential interest of prebiotics. Int Dairy J 2010;20:277–80.
[39] Kelemen E, Jakab K, Váradi G, et al. Non-supralethal mitobronitol/cytarabine/cyclophosphamide conditioning without irradiation before bone marrow transplantation for accelerated chronic granulocytic leukemia: apparent absence of acute graft-versus-host disease. Leukemia 1993;7:939–45.
[40] Evenepoel P, Bammens B, Verbeke K, et al. Acarbose treatment lowers generation and serum concentrations of the protein-bound solute p-cresol: a pilot study. Kidney Int 2006;70:192–8.
[41] Springer, Boler BMV, Fahey GCJr. Prebiotics of plant and microbial origin. Direct-Fed Microbials and Prebiotics for Animals 2012;13–26.
[42] Pranckute R, Kaunietis A, Kuisiene N, et al. Development of synbiotics with inulin, palatinose, α-cyclodextrin and probiotic bacteria. Pol J Microbiol 2014;63:33–41.
[43] Marcolino VA, Zanin GM, Durrant LR, et al. Interaction of curcumin and bixin with β-cyclodextrin: complexation methods, stability, and applications in food. J Agr Food Chem 2011;59:3348–57.