Identification of Candidate Genes Related to Inflammatory Bowel Disease Using Minimum Redundancy Maximum Relevance, Incremental Feature Selection, and the Shortest-Path Approach

Fei Yuan,1 Yu-Hang Zhang,2 Xiang-Yin Kong,2 and Yu-Dong Cai3

1Department of Science & Technology, Binzhou Medical University Hospital, Binzhou 256603, Shandong, China
2Institute of Health Sciences, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China
3School of Life Sciences, Shanghai University, Shanghai 200444, China

Correspondence should be addressed to Fei Yuan; snowhawkyrf@outlook.com

Received 26 August 2016; Accepted 11 January 2017; Published 14 February 2017

Academic Editor: Mikihiro Fujiya

Copyright © 2017 Fei Yuan et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Identification of disease genes is a hot topic in biomedicine and genomics. However, it is a challenging problem because of the complexity of diseases. Inflammatory bowel disease (IBD) is an idiopathic disease caused by a dysregulated immune response to host intestinal microflora. It has been proven to be associated with the development of intestinal malignancies. Although the specific pathological characteristics and genetic background of IBD have been partially revealed, it is still an overdetermined disease and the blueprint of all genetic variants still needs to be improved. In this study, a novel computational method was built to identify genes related to IBD. Samples from two subtypes of IBD (ulcerative colitis and Crohn's disease) and normal samples were employed. By analyzing the gene expression profiles of these samples using minimum redundancy maximum relevance and incremental feature selection, 21 genes were obtained that could effectively distinguish samples from the two subtypes of IBD and the normal samples. Then, the shortest-path approach was used to search for an additional 20 genes in a large network constructed using protein-protein interactions based on the above-mentioned 21 genes. Analyses of the 41 genes obtained indicate that they are closely associated with this disease.

1. Introduction

Inflammatory bowel disease (IBD) is a common systemic disease that involves the intestinal tissue [1]. It usually refers to chronic conditions that lead to intestinal inflammation and lesions. With the gradual development of inflammation, the intestinal walls become swollen, inflamed, and ulcerogenic [2]. Due to such lesions, several classical symptoms have been considered to be diagnostic indicators. Abdominal pain or cramping, diarrhea multiple times per day, and bloody stools are all classical symptoms of IBD [3]. Such severe symptoms are induced by violent unhealthy inflammation reactions and lesions in the intestinal tissue. Additionally, several complications outside the digestive tract may also be induced by IBD. Mouth sores and skin problems have both been reported in IBD patients [4]. Furthermore, arthritis is also related to IBD, as well as eye problems [5, 6].

As we have mentioned above, several complications have been identified in IBD patients. Such severe complications and related chronic characteristics strongly increase the risk of death [4–6]. In 2013 alone, thousands of people in the world died from IBD [7]. Additionally, IBD has been proven to be associated with colorectal cancer, with a high mortality. Apart from the risk of death, IBD is a lifetime disease, and life with IBD can be quite challenging. The complications associated with IBD and disease relapse severely impact the quality of life [8]. Therefore, the prevention, diagnosis, and treatment of IBD are quite crucial. It is known that IBD is a widespread disease that can develop at any stage of life. However, the disease usually initiates during the teenage years or the early
adulthood of the patients [8]. As we mentioned above, genetic factors participate in the initiation and progression of IBD [9, 10]. Therefore, people with a family history of IBD are at least ten times more likely to suffer from it. Racial factors also contribute to the morbidity of IBD [11].

Although IBD is a very severe and widespread disease, the essential mechanism behind the disease has not been demonstrated clearly. Most people believe that some types of exogenous materials trigger the initiation of inflammation [12, 13]. However, genetic factors also may contribute to the progression of such disease. Several specific genes have been linked to IBD. Pathogenic genes such as IL23R and IL12B play a crucial role in the intestinal immune system, which may induce the initiation of IBD [14, 15]. Several transcriptional factors also contribute to disease progression. The transcription factor NKX2-3 regulates the correct localization of lymphocytes and may further contribute to the immune response in intestinal tissue that induces IBD [16]. Several genes such as ZNF365 and PTGER4 show diversity in different subtypes of IBD and contribute to IBD through their respective methods and pathways [17, 18].

As mentioned above, IBD has several subtypes. Basically, there are two main clinical classifications of IBD: Crohn’s disease and ulcerative colitis [17]. Both classifications share the basic symptoms of IBD. However, Crohn's disease can occur anywhere along the digestive tract and typically appears as “skip lesions” between healthy areas [19], while the other type, ulcerative colitis, only involves the colon and rectum. Inflammation and ulcers typically affect only the innermost lining in these areas, with more superficial lesions than those with Crohn's disease [20]. Apart from the differences in clinical symptoms, genetic diversity is also observed between Crohn's disease and ulcerative colitis. Although they share most of the disease-causing genes, genes like ATG16L1, PTGER4, IRGM, and NOD2 have been proven to be specifically related to Crohn's disease but independent with ulcerative colitis [18, 21–23]. The roles of genes such as SLC22A5, ZNF365, and PTPN2 in ulcerative colitis are still unclear, even though they have been proven to be strongly related to Crohn's disease [24, 25].

Because genetic factors have been shown to be related to IBD and its specific subtypes, we developed a new computational method to screen differential expressing genes among different clusters based on a database for Crohn's disease and ulcerative colitis. From the Gene Expression Omnibus (GEO), we obtained the gene expression profiles (information often used to deduce and understand gene functions) for 59 Crohn's disease, 26 ulcerative colitis, and 42 normal samples. Each sample was represented using the expression levels of 12,754 genes. Two feature selection methods, minimum redundancy maximum relevance (mRMR) and incremental feature selection (IFS) [26], and a basic machine learning algorithm, sequential minimal optimization (SMO) [27, 28], were adopted to analyze the gene expression profiles and extract 21 promising candidate genes that could be used to distinguish the samples from the two subtypes of IBD and the normal samples; that is, they may be related to IBD. Furthermore, based on these 21 genes, the shortest-path (SP) approach was employed to identify additional 20 genes in a network constructed using protein-protein interaction (PPI) information. It was concluded that the 41 (21 + 20) genes obtained are closely associated with IBD and can be used to clearly distinguish healthy people from those who have IBD and to identify the subtypes of IBD.

2. Materials and Methods

2.1. Dataset. We downloaded the gene expression profiles of 59 Crohn's disease, 26 ulcerative colitis, and 42 normal samples from GEO under accession number GSE3365 [29]. The expression levels of 12,754 genes were measured using an Affymetrix Human Genome U133A Array. The gene expression profiles were quantile normalized. Each sample was represented using the expression levels of 12,754 genes; that is, each sample was encoded into a 12754-D vector. These features/genes were analyzed to identify the genes that can best discriminate the samples from these three different classes.

2.2. mRMR Method. It is known that some genes can effectively help us discriminate the samples from the three different classes mentioned in Section 2.1, while others offer few or no contributions. To identify these genes, the mRMR method, proposed by Peng et al. [26], was adopted to analyze the gene expression data. The mRMR method employed two criteria, Max-Relevance and Min-Redundancy, to analyze the features. Using the Max-Relevance criterion, the MaxRel feature list can be obtained, in which features are sorted by measuring the relevance between them and sample class labels. Features with high relevance receive high ranks, whereas those with low relevance receive low ranks. It is clearly seen that the rank of a feature in the MaxRel feature list indicates its single contribution to classification. Furthermore, another list, namely, the mRMR feature list, was created using both Max-Relevance and Min-Redundancy criteria. The rank of a feature in this list is determined using the relevance between it and sample class labels and the redundancies between it and the features listed before it. The MaxRel feature list and mRMR feature list in this study were formulated as follows:

MaxRel features list is

\[ F_{\text{MaxRel}} = \left[ f_1^M, f_2^M, \ldots, f_N^M \right] \]

mRMR features list is

\[ F_{\text{mRMR}} = \left[ f_1^m, f_2^m, \ldots, f_N^m \right], \]

where \( N \) represents the total number of features. Many investigators have used the mRMR method to analyze various complicated biological systems [30, 31], and it is deemed to be a useful tool for extracting important information from a complicated system. Readers can refer to Peng et al.'s paper [26] for the detailed procedures and principle of this method.

2.3. Prediction Engine. SMO is a type of support vector machine that uses Platt's sequential minimal optimization
algorithm to train and optimize the support vector classifier. The kernels can be polynomial or Gaussian [27, 28]. For implementing our method, we employed the classifier SMO implemented in Weka [32] as the prediction engine.

2.4. Tenfold Cross-Validation. Tenfold cross-validation [33] is a type of cross-validation method that is widely used to examine the performance of a classifier on a given dataset. The given dataset is randomly and equally divided into ten partitions. Samples in each partition are singled out in turn as the test data, while other samples are used to train the classifier. Compared to the jackknife test [34, 35], another popular cross-validation method, this method involves a lower amount of computational time and always yields similar results. Thus, it was used in this study for evaluating the performance of the current prediction engine.

2.5. IFS Method. Using the mRMR method, features/genes were sorted and listed in the MaxRel feature list and mRMR feature list. Because the MaxRel feature list sorted features/genes by only measuring their own contributions to classification, the combination of some features/genes with features/genes by only measuring their own contributions to mRMR feature list. Because the MaxRel feature list sorted the performance of the current prediction engine.

2.6. Network Construction from PPI Information. The optimal gene set \( G_{\text{optimal}} \) containing some genes closely related to IBD can be obtained using the mRMR and IFS methods. To further mine for other related genes, we constructed a large network from the PPI data and searched for additional candidate genes in the network.

To construct the network, we downloaded the file “protein.links.v9.1.txt.gz” containing the PPI information from STRING (Search Tool for the Retrieval of Interacting Genes/Proteins, version 9.1, http://www.string-db.org/), from which the human PPI data were extracted by identifying lines starting with “9606.” A total of 2,425,314 human PPIs involving 20,770 proteins represented using Ensembl IDs were obtained. According to STRING (http://string-db.org/) [36, 37], these PPIs are derived from the following sources: (i) genomic context, (ii) high-throughput experiments, (iii) (conserved) coexpression, and (iv) previous knowledge. Thus, the obtained PPIs contained actual PPIs validated using experiments and predicted PPIs, suggesting that they can be used to widely measure the physical and functional relationships between proteins. Each PPI contained two proteins represented using Ensembl IDs and one score that indicates the strength of the interaction with a range between 150 and 999. The constructed network had 20,770 proteins as nodes. Two nodes were adjacent if and only if the corresponding proteins comprise an interaction that is contained in the 2,425,314 human PPIs. Furthermore, the interaction score was also added to the network. Each edge was assigned a weight defined to be 1,000 minus the corresponding interaction score.

2.7. SP Approach for Searching for Additional Candidates. Network method is an important type of approaches for investigation of disease genes, such as methods based on guilt-by-association (GBA) [38–40] and Random Walk with Restart (RWR) [41–43]. This section proposed another network method for identifying novel disease genes.

It has been elaborated in some previous studies [44–46] that two proteins in an interaction are more likely to share similar functions. It can be induced that the interactive proteins of the proteins encoded by genes in \( G_{\text{optimal}} \) are also related to IBD. Furthermore, if we consider a series of proteins \( p_1, p_2, \ldots, p_s \) such that the consecutive proteins comprise a PPI with a high score and \( p_1, p_s \) are proteins encoded by genes in \( G_{\text{optimal}} \), \( p_2, p_3, \ldots, p_{s-1} \) may also be related to IBD. From the construction of the network mentioned in Section 2.6, the corresponding nodes of \( p_1, p_2, \ldots, p_s \) may comprise a shortest path connecting \( p_1 \) and \( p_s \). Therefore, for any two genes in \( G_{\text{optimal}} \), we searched the shortest path connecting these two genes, thereby collecting a number of shortest paths. Because the endpoints of these paths represented proteins encoded by genes in \( G_{\text{optimal}} \) genes on these paths may be related to IBD. Thus, we extracted inner nodes on the obtained shortest paths and their corresponding genes can be obtained. To identify novel genes related to IBD, genes in \( G_{\text{optimal}} \) were excluded from the obtained genes. The remaining genes were called shortest-path genes for convenience. To identify these shortest-path genes, a measurement, namely, the betweenness [47], was recorded for each shortest-path gene, and it was defined to be the number of shortest paths containing the shortest-path gene.

Because some nodes occupied general hubs in the constructed network, the corresponding genes may always be selected even if we searched for the shortest path connecting any pair of randomly selected genes; some of these genes may be selected as the shortest-path genes obtained as described above. In fact, they have few or no associations with IBD. Thus, a permutation test is necessary to control for this type of gene. The procedures used are as follows:

(1) Randomly produce 1,000 gene sets, say \( G_1, G_2, \ldots, G_{1000} \), where the size of each set is the same as that of \( G_{\text{optimal}} \).

(2) For each \( G_i \), search for all the shortest paths connecting any pair of genes in \( G_i \) and count the betweenness of the shortest-path gene based on these paths.

(3) A total of 1,000 betweenness scores on 1,000 randomly produced gene sets can be obtained for each
shortest-path gene. After comparing the betweenness on $G_{optimal}$, we calculate another measurement, the permutation FDR, for each shortest-path gene, which is defined to be "the number of betweenness scores on randomly produced gene sets that was larger than that on $G_{optimal}$"/1000.

(4) Because it is implied that shortest-path genes with high permutation FDRs are general hubs in the network and not specific to IBD, those with permutation FDRs larger than or equal to 0.05 are excluded. The remaining genes are termed candidate genes.

To select genes with core relationships with IBD from the candidate genes, the human PPIs and their interaction scores were directly used. For each candidate gene $g$, we checked the scores of the interactions between $g$ and genes in $G_{optimal}$ and selected the maximum value among them as the maximum interaction score of $g$. If a candidate gene has a high maximum interaction score, this suggests that it is highly related to at least one gene in $G_{optimal}$, indicating that it is more likely to be related to IBD. As 900 is set to be the threshold of the highest confidence cutoff in STRING, we also set 900 as the threshold for the maximum interaction score; that is, genes with maximum interaction scores no less than 900 were finally selected as the candidate genes in this study.

3. Results and Discussion

3.1. Results of the mRMR and IFS Methods. The mRMR method was executed on the dataset containing 59 Crohn’s disease, 26 ulcerative colitis, and 42 normal samples, and each sample was represented using the expression levels of 12,754 genes, thereby yielding the MaxRel feature list and mRMR feature list, which are provided in Supplementary Material I in Supplementary Material available online at https://doi.org/10.1155/2017/5741948.

To extract the optimal gene sets for discriminating the samples from two subtypes of IBD and normal samples, the IFS method was used with the mRMR feature list obtained using the mRMR method and SMO as the prediction engine. To reduce computational time and account for the fact that genes with important contributions for discriminating samples from two subtypes of IBD and normal samples are few in number, we only investigated the first 2,000 feature sets. According to the procedures of the IFS method, each feature set can yield four accuracies: three accuracies for the three classes and the total prediction accuracy. All of these are provided in Supplementary Material II. Furthermore, an IFS curve was plotted by representing the total prediction accuracy along $y$-axis and the size of the feature set, that is, the number of features participating in the classification, along $x$-axis, as shown in Figure 1. It can be seen that the highest total prediction accuracy was 97.64% using the 1170th feature set. The corresponding accuracies for the three classes were 100%, 92.31%, and 97.62%, respectively. Although the accuracies were quite good, the involved features/genes were too many in number, which is not realistic. By carefully checking the IFS curve shown in Figure 1, we observe a sharp increasing trend with more and more features participating in the classification at the beginning of the curve with a rather high total prediction accuracy (93.70%) using the 21st feature set. Then the curve is unstable; increasing trends and decreasing trend occur in succession. Thus, we believe that the first 21 features in the mRMR feature list are more important for discriminating the samples from two subtypes of IBD and normal samples than others and set the optimal gene set $G_{optimal}$ to be the 21st feature set. These 21 genes are listed in Table I. The associations between these 21 genes and IBD are elaborated in Section 3.4. However, some important IBD-related genes may not be omitted using the mRMR and IFS methods. Based on these genes, the SP approach was applied to discover additional genes related to IBD, which is described in the following sections.

3.2. Shortest-Path Genes. As mentioned in Section 3.1, 21 genes were obtained and deemed to be important for discriminating the samples from the two subtypes of IBD and the normal samples. To further identify more candidate genes, we constructed a large network, as described in Section 2.6. These 21 genes were mapped to 20 genes in the network. We searched for all shortest paths connecting any pair of 20 genes, resulting in 190 paths. The graph of these 190 paths is shown in Figure 2, where we can see that there are 110 Ensembl genes on these paths other than the 21 genes obtained in Section 3.1. By mapping to their gene symbols, we obtained 107 shortest-path genes. These genes and their betweenness are listed in Supplementary Material III.

3.3. Additional Candidate Genes. According to Section 2.7, a permutation test was executed to exclude general genes in the network. The obtained permutation FDRs of 107 shortest-path genes are also provided in Supplementary Material III. By setting the threshold of the permutation FDR to be 0.05, 57 candidate genes were obtained, which are listed in Supplementary Material IV.

To select the core genes among the 57 candidate genes, the maximum interaction score of each candidate gene was calculated. These values are also provided in Supplementary Material IV. The threshold of the maximum interaction score...
was set to 900, resulting in 20 candidate genes that are listed in Table 2.

### 3.4. Analysis of Candidate Genes

Based on feature analysis of 59 Crohn’s disease, 26 ulcerative colitis, and 42 normal samples, we obtained 21 genes, listed in Table 1, which may be related to IBD and can help distinguish healthy people from those who have two subtypes of IBD. Furthermore, according to the above 21 genes and the SP approach, we obtained additional 20 candidate genes, listed in Table 2. These genes are also thought to be related to IBD. This section provides some evidence for this claim.

We combined two candidate gene sets and analyzed the biological meaning behind them using Functional Annotation Bioinformatics Microarray Analysis (DAVID) (version 6.7, https://david.ncifcrf.gov/) [48]. The obtained results are provided in Supplementary Material V. According to the results yielded by DAVID, crucial gene ontology (GO) terms and KEGG pathways like hsa04660 (T cell receptor signaling pathway), GO: 0001775 (cell activation), and GO: 0045449 (regulation of transcription) were screened out to be enriched by 41 candidate genes. In addition, the results also gave clues for clustering 41 candidate genes into some groups, which provided convenience for analyzing candidate genes.

#### 3.4.1. Candidate Genes Contributing to T Cell Receptor Signaling Pathway (hsa04660)

As mentioned above, IBD is a severe disease induced by inflammation reactions [1]. Considering the core regulatory role of T cells in immune system, it is quite reasonable that various candidate genes contribute to such pathway. Based on SP approach, we identified a specific gene FOS. It is also a tumor-associated gene, which encodes a leucine zipper protein that can dimerize with proteins of the JUN family, thereby forming the transcription factor complex AP-1 [49]. Related to crucial pathways such as NF-kB and MAPK, FOS is quite significant in inflammation initiation, especially in the digestive tract [50, 51]. Another calcium-associated gene PLCG1 was also discovered. PLCG1 participates in the intracellular transduction of receptor-mediated tyrosine kinase activators and may participate in the inflammation reaction through a specific function [52, 53]. LCK (also known as p56lck) is another predicted IBD-related gene that encodes a functional tyrosine kinase. Similar to ZAP70, LCK also regulates the metabolism and maturation of T cells and may further regulate the inflammation process [54, 55]. In terms of IBD, LCK has been reported to be associated with ulcerative colitis but not with Crohn’s disease [56]. Our predicted gene ZAP70 is a protein tyrosine kinase participating in the development and activation of T cells [57, 58].

<table>
<thead>
<tr>
<th>GO term/KEGG pathway ID</th>
<th>Description</th>
<th>Rank</th>
<th>Gene symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>hsa04660</td>
<td>T cell receptor signaling pathway</td>
<td>3</td>
<td>CD247</td>
<td>CD247 molecule</td>
</tr>
<tr>
<td></td>
<td></td>
<td>19</td>
<td>CD4</td>
<td>CD4 molecule</td>
</tr>
<tr>
<td>GO: 0001775</td>
<td>Cell activation</td>
<td>5</td>
<td>PF4</td>
<td>Platelet factor 4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>ZNF207</td>
<td>Zinc finger protein 207</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8</td>
<td>EGR3</td>
<td>Early growth response 3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>SLTM</td>
<td>SAFB-like, transcription modulator</td>
</tr>
<tr>
<td></td>
<td>Regulation of transcription</td>
<td>14</td>
<td>CNOT8</td>
<td>CCR4-NOT transcription complex, subunit 8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>13</td>
<td>THIL (NELFCD)</td>
<td>Negative elongation factor complex member C/D</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9</td>
<td>HMGBI</td>
<td>High mobility group 1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>12</td>
<td>UBE2I</td>
<td>Ubiquitin-conjugating enzyme E2I</td>
</tr>
<tr>
<td>GO: 0007243</td>
<td>Protein kinase cascade</td>
<td>2</td>
<td>MARK2</td>
<td>MAP/microtubule affinity-regulating kinase 2</td>
</tr>
<tr>
<td>GO: 0031226</td>
<td>Intrinsic to plasma membrane</td>
<td>6</td>
<td>FOLR1</td>
<td>Folate receptor 1 (adult)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>18</td>
<td>SLC22A4</td>
<td>Solute carrier family 22 (organic cation/zwitterion transporter), member 4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10</td>
<td>LEPROT</td>
<td>Leptin receptor overlapping transcript</td>
</tr>
<tr>
<td></td>
<td></td>
<td>16</td>
<td>CLEC1B</td>
<td>C-type lectin domain family 1, member B</td>
</tr>
<tr>
<td>GO: 0006915</td>
<td>Apoptosis</td>
<td>21</td>
<td>RHOT2</td>
<td>ras homolog family member T2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7</td>
<td>BLCAP</td>
<td>Bladder cancer associated protein</td>
</tr>
<tr>
<td></td>
<td>Non-grouped genes</td>
<td>20</td>
<td>ANXA11</td>
<td>Annexin A11</td>
</tr>
<tr>
<td></td>
<td></td>
<td>15</td>
<td>OGT</td>
<td>O-linked N-acetylglucosamine (GlcNAc) transferase</td>
</tr>
<tr>
<td></td>
<td></td>
<td>17</td>
<td>USP1</td>
<td>Ubiquitin specific peptidase like 1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>11</td>
<td>HIST1H2AC</td>
<td>Histone cluster 1, H2ac</td>
</tr>
</tbody>
</table>

a: this column indicates the ranks of related features in the mRMR feature list.
Figure 2: The graph consisting of 190 shortest paths connecting any two genes in the optimal gene set. The yellow diamonds represent genes in the optimal gene set. The blue diamonds represent shortest-path genes. The numbers on the edges represent the edge weights in the network.
<table>
<thead>
<tr>
<th>GO term/KEGG pathway ID</th>
<th>Description</th>
<th>Gene symbol</th>
<th>Ensembl ID</th>
<th>Description</th>
<th>Betweenness</th>
<th>Permutation FDR</th>
<th>Maximum interaction score</th>
<th>Most related gene in the optimal gene set</th>
</tr>
</thead>
<tbody>
<tr>
<td>hsa04660</td>
<td>T cell receptor signaling pathway</td>
<td>FOS</td>
<td>ENSP00000306245</td>
<td>FBFI murine osteosarcoma viral oncogene homolog</td>
<td>20</td>
<td>0.035</td>
<td>950</td>
<td>CD4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PLCG1</td>
<td>ENSP00000244007</td>
<td>Phospholipase C, gamma 1</td>
<td>19</td>
<td>0.022</td>
<td>927</td>
<td>CD4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>LCK</td>
<td>ENSP00000337825</td>
<td>LCK protooncogene, Src family tyrosine kinase</td>
<td>21</td>
<td>0.02</td>
<td>999</td>
<td>CD4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ZAP70</td>
<td>ENSP00000264972</td>
<td>Zeta-chain (TCR) associated protein kinase 70 kDa</td>
<td>16</td>
<td>0.016</td>
<td>999</td>
<td>CD247</td>
</tr>
<tr>
<td>GO:000775</td>
<td>Cell activation</td>
<td>YWHAZ</td>
<td>ENSP00000309503</td>
<td>Tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, zeta</td>
<td>19</td>
<td>0.009</td>
<td>962</td>
<td>MARK2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TLR4</td>
<td>ENSP00000363089</td>
<td>Toll-like receptor 4</td>
<td>19</td>
<td>&lt;0.001</td>
<td>970</td>
<td>HMGB1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>F2</td>
<td>ENSP00000308541</td>
<td>Coagulation factor II (thrombin)</td>
<td>19</td>
<td>&lt;0.001</td>
<td>953</td>
<td>PF4</td>
</tr>
<tr>
<td>GO:0045449</td>
<td>Regulation of transcription</td>
<td>HCFC1</td>
<td>ENSP00000309535</td>
<td>Host cell factor 1</td>
<td>36</td>
<td>&lt;0.001</td>
<td>997</td>
<td>OGT</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CNOT1</td>
<td>ENSP00000320949</td>
<td>CCR4-NOT transcription complex, subunit 1</td>
<td>6</td>
<td>0.004</td>
<td>998</td>
<td>CNOT8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CNOT4</td>
<td>ENSP00000354673</td>
<td>CCR4-NOT transcription complex, subunit 4</td>
<td>6</td>
<td>0.008</td>
<td>987</td>
<td>CNOT8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TRAK1</td>
<td>ENSP00000328998</td>
<td>Trafficking protein, kinesin binding 1</td>
<td>19</td>
<td>&lt;0.001</td>
<td>946</td>
<td>OGT</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HDAC1</td>
<td>ENSP00000362649</td>
<td>Histone deacetylase 1</td>
<td>19</td>
<td>0.031</td>
<td>967</td>
<td>UBE2I</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BTG1</td>
<td>ENSP00000256015</td>
<td>B-cell translocation gene 1, anti-proliferative</td>
<td>13</td>
<td>0.006</td>
<td>981</td>
<td>CNOT8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RUNX1</td>
<td>ENSP00000303035</td>
<td>Runt-related transcription factor 1</td>
<td>19</td>
<td>&lt;0.001</td>
<td>927</td>
<td>SLC22A4</td>
</tr>
<tr>
<td>GO:0032126</td>
<td>Intrinsic to plasma membrane</td>
<td>THBD</td>
<td>ENSP00000366307</td>
<td>Thrombomodulin</td>
<td>19</td>
<td>&lt;0.001</td>
<td>985</td>
<td>PF4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>FASLG</td>
<td>ENSP00000356694</td>
<td>Fas ligand (TNF superfamily, member 6)</td>
<td>19</td>
<td>&lt;0.001</td>
<td>985</td>
<td>EGR3</td>
</tr>
<tr>
<td>GO:0042127</td>
<td>Regulation of cell proliferation</td>
<td>STK11</td>
<td>ENSP00000324856</td>
<td>Serine/threonine kinase 11</td>
<td>19</td>
<td>&lt;0.001</td>
<td>986</td>
<td>MARK2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S100A6</td>
<td>ENSP00000357708</td>
<td>S100 calcium binding protein A6</td>
<td>19</td>
<td>&lt;0.001</td>
<td>987</td>
<td>ANXA11</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SERPINE1</td>
<td>ENSP00000223095</td>
<td>Serpin peptidase inhibitor, clade E (nexin, plasminogen activator inhibitor type 1), member 1</td>
<td>18</td>
<td>0.034</td>
<td>933</td>
<td>PF4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>VEGFC</td>
<td>ENSP00000280193</td>
<td>Vascular endothelial growth factor C</td>
<td>19</td>
<td>&lt;0.001</td>
<td>919</td>
<td>PF4</td>
</tr>
</tbody>
</table>
ZAP70 has been reported to be associated with a specific subtype of IBD, Crohn’s disease, but not ulcerative colitis [59]. Therefore, the expression level of ZAP70 can be a useful biomarker for distinguishing different subtypes of IBD.

While based on the mRMR and IFS method, we also identified a group of candidate genes. Among them, CD247 (rank 3 in the mRMR feature list) and CD4 (rank 19 in the mRMR feature list) are both crucial genes for T cells and have been confirmed to further regulate the inflammation reaction [60, 61]. Our predicted gene CD4 is characteristic expressed in IBD [62]. However, CD4 has also been reported as a differentially expressed gene in Crohn’s disease and ulcerative colitis, and it may further serve as a new biomarker for distinguishing these two diseases [63]. Based on our functional clustering, various screened and predicted genes also are enriched in a similar GO term, GO: 0042101 which describes T cell receptor complex as a cellular component, validating the enrichment of T cell receptor signaling pathway of our screened out IBD associated genes.

3.4.2. Candidate Genes Contributing to Cell Activation (GO: 0001775). In Table 2, a highly conserved monooxygenase-associated protein YWHAZ was identified as a functional protein in intestinal bowel disease. Such gene is a crucial housekeeping gene that has been proven to be a suitable normalizer for bowel inflammation and cancer [60]. As a functional factor of innate immune response, which is also crucial in intestinal tissues, TLR4 is predicted to be associated with IBD. TLR4 has been reported as a crucial factor in the innate immune barrier of the intestine [61]. Such factors can be activated by specific factors (FFA, etc.) and further induce the initiation of IBD [64, 65]. Another thrombin-associated gene, F2 (coagulation factor II) was also identified by the SP approach. F2 and THBD are both coagulation-associated genes. The coagulation process is reported to be associated with Crohn’s disease but not with ulcerative colitis, which reflects the differences between various subtypes of IBD [66]. There are two major subtypes of IBD: Crohn’s disease and ulcerative colitis. Some candidate genes yielded by mRMR and IFS methods may distinguish these two subtypes. PF4 (rank 5 in the mRMR feature list), a crucial diagnostic biomarker for IBD, has been clearly reported to be overexpressed in Crohn’s disease and thought that it does not play a clear role in ulcerative colitis [67, 68]. PF4 can also separate IBD from normal inflammation, which is crucial for diagnosis [69].

3.4.3. Candidate Genes Contributing to Regulation of Transcription (GO: 0045449). Among the 41 candidate genes, quite a lot of genes contribute to the regulation of transcription, implying the complicated endogenous pathological factors of IBD on multiple levels. Based on mRMR and IFS methods, the candidate gene ZNF207 (rank 1 in the mRMR feature list), which is a specific microtubule-associated zinc finger protein, may regulate the inflammation of IBD [70]. As a regulator of mitotic chromosome alignment, ZNF207 has been reported to be related to another type of inflammation disorder, chronic obstructive pulmonary disease (COPD). Since both COPD and IBD are localized inflammation involving the mucosal tissue, ZNF207 as our candidate gene may also contribute to inflammatory bowel disease [71]. As a T cell regulator, EGR3 (rank 8 in the mRMR feature list) was also identified. As a member of the EGR family, EGR3 may be a crucial transcriptional factor for T cells, with high similarity with EGR2 [72]. SLTM (rank 4 in the mRMR feature list) acts as a general inhibitor of transcription that eventually leads to apoptosis via the regulation of telomere [73]. Because IBD is associated with abnormal cell death, SLTM may participate in IBD through the regulation of the apoptosis of intestinal cells [74]. CNOT8 (rank 14 in the mRMR feature list) is a significant predicted gene that interacts with BTG, the regulator of the cell cycle, especially in B cells [75]. Therefore, CNOT8 may indirectly participate in the intestinal inflammation reaction [75, 76]. THIL (rank 13 in the mRMR feature list), as a negative elongation factor complex member CI/D (NELFCD), promotes the proliferation of intestinal cells and has been proved to induce carcinoma progression [77]. As a regulator of B cells, HMGB1 (rank 9 in the mRMR feature list) and its homolog HMGB2 constitute a complex that is differentially expressed in Crohn’s disease and ulcerative colitis [78, 79]. Such a complex has also been reported as a new marker of IBD and may be a sensitive marker of mucosal inflammation [80]. As we have mentioned above, our predicted gene CD4 is characteristically expressed in IBD [62]. However, CD4 has also been reported as a differentially expressed gene in Crohn’s disease and ulcerative colitis, and it may further serve as a new biomarker for distinguishing these two diseases [63]. UBE2I (rank 12 in the mRMR feature list) also regulates the proliferation of intestinal cells [81]. Unlike FOLR1, which we will analyze below, UBE2I is a major part of the SUMO ligases and further promotes the proliferation of intestinal cells via multiple means even under pathological conditions [81, 82].

For the candidate genes obtained by the SP approach, HCF1 is a functional nuclear activator. As a unique cleavage signal, it has been reported to be associated with cell cycle regulation and may have a specific function in tumorigenesis [83, 84]. As a part of the CCR4-NOT complex, CNOT1 is a crucial immune associated gene that is a major cellular mRNA deadenylase and has been reported to participate in several processes related to immune reactions [85]. Regulated by the CCR4-NOT complex, a crucial microRNA, miR155, has been reported to be directly associated with inflammation, which may further reveal the tight connection between CNOT1 and the inflammation reaction [86, 87]. Such functional genes may also participate in the initiation of inflammation and tumors. CNOT4 is also a part of the CCR-NOT complex, and CNOT4 may act similarly to CNOT1 and contribute to the regulation of the immune reaction [85]. As a functional factor of innate immune response, which is also crucial in intestinal tissues, TRAK1 is a regulatory gene that may be related to endosome-to-lysosome trafficking and EGF-EGFR interaction [88]. Such an EGF-EGFR interaction is definitely associated with the initiation of bowel inflammation [89]. The candidate gene HDAC1 regulates the acetylation of specific genes and further participates in the regulation of corresponding functions [90]. Gene acetylation and deacetylation are functional regulatory methods for
cell metabolism, which have been identified in IBD [91–93]. Therefore, HDAC1 may play a regulatory role in the initiation and progression of intestinal bowel diseases. BTG1 is a functional regulatory gene associated with cell growth and differentiation. Similar to FASLG, it also regulates the apoptosis of specific target cells and may further regulate specific cytokines associated with inflammation such as IFN-β [94]. Histone deacetylase is commonly used to modify the epigenetic status and regulate gene expression [95]. RUNXI, known as runt-related transcription factor 1, is quite crucial in the development of normal hematopoiesis as a part of CBF (core binding factor). Associated with T cell function and TGF-β, RUNX1 has been proven to be quite crucial in inflammation initiation [96, 97]. Considering the strong relationship between IBD and immune reaction, RUNX1, which regulates the function of T cells, may also participate in the initiation of IBD [98].

3.4.4. Candidate Genes Contributing to Protein Kinase Cascade (GO: 0007243). Four functional genes have been clustered into such group. Genes like F2, ZAP70, and TLR4 have already been analyzed above. The gene MARK2 (rank 2 in the mRMR feature list) may also contribute to the initiation and progression of IBD by interfering with the protein kinase cascade. Inflammation is a basic pathological process regulated by the immune system [99]. Therefore, the immune system plays an irreplaceable role in IBD [100]. Several predicted genes have been confirmed to be associated with the immune system and participate in the immune reaction. MARK2 is a serine/threonine-protein kinase that is the major regulator of cell polarity in epithelial cells, including intestinal epithelial cells. Since immune cells in intestinal system have been proven to be regulated by such gene, the abnormal expression and effect of MARK2 may contribute to the unusual activation of focal inflammatory reaction in the digestive system, which may further promote IBD [101].

3.4.5. Candidate Genes Contributing to Intrinsic to Plasma Membrane (GO: 0031226). Among the candidate genes obtained by the SP approach, THBD is an endothelial-specific type I membrane receptor that binds thrombin [102]. As a specific protein in coagulation mechanisms, this receptor has also been reported as a potential inflammation mediator and may have a specific function in IBD [103, 104]. We also predicted a specific member of the TNF family, FASLG, as a candidate gene. FASLG has been proven to be involved in the induction of apoptosis triggered by binding to FAS [105]. Members of the TNF family have been widely reported to participate in IBDs by regulating the apoptosis of specific local cells [106, 107].

For candidate genes listed in Table 1, FOLR1 (rank 6 in the mRMR feature list), the folate receptor, participates in intestinal inflammation via the regulation of folate. Folate is associated with cell apoptosis in bowel tissues and has been reported to be crucial in colonic epithelial cell proliferation implying its potential role in inflammatory bowel diseases [108, 109]. SLC22A4 (rank 18 in the mRMR feature list) is a homolog of SLC22A5, which has been reported to be crucial in Crohn’s disease and is also overexpressed in this disease [110]. However, just like SLC22A5, SLC22A4 has not been confirmed to be overexpressed in ulcerative colitis [111, 112]. The genes mentioned above can distinguish IBD subtypes at the genetic level and may serve as new markers for the classification of inflammation in intestinal tissues. As a receptor of significant biological signals, LEPROT (rank 10 in the mRMR feature list) encodes a crucial receptor of GH and has been reported to be associated with the initiation of inflammation in the intestine in mice [113]. IBD has been regarded to be the result of immune systematic disorders and autoimmune reactions [1, 100]. Another gene, CLEC1B (rank 16 in the mRMR feature list), also participates in the development of IBD via the regulation of the intestinal immune system, especially the proliferation of NK cells and the formation of lymph nodes [114]. Apart from NK cells, activated cell is also a major part of the immune system and has been shown to be related to IBD [115, 116].

3.4.6. Candidate Genes Contributing to Apoptosis (GO: 0006915). Some candidate genes have been confirmed to participate in the apoptosis processes during the pathological processes of IBD. Apart from genes like SLTM, LCK, F2, FASLG, and BLCAP which have just analyzed above, the candidate gene RHOT2 (rank 21 in the mRMR feature list), a mitochondrial GTPase involved in mitochondrial trafficking, has been proven to be crucial regulator of Ca2+ in T cells. Thus, RHOT2 may also contribute to IBD [117]. IBD is a common disease involving the digestive system, especially the intestinal tissue [1]. However, IBD has also been shown to be associated with carcinoma in the digestive system, especially colorectal cancer [100]. Several of our predicted genes are also involved in tumor initiation, where cells may have mutated in precancerous lesions, including severe IBD. Most of these genes are related to cell proliferation. BLCAP (rank 7 in the mRMR feature list), which was first reported in bladder cancer, regulates the proliferation of cells that are quite common in intestinal tissue of IBD patients [118].

3.4.7. Candidate Genes Contributing to Regulation of Cell Proliferation (GO: 0042127). Among the candidate genes listed in Table 2, several have been confirmed to contribute to cell proliferation, implying the potential role of that during IBD initiation and progression. STKII, a functional serine/threonine kinase, regulates the polarity of cells and may participate in tumor suppression [119]. NF-kB is a crucial transcriptional factor that participates in the inflammation process [120]. STKII (also known as LKB1) directly regulates the function of NF-kB and is definitely associated with inflammation [121]. STKII also regulates the proliferation and maturation of intestinal cells, which indirectly reflects the regulatory function of STKII in intestinal tissues. The calcium binding protein SI00A6 is also on our predicted list, and it is located in the cytoplasm and nucleus of a wide range of cells. SI00A6 regulates the progression of the cell cycle and the differentiation of specific cells [122]. Considering the tight relationship between IBD and cancer, some of our predicted genes are also associated with tumor initiation [123]. We
also predict as a candidate gene a serine proteinase inhibitor SERPINE1, which encodes the principal inhibitor of tissue plasminogen activator (tPA) and urokinase (uPA). Tissue plasminogen activator and urokinase are both associated with inflammation and the process of wound healing [124]. SERPINE1 and proteins in the downstream of its specific pathway have also been reported to be directly associated with IBD as a functional regulator [125, 126]. As a candidate gene, we also predicted an angiogenesis-associated gene VEGFC, which regulates angiogenesis and endothelial cell growth [127, 128]. VEGFC has been reported to participate in several intestinal disorders including IBD and some specific digestive tract cancers [129, 130].

3.4.8. Other Candidate Genes. Four candidate genes obtained by mRMR and IFS methods were not clustered into any above group. The candidate gene ANXA11 (rank 20 in the mRMR feature list) is a predicted gene that regulates the autoimmune reaction. Such a gene has been reported to be related to several autoimmune disorders and may further participate in intestinal inflammation [131]. As a part of the MLL complex, OGT (rank 15 in the mRMR feature list) regulates the cell cycle of intestinal cells, including immune cells [132]. Therefore, the abnormality of the OGT gene may induce IBD in various downstream pathways. Another candidate gene, USPL1 (rank 17 in the mRMR feature list), also participates in intestinal inflammation reaction via the SUMO complex [133]. Large-scale mapping of human protein-protein interactions by mass spectrometry revealed several genes associated with inflammation, especially in the intestine [76]. The last gene, HISTH2AC (rank 11 in the mRMR feature list), is also a candidate gene for cancer. Such gene was first reported in breast cancer and regulates the proliferation of tissue cells, similar to BLCAP [134].

3.5. Comparison of Other Methods. To indicate the effectiveness of the proposed method and the reliability of the obtained genes, we compared our method with other methods. Before making the comparison, 77 validated IBD-related genes were retrieved from [135], which are provided in Supplementary Material VI. These genes were used to test the results yielded by our method and other methods.

DisGeNET (Version 4.0) [136] is a discovery platform that collects gene-disease associations from several public data sources and the literature. Here, it was used to search IBD-related genes. The obtained material is provided in Supplementary Material VII, from which we extracted 100 genes with high confidence (score > 0.1) as the predicted genes of this method. DAVID 6.7 (https://david.ncifcrf.gov/) [48] was employed again to analyze the biological meanings behind the validated genes, predicted genes by our method, and predicted genes by DisGeNET. The enriched gene ontology (GO) terms and KEGG pathways for three gene lists are listed in Supplementary Material V. It can be observed that 209 GO terms and KEGG pathways were enriched by 77 validated genes, while, for predicted genes by our method and DisGeNET, we obtained 154 and 314 GO terms and KEGG pathways, respectively. For the 154 GO terms and KEGG pathways enriched by 41 predicted genes of our method, 51 (51/154 = 33.12%) were also enriched by 77 validated genes, while there were 117 (117/314 = 37.26%) GO terms and KEGG pathways enriched by both 77 validated genes and 100 predicted genes of DisGeNET.

At a first glance, the performance of the DisGeNET is superior to our method. However, our method still has its advantages. According to our method, 21 genes were extracted by analyzing the gene expression profiles using mRMR, IFS, and SMO methods. In fact, these genes can only help us to distinguish two subtypes of IBD (rather than all subtypes of IBD) and normal samples. Thus, they are parts of IBD-related genes even if they are really IBD-related genes. 20 additional candidate genes were further obtained based on these genes, thereby accessing 41 predicted genes. These 41 predicted genes, in fact, are deemed to be related to two subtypes of IBD rather other all IBD subtypes. On the other hand, 100 predicted genes yielded by DisGeNET considered all subtypes of IBD. It is an important reason why DisGeNET gave the better performance. However, the performance of our method is only slightly lower than that of DisGeNET. Therefore, we believe that the proposed method is still quite effective and the obtained genes can be important and reliable materials for the investigation of IBD.

4. Conclusions

This contribution provides a novel computational method to identify genes related to IBD, which consists of two main steps: (1) analyzing the gene expression profiles and extracting important genes for IBD and (2) applying the shortest-path approach to the network constructed using protein-protein interactions and identifying additional related genes. By analyzing the obtained genes, it is concluded that they have special relationships with IBD, implying that our method is effective. It is also believed that our method has potential applicability for the investigation of other diseases.

Competing Interests

The authors declare that there is no conflict of interests regarding the publication of this article.

Acknowledgments

This paper is supported by National Natural Science Foundation of China (31371335).

References


M. Jiang, Y. Chen, Y.-D. Cai, K.-Y. Feng, and K.-C. Chou, “Predicting anatomical therapeutic chemical (ATC) classification of drugs by integrating chemical-chemical interactions and similarities,” *PLOS ONE*, vol. 7, no. 4, Article ID e35254, 2012.


[77] B. Carvalho, C. Postma, S. Mongera et al., “Multiple putative oncogenes at the chromosome 20q amplicon contribute to colorectal adenoma to carcinoma progression,” *Gut*, vol. 58, no. 1, pp. 79–89, 2009.


