Review Article

It Medical Team
http://www.imedpub.com

Molecular Enzymology and Drug Targets
2572-5475

2022
Vol. 8 No. 4: 108

Review of Protein Pathway System by Utilizing Compound Resources

Yu-Dong Cai*

Institute of Systems Biology, Shanghai University, Shanghai 200444, China

Corresponding author: Yu-Dong Cai

kcchou@gordonlifescience.org

Institute of Systems Biology, Shanghai University, Shanghai 200444, China

Citation: Cai YD (2021) Review of Protein Pathway System by Utilizing Compound Resources. Mol Enzy Drug Targ, Vol. 8 No. 4: 108.

Abstract

Basic issue with proteomics and systems biology. An enormous quantity of knowledge about many organisms has been gathered during the last ten years, both at the genetic and metabolic levels. Collected, organised, and methodically kept in a variety of niche databases, including KEGG,EcoCyc, MetaCyc, ENZYME, and BRENDA. With the use of these data, it is now possible to such a crucial issue. In this article, we examined established regulatory mechanisms in utilising various (biological and visual) features extracted from each of the 17,069 protein-formed systems, 169 of which are recognised regulatory pathways, or positive pathways. KEGG pathways were used, however 16,900 of them were unfavorable—that is, they weren't formed as biologically meaningful route. Through cross-validation, it was discovered that the overall success rate in identifying the good paths was 79.88 percent. Although the results are still preliminary, it is hoped that this innovative technique and good outcome will inspire further research into this crucial issue.

Keywords: Protein-forming system; regulatory pathway; minimum redundancy maximum relevance; gene ontology; biological graphic feature

Received: 01-Aug-2022, Manuscript No. IPMEDT-22-12914; **Editor assigned:** 03-Aug-2022, Preqc No. IPMEDT-22-12914(PQ); **Reviewed:** 17-Aug-2022, QC: IPMEDT-22-12914; **Revised:** 22-Aug-2022, Manuscript No. IPMEDT-22-12914 (R); **Published:** 29-Aug-2022, DOI: 10.36648/2572-5475-08.04-108

Introduction

Metagenomes and individual genomes are two examples of large-scale datasets that have grown in size due to the advent of high-throughput experimental techniques, necessitating renewed attempts to create computational techniques for a more accurate biological interpretation of all this data. In particular databases that are accessible on multiple websites, a significant amount of knowledge about various organisms has been gathered and methodically recorded at the genetic and metabolic levels [1]. KEGG is a popular knowledge database that contains graphical diagrams of biochemical pathways, including the majority of recognized metabolic pathways and some recognized regulatory pathways. It is used for the systematic analysis of gene functions in terms of interactions between genes and molecules. A new global map of metabolic pathways, which is effectively a combination of roughly 120 previous route maps, has recently been added to KEGG PATHWAY [2]. The Kyoto Encyclopedia of Genes and Genomes, or KEGG, is a well-known database for the systematic analysis of gene functions in terms of interactions between genes and molecules. It is made up of graphical representations of biochemical pathways, including the majority of recognized metabolic pathways and some recognised regulatory pathways [3]. A new global map of metabolic pathways, which is effectively a combination of roughly 120 previous route maps, has recently been added to KEGG PATHWAY. An ontology database called KEGG BRITE depicts the functional hierarchies of diverse biological items, such as molecules, cells, organisms, diseases, and medications, as well as the connections between them. Experimental knowledge is arranged in these databases. An ontology database known as KEGG BRITE represents the functional hierarchies of numerous biological items, such as molecules, cells, organisms, diseases, and medications, as well as the connections between them. Numerous studies from numerous research facilities around the world have shown that mathematical analysis, computational modeling, and the application of cutting-edge physical concepts can be used to address significant issues in biology and medicine, such as protein structural class prediction modeling of targeted proteins for drug design. Encouraged by these encouraging findings, the current investigation was started to address a crucial issue in system biology and proteomics. The routes lacking biological characteristics or GO information were eliminated. Additionally, pathways with less than three proteins were disregarded. As a consequence, 169 protein-forming or regulatory pathways were discovered; these are known as "positive pathways." The two paths listed below were used to produce data on negative pathways: Proteins were chosen at random to be the nodes of a graph, after which some random arcs were made between the protein nodes. The size distribution of the arcs in the positive paths was used to determine how many arcs should be present in each pathway. This study created 100 times more negative pathways than good ones since positive pathways are extremely uncommon compared to the huge majority of negative paths. The distribution of the arcs in the positive pathways is shown in Online Supporting Information S2 together with the 16,900 negative pathways that were thus obtained. Numerous prior research on a range of significant biological topics, such as enzyme-catalyzed processes, protein folding kinetics, and suppression of HIV-1 reverse transcriptase, have shown that the use of graphic tools to explore biological systems can yield helpful intuitive insights. Drug metabolic systems, processive nucleic acid polymerases and nucleases inhibition kinetics, and recently, graphical techniques have also been used to address a variety of medical and biological. Proteins are the vertices in networks that have been parsed from KGML files, and the arcs show the relationships between the protein vertices. Since the relationship between two proteins is directional, each graph is a directed graph or digraph. For example, protein P1 can sometimes control protein P2, while P2 cannot always regulate P1. In this study, 264 features of biological properties were derived from biochemical properties and physicochemical properties, including amino acid compositions, hydrophobicity, normalized van der Waals volume, polarity, polarizability, solvent accessibility, and secondary structure. Each directed graph that represents a pathway had 88 graph features extracted from it [4].

Discussion

Online Supplemental Material is in a 352-D (dimensional) space, we may similarly identify each of the 16,900 negative pathways uniquely, just as we did for the 169 positive pathways. Because the associated file is too big to submit, the exact results for the 16,900 negative paths are not displayed in this instance. It is nevertheless accessible upon request. Many graph properties were actually derived, and they were taken from an undirected graph. Every pathway in this study can be thought of as a directed graph, where the vertices represent proteins and the arcs represent relationships. As will be further detailed, the arcs are weighted according to the possibility that they may interact with one another. Following groupings were created from the 352 features [5]. Normalized van der Waals volume, polarity, and polarizability, along with hydrophobicity: Each of these physicochemical parameters can provide 42 traits. As features from other properties can be obtained in a manner similar to this, we will only describe how to obtain features from the hydrophobicity property here. Each amino acid is classified as either polar (P), neutral (N), or hydrophobic (H) (H). A protein pseudo-sequence is the sequence that results from replacing each amino acid in a given protein sequence with P, N, or H. The percentage of P, N, and H in the entire pseudo-sequence is known as composition (C). The shifting frequency between any two characters is known as a transition (T). Maximum relevancy with the least amount of redundancy. Maximum relevance would ensure selection of the features most important to classification, while minimum redundancy would ensure elimination of features previously covered by the features chosen [6]. One feature at a time was chosen by mRMR and added to the selected list during the selection process. A characteristic with the greatest relevance and the least amount of duplication was chosen in each round. As a result, we were able to compile an orderly list of all the characteristics that were chosen (**Figure 1**).

Gene ontology

To indicate how likely it is that an interaction between two proteins will occur; some properties require the arc weight. We used the gene ontology consortium (GO) to represent each protein in order to compute the edge weight of two interacting proteins. "Ontology" relates to the subject of existence and is a specification of a conceptualization. The three requirements of molecular function, biological process, and cellular component determine the existence of GO [7]. Because these three criteria indicate the attribute of gene, gene product, gene-product groups, and core properties reflecting the sub cellular localization, the GO consortium is seen to be a very effective and useful tool for examining protein-protein interactions. There is a description of the GO (gene ontology) encoding process.

Minimum redundancy maximum relevance (mRMR)

Selection might minimise the feature dimensions to increase a learning machine's effectiveness. The mRMR technique, first put forth by Peng, can be used to execute the concrete procedure. This is due to its ability to strike a balance between minimal repetition and high relevance. Maximum relevance would ensure selection of the features most important to classification, while

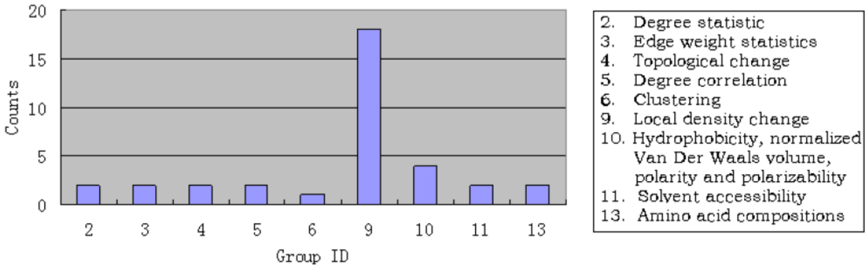

Figure 1 Illustration to show the distribution of features.

minimum redundancy would ensure elimination of features previously covered by the features chosen [8]. One feature at a time was chosen by mRMR and added to the selected list during the selection process. A characteristic with the greatest relevance and the least amount of duplication was chosen in each round. As a result, we were able to compile an orderly list of all the characteristics that were chosen. During the redundancy calculation and relevance, the mutual information (MI) was adopted. Let Ω denote the whole feature set. The selected feature set with m features is denoted by Ωs, and the rest of n features are denoted by Ωr. The relevance of a feature f and the target variable h can be computed as I (f, h), the redundancy between a feature f and the selected Ωs is computed.

Nearest neighbour algorithm

If there are m training paths in the NN algorithm, each of which is either positive or negative, it is necessary to evaluate whether a query protein system forms a positive or negative pathway. The new pathway's closest neighbour is identified after the distances between each of the m paths and it is calculated [9]. The query protein system is given a positive or negative pathway assignment depending on whether the nearest neighbour is positive or negative.

Jack-knife cross-validation

The jackknife test was used to evaluate the prediction model. The following three cross-validation techniques are frequently used in statistical prediction to assess a predictor's accuracy. However, as clarified and demonstrated by the three cross-validation methods, the jack knife test is thought to be the most objective and can always produce a singular result for a given benchmark dataset; as a result, it is increasingly used and widely acknowledged by researchers to assess the accuracy of different predictors. As a result, the jack-knife test was used in this study to evaluate the efficacy of our prediction system as well. Each statistical sample in the benchmark dataset was separately chosen as the prediction target during the jack-knifing process, and the remaining samples were used.

Incremental feature selection (IFS)

By using the Fi characteristics in conjunction with the NN algorithm, we were able to accurately predict the positive paths as measured by jack-knife cross-validation. As a result, a curve known as the IFS curve was created, with identification accuracy serving as both its y-axis and its x-axis. From was used to obtain the mRMR software [10]. It was executed using the default settings. Through the use of the mRMR programme, the following two feature lists were obtained: MaxRel features list and mRMR features list. We researched the top 10 percent of features for the MaxRel feature list (35 in total). The distribution of these traits is displayed. It is clear that 27 (77.1%) of the characteristics originate from the pathway graph, demonstrating that among the chosen elements, graph features play the largest role in the formation of regulatory pathways. 18 (51.43%) of the 27 features came from the 9th feature group, which captures the essence of the similarity in question and suggests that related proteins can be regulated by the same protein.

Conclusion

We looked at 352 attributes that were taken from both the positive and negative paths that were developed. The 352 features were split into two categories: 264 were obtained from biological properties of proteins, and 88 were graph ones, which meant that each pathway was regarded as a graph. These features were analysed using the mRMR (minimal redundancy maximum relevance) and IFS (incremental feature selection) approaches. A jackknife test and nearest neighbour algorithm were employed to assess how well our model identified the promising pathways. In the end, it was determined that 22 criteria were crucial for classifying the data. These results might serve as inspiration for additional research on this crucial and difficult subject.

Dimension and density of the graph. Let's say the formula for a pathway's graph is G = (V, E). In which V stands for the vertices and E for the arcs. The amount of proteins in the pathway determines the graph's size. Degree information. The quantity of a vertex's in- or out-neighbours is referred to as the vertex's in- or out-degree. The mean in-degree, variance in in-degree, median in-degree, maximum in-degree, mean out-degree, variance in out-degree, median out-degree, and maximum out-degree were all taken into account in this study as features. Weight statistics for edges. Consider the weighted pathway graph G = (V, w (E)) where each arc is given a weight w between. When w (e) = 0, it is conceivable for some arcs of eE; in two cases, we extracted features. We looked at 352 attributes that were taken from both the positive and negative paths that were developed. The 352 features were split into two categories: 264 were obtained from biological properties of proteins, and 88 were graph ones, which meant that each pathway was regarded as a graph. These features were analysed using the mRMR (minimal redundancy maximum relevance) and IFS (incremental feature selection) approaches. A jackknife test and nearest neighbour algorithm were employed to assess how well our model identified the promising pathways. In the end, it was determined that 22 criteria were crucial for classifying the data. These results might serve as inspiration for additional research on this crucial and difficult subject. The percentage of P, N, and H in the entire pseudo-sequence is known as composition (C). The shifting frequency between any two characters is known as a transition (T) (such as P and N, P and H, N and H). The distribution (D) is the portion of the pseudo-sequence that is required to contain the first, first 25%, first 50%, last 75%, and last 75% of the Ps, Ns, and Hs, respectively. In conclusion (C), (T), and (D) have three, three, and fifteen properties, respectively. 42 features total—21 plus 2—are obtained. Compositions of amino acids: the proportion of each amino acid in the entire sequence. In total, 40 features concerning the composition of amino acids are extracted (20 x 2). Based on hybrid qualities, graph properties, and biochemical and physical properties, the KEGG pathway network analysis approach. The "out local density" and "in local density" characteristics, both of which were connected with the change of the number of edges when various weight cut-offs were applied to the graph, were found to be the features contributing most to the formation of pathways. Therefore, if higher weight cut-offs were used, more edges might still be present in the positive graph. The "topological mean," which represents different protein topologies in the regulatory pathway, is another graph feature that contributes more to the pathway. A full graph has a maximum topological mean for a non-broken graph, while a linear graph (in which the proteins trace a linear path) has a minimum topological mean. A densely-connected graph always has higher topological mean, indicating a higher likelihood to form a regulatory pathway. Determinants of the regulatory networks. Protein conformation, and consequently their interactions and binding sites, were significantly impacted by the distribution of polarity in protein structures.

Acknowledgement

None

Conflict of Interest

No conflict of interest

References

1 Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M et al. (2008) KEGG for linking genomes to life and the environment Nucl Acid Res 36: D480-D484.

2 Klukas C, Schreiber F (2007) Dynamic exploration and editing of KEGG pathway diagrams. Bioinformatics 23: 344-350.

3 Caspi R, Foerster H, Fulcher C, Hopkinson R, Ingraham J et al. (2006) A multiorganism database of metabolic pathways and enzymes. Nucl Acid Res 34: D511-D516.

4 Caspi R, Foerster H, Fulcher C, Kaipa P, Krummenacker M et al. (2008) The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. Nucl Acid Res 36: D623-D631.

5 Zhou GP (2001) Some insights into protein structural class prediction. Protein Struct Funct Genet 44: 57-59.

6 Chou KC, Zhang CT (1995) Prediction of protein structural classes. Crit Rev Biochem Molec Biol 30: 275-349.

7 Zhou GP, Troy FA (2006) NMR studies on how the binding complex of polyisoprenol recognition sequence peptides and polyisoprenols can modulate membrane structure. Curr Protein Pept Sci 6: 399-411.

8 Chou KC (2004) Review: Structural bioinformatics and its impact to biomedical science. Curr Med Chem 11: 2105-2134.

9 Sharma AK, Zhou GP, Kupferman J, Surks HK, Christensen EN et al. (2008) Probing the interaction between the coiled coil leucine zipper of cGMP-dependent protein kinase Ialpha and the C terminus of the myosin binding subunit of the myosin light chain phosphatase. J Biol Chem 283: 32860-32869.

10 Zhou GP, Surks HK, Schnell JR, Chou JJ, Mendelsohn ME et al. (2006) The Three Dimensional Structure of the cGMP-Dependent Protein Kinase I-α Leucine Zipper Domain and Its Interaction with the Myosin Binding Subunit. Blood 104: 963-968.