Analysis on hydrophobic properties of proteins based on Giraph

Yongyu Fan[1, a], Ruoming He[2,b] and Xuesong Wang[3,c]

[1]Sino-Dutch Biomedical and Information Engineering School, Northeastern University, Shenyang, Liaoning Province, China

[a]365190168@qq.com, [b]ruominghe@163.com, [c]a18666667@gmail.com

Keywords: analysis on hydrophobic properties; protein; graph model; Giraph

Abstract. The previous experiments have shown that hydrophobic properties of 20 kinds of human's amino acid are different. It was of great significance to study whether hydrophobicity of proteins had a role in the research of the intimate relationship of proteins. Besides, it is also very important to balance the proteins'structure and function. Because of the highly complex data produced by hydrophobic proteins, it is suggested to build a computational graph model to solve this problem. In this paper, Giraph platform is used to calculate the data that come from protein interaction database (DIP) . After analyzing the big graph models composed of protein data, this thesis compared the results with those of Cytoscape which is a traditional protein analysis software and finally came to the same conclusions. What's more, Cytoscape didn't work when the data of proteins were huge. But Giraph turned out to be more accurate and efficient.

Introduction

Currently, some experiments have proved that the hydrophobic properties of 20 kinds of amino acids in body are different. Taking advantage of these characteristics, people can use the method of hydrophobic graphs to predict different epitomes on the surface of proteins as well as some membrane proteins. People can also research the intimate relationship among them by analyzing different features of hydrophobicity of proteins. And it is also important for the study of the proteins' formation and structure.

Due to the high complexity of the protein data, it needs to be calculated and solved by using the network model. The network model can be abstracted to be a graph. Because of the large amount of protein data, big graph model is used to solve the problem. In these big graphs of proteins, nodes represent a variety of genes, edges on behalf of all kinds of interactions among the genes. Therefore, it is required to use the big graph model to explore the relationship of proteins. In this way, human being can recognize the phenomenon of life and learn the constructions of the life and principles of the evolution.

Because of the large-scale diagram from the proteins, it turns out to be a serious problem to handle massive amounts of data. Although it has many researches on the techniques solving graph problems, the technology of big graph issues is still not mature. Cloud computing can provide a distributed storage mode which makes it possible to store and deal the large-scale data. BSP is a fundamental model in cloud computing which has a strong running speed and can process huge data efficiently. Giraph platform is based on BSP model which has not only the excellent performance of iterative calculation but also the efficient processing speed. So we use it as a tool to deal with the hydrophobic protein problem.

This paper is based on using the Giraph platform to analyze the big graph of proteins which makes experience on the study of protein hydrophobic. By abstracting the proteins as a big graph, nodes in the graph represent various genes that compose the proteins. Edges are on behalf of the interaction among these various genes. By analyzing the big graph, we can get the conclusion on hydrophobic of proteins. Using the Giraph platform to analyze the big graph, it can give a correct conclusion with little time. The contribution of this paper concluding:

(1) This paper abstracted the protein interaction network into a big graph model. To solve the problems of the hydrophobicity of proteins can be replaced by analyzing the big graph.

(2) This paper used the Giraph platform to analyze the big graph model formed by proteins. Enhancing the efficiency of solving problems on big graph models.
(3) This paper compared the experimental results with the results produced by the Cytoscape software. Verifying the accuracy of results and prove the feasibility of the experimental method.

This paper contains four parts. Part one is the introduction of related works on hydrophobicity. This part mainly describes the technique background and their disadvantages. Part two is analysis on technique. It mainly introduces the technique on the study of hydrophobicity. Part three is the result, which proves the accuracy and efficiency of the experiment. Part four is the conclusion on study about hydrophobic properties of proteins based on Giraph.

Related work

At present, the existing methods of detecting the protein hydrophobicity are all biological methods for gene sequencing. In the reference[1-2], the 8 - aniline - 1 - naphthalene sulfuric acid was used as probe. A theory that Octadecatetraenoic acid was suitable for detecting some protein was raised in reference[3-4]. Then, Canadian scientists Haskard et al[5-7] put forward an idea that PRODAN probe can be used to detect the protein surface hydrophobicity. However, considering the cases that the hydrophobic values detected by CPA and ANS were usually different cannot be explained reasonably, and the drew back that when using the probe method, the interaction between ion probe and protein may be take place, the Fluorescent probe method was not the optimal scheme to solve the problem of protein hydrophobicity.

A method that used lauryl sodium sulfate compound to detect the insolubility protein hydrophobicity was presented in reference [8].Scientists Smith et al[9-10] utilized the combining capacity of triglyceride and protein to detect the protein hydrophobicity. Due to the difficulties in theoretical explanations, this type of the methods is still not optimal.

Cytoscape, a software used to form big graphs, is used in analyzing the character of protein. This method can not only avoid the shortcoming that there may be interaction between protein and biological reagents but also provide a better envision for solving the protein related problem. However, the processible data size of this method is relatively small and manual computation is necessary. Thus, it is not suitable for the big data processing problem.

Protein hydrophobicity analysis based on Giraph platform

Introduction to Giraph platform

Giraph is a big graph calculation system using iteration for calculation. Its operation is based on Hadoop platform based on BSP(Bulk Synchronous Parallel) model[11]. It can realize the super-large scale calculation of the relationship between edges and nodes because its algorithm is a superstep operation of multiple iterations whose calculation efficiency is very high.

The graph model on protein hydrophobicity

The main character of the protein is presented in the different space structure of the polypeptide which constitutes protein. Taking protein hydrophobicity as an example, protein consists of 20 different kinds of amino acids. These 20 kinds of amino acids can be seen as different nodes in the hydrophobicity graph model. Thus, the space structure of certain protein can be determined by a definite 2 or 3 dimensional location information of the nodes, and the distance between two proteins can be determined by the length information of the ledges in the big graph model. Thus, the biological significance of the protein can be transferred to the calculable information in the graph model.

However, the interaction among proteins consists of hydrogen bond, disulfide bond, peptide bond, Van der Waals force and ionic bond etc. These interactions correspond to the different weights relationship among every edge in the graph according to their intensities and patterns. As the result, the protein hydrophobicity can be transferred to pure mathematical graph model. The calculation capacity of Giraph can be used to calculate the hydrophobicity relationship of the big molecule proteins.

The analytical method of protein hydrophobicity based on Giraph

The calculation process of Giraph can be seen as multiple iterations which are called superstep in the BSP. After judging the validity of the initial value of the superstep(1st-2nd row), every task does the calculation of their own nodes and estimate the minimal value of the current step(3rd-4th row). Then the result is sent to the next superstep. If the minimum value of the last step is smaller than the current value, it is supposed to add the distance between this node and next node to the current value and sending it to the next node. Instead, it is replaced to use the minimum value (5th-7th row).

```
1）If  Superstep == 0
2）        Superstep = NodeNum；
3）For messageNum = 0 to N
4）        minDist = Min(minDist, message.get())；
5）For edge       = 1 to N
6）        If (minDist < Value)
7）                distance = minDist + edge；
```

Experimental analysis

This chapter introduces the source of experimental data and carries out specific analysis and proof of the correctness and validity of the experimental method. In the end it gives out a brief summary of the experiment.

Experimental data

The experimental data of the text are provided by Database of Interacting Proteins (DIP), which has included around 18,000 items of information on protein interactions. The database uses opening mechanism, which means the contents can update with the increase of molecular interactions information, and the reliability of experimental data could complete tests on genome level in time as well based on the development of data quality evaluation method. Therefore, it is widely used in the field of researching protein, from which the experiment downloaded large amounts of data, like the protein MrpF, protein PhoA and protein homo as source.

Correctness analysis

First, researches show that the hydrophobic property of proteins can only be measured by biological methods such as genetic sequencing, and it is difficult to calculate the shortest path between the hydrophobic residues mathematically. Compared with practical issues, it is suggested to use programming ideas of Giraph to work out the length of corresponding nodes. Figure one below shows the shortest path between two nodes in the big graph composed by Homo protein.

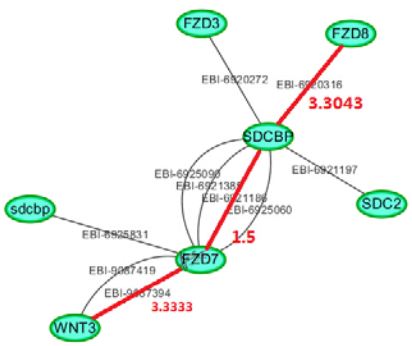

Fig.1 shortest path between nodes FZD8 and WNT3 in protein Homo

Second, compare Giraph programming with traditional protein analysis tool, the Cytoscape. By using the Cytoscape and manual calculation, it turns out that the distance between node G3H9I8 and SMO of protein homo is 25.3. The answer of Giraph is 25.3222, which has proved the correctness of the conclusion that the hydrophobic property of proteins is related to the shortest path. Table one below shows the comparability of Giraph and Cytoscaoe for distances based on 8 groups.

Tab.1 the comparability of Giraph and Cytoscaoe for distances based on 8 groups

	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Group 7	Group 8
Giraph	8.14	8.33	7.11	11.83	7.35	8.26	7.53	7.33
Cytoscape	8.1373	8.3333	7.1111	11.8333	7.3535	8.2626	7.5333	7.3333

Efficiency analysis

Though the Cytoscape software reveals all nodes and the distance between them, it lacks the function of calculating the shortest path directly. Only through manual calculation could the distance between required nodes be calculated, which is time-consuming, laborious, and it reduces the accuracy of the results. However, it is very convenient and accurate to calculate the results of the experiment through the Giraph platform. Table two below shows when the amount of data keeps increasing the time of Giraph. Apparently, Cytoscape turns out to be difficult to calculate, which easily proves the efficiency of the two methods when using them to calculate the shortest path between nodes.

Tab.2 the time of Giraph and Cytoscaoe for calculating distances based on 8 groups

	Group1	Group2	Group3	Group4	Group5	Group6	Group7	Group8
Giraph	6.007[s]	12.51[s]	57.4[s]	92[s]	3[min]	5min20[s]	6min15[s]	7[min]
Cytoscape								

Moreover, the feasibility of artificial calculation reduces when the amount of proteins data increases. But no matter how the amount of data increases, the Giraph can get results automatically and efficiently only by increasing the number of cluster. It gives full play to the advantage of efficiency on large amount of data processing and analysis technology, and improves efficiency on solving problems.

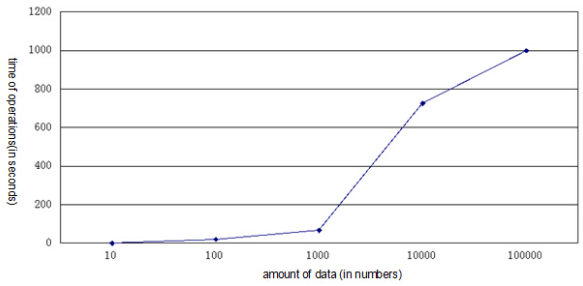

Fig.2 Time-cost by calculating different numbers of proteins with Giraph

Summary of experiment

With the development of proteomics and the increasing amount of proteins data, more and more proteins have been discovered to have hydrophobic relationships and can interact to form big graph model. Former researches, however, could only carry out experimental analysis on the hydrophobic property of proteins through biological methods such as genetic sequencing, which makes experimental instruments expensive and the process time-consuming and laborious. Besides, the feasibility of experiment was impeded in the case of large amounts of data. But by making use of the Giraph platform for programming and calculating, the idea of the paper, any number of proteins can be calculated accurately and efficiently, which reduces the cost greatly and improves work efficiency.

Conclusion

This paper uses the idea of building cluster and constructing the Giraph platform for programming and sets cloud computing as the environment of solving and processing large amount of data. In comparison with traditional protein analysis software-Cytoscape, it has been verified,

through extraction, analysis, processing and calculating on large amount of proteins data, that the hydrophobic property of proteins is related to the shortest path between them. Also, the accuracy and efficiency of the Giraph platform in processing large amount of data are reflected through crosswise and longitudinal contrast with the experimental data. This makes up for the blank of big graph model processing technology in biomedical application.

Acknowledgement

Thanks for the financial support of Sivi project of Northeastern University.

References

[1]Sakono M, Seko A, Takeda Y, et al. Glycan specificity of a testis-specific lectin chaperone calmegin and effects of hydrophobic interactions[J]. Biochimica et Biophysica Acta (BBA)-General Subjects, 2014.

[2]Gonzalez W G, Miksovska J. Application of ANS fluorescent probes to identify hydrophobic sites on the surface of DREAM[J]. Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics, 2014.

[3]Nor Afizah M, Rizvi S S H. Functional properties of whey protein concentrate texturized at acidic pH: Effect of extrusion temperature[J]. LWT-Food Science and Technology, 2014, 57(1): 290-298.

[4] Zhang W. Emulsifying Properties of Deamidated Barley Protein Fractions[D]. University of Alberta, 2014.

[5]Haskard C A, Li-Chan E C Y. Hydrophobicity of bovine serum albumin and ovalbumin determined using uncharged (PRODAN) and anionic (ANS-) fluorescent probes[J]. Journal of Agricultural and Food Chemistry, 1998, 46(7): 2671-2677.

[6]Mao X Y, Tong P S, Gualco S, et al. Effect of NaCl addition during diafiltration on the solubility, hydrophobicity, and disulfide bonds of 80% milk protein concentrate powder[J]. Journal of dairy science, 2012, 95(7): 3481-3488.

[7]Benjamin O, Lassé M, Silcock P, et al. Effect of pectin adsorption on the hydrophobic binding sites of β-lactoglobulin in solution and in emulsion systems[J]. International Dairy Journal, 2012, 26(1): 36-40.

[8]Bhuyan A K. On the mechanism of SDS‐induced protein denaturation[J]. Biopolymers, 2010, 93(2): 186-199.

[9]Smith L M, Fantozzlb P, Crevelinga R K. Study of triglyceride-protein interaction using a microemulsion-filtration method[J]. Journal of the American Oil Chemists’ Society, 1983, 60(5): 960-967.

[10]Farese Jr R V, Cases S, Smith S J. Triglyceride synthesis: insights from the cloning of diacylglycerol acyltransferase[J]. Current opinion in lipidology, 2000, 11(3): 229-234.

[11]Valiant Leslie G. A bridging model for parallel computation. Communication of the ACM, 1990, 33(3) : 103-111