Proceedings of the 2013 Winter Simulation Conference
R. Pasupathy, S.-H. Kim, A. Tolk, R. Hill, and M. E. Kuhl, eds

LEARNING PRIMARY FEATURE IN COMPRESSIVE SAMPLING SPACE: A SPARSE REPRESENTATION STUDY

Ya Nan Zhang
Jian Dong Ding
Feng Jin
Wen Jun Yin
Zhi Bo Zhu

IBM China Research Laboratory
Keyuan Road 399
Shanghai, 201203, CHINA

ABSTRACT

In most biological metabolic processes, protein-protein interactions (PPIs) play a vital important role, the identification of which has been attracting much effort and devotion. Nevertheless, there are still many difficulties because of lacking enough information such as protein homology, protein structure and so on. Accordingly, a novel sequence-based computational method is proposed to predict PPIs and has achieved a promising performance. This method was put forward by incorporating primary feature representation with compressed learning theory framework. When applied to the PPI data of yeast Saccharomyces cerevisiae, it shows an inspiring result and also performs well in an independent dataset. Our results not only demonstrate compressed learning theory framework is suitable for PPIs prediction, but also imply that it has potential applications in many other bioinformatics problems.

1 INTRODUCTION

The advent of high-throughput technologies generate a large number of gene sequences and protein sequence data , from which mining effective information is the main goal that computational biologists have been diligently seeking for. Furthermore, the amino acid sequence profile also provides a deep insight into protein function, at the core of which is protein-protein interactions that exist widely in metabolic pathways, transcription regulation and signaling cascades and so on. Therefore, as one of the major challenges in post-genomic era, identification of protein-protein interactions is crucial for elucidating protein functions and further understanding various biological processes in a cell. Recently an impressive set of experimental methods have been introduced for identification of PPI, including yeast two-hybrid systems, mass spectrometry, protein chips and so on.

However, experimental methods only reveal a small portion of complete PPI networks, and considering they are always time-consuming and thus have a lower efficiency. Hence, computational methods for PPI prediction are necessary and therefore many computational methods have been proposed.

2 MATERIALS AND METHODS

In this paper, the PPI data was firstly collected from Saccharomyces cerevisiae core subset of database of interacting proteins(DIP). Subsequently a new method based on auto covariance(AC) and compressed learning theory was proposed. AC is one of feature extraction and representation methods of ami-

no acids sequence, which integrates neighboring effect among amino acids and physicochemical properties of distinct amino acid residues. Hence, we could acquire feature vector for every sample in PPIs through the using of AC feature representation. Compressive sampling (CS), also called Compressed sensing, involves sampling signals in a non-traditional way-each observation is obtained by projecting the signal onto a randomly chosen vector. In this paper, feature vectors of n dimensions obtained from AC could be recognized as signals. The goal of compressed sensing is then to provide a $m \times n$ measurement matrix A, with the number of measurements m as small as possible. Consequently the matrix A is used to transform all feature vectors of n dimensions to feature vector domain of m dimension.

In order to compare different performances of various kinds of machine learning algorithm on obtained measurement domain, some typical learning algorithms including SVM, random forest, k-nearest neighboring (KNN) were selected and then three compressed learning model: CS-SVM, CS-random forest and CS-KNN were constructed, based on the PPI data of yeast Saccharomyces cerevisiae which had been tested in cross-validation and yielded promising prediction accuracy. Furthermore, another independent dataset constructed by other yeast PPIs was employed to further evaluate these models.

3 CONCULSIONS

In this paper, compressed learning framework is introduced into bioinformatics field and discusses its application in PPI prediction. Assessment results demonstrate efficient application of compressive learning framework in bioinformatics. Moreover, the ability that compressed learning algorithm may lower risk of overfitting problem implies a promising application prospect in many other problems of bioinformatics.

4 ACKNOWLEDGEMENT

This work was supported by the Science and Technology Project (Contact No.8300022813) from the State Grid Corporation of China.

REFERENCES

Fields, S. and O. Song, A novel genetic system to detect protein-protein interactions. Nature, 1989. 340(6230): p. 245-6.

Clb, C., Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature. 415: p. 180-183.

Donoho, D.L., Compressed sensing. Ieee Transactions on Information Theory, 2006. 52(4): p. 1289-1306.

Wold, S., et al., DNA and peptide sequences and chemical processes multivariately modelled by principal component analysis and partial least-squares projections to latent structures. Analytica Chimica Acta, 1993. 277(2): p. 239-253.

Xenarios, I., et al., DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. Nucleic acids research, 2002. 30(1): p. 303-5.

Bock, J. and D. Gough, Predicting protein-protein interactions from primary structure. Bioinformatics (Oxford, England), 2001. 17(5): p. 455.