IPASJ International Journal of Computer Science (IIJCS)
Web Site: http://www.ipasj.org/IIJCS/IIJCS.htm
A Publisher for Research Motivation ..... Email: editoriijcs@ipasj.org
Volume 5, Issue 2, February 2017 ISSN 2321-5992

Development of Predictor for Sequence Derived Features from Amino Acid Sequence using Associate Rule Mining

Miss. Bhavna Pachori

Vardhaman College of Engineering, Hyderabad

ABSTRACT

A review of the literature on successful implementation of ERP reveals that there are many case studies undertaken by researches, but very few have empirically examined the success factors of ERP implementation. While most of those empirical studies were undertaken in Western countries, very few had examined the implementations in Middle Eastern countries and none in Saudi Arabia. Factors and challenges of ERP implementation in developing countries differ from those of Western countries. Hence a gap in the literature that examines Middle Eastern countries exists. This study is motivated to fill such gap by going beyond case study and boundaries of Western counties to empirically examine the determinants of successful ERP implementation in Saudi Arabia. The main purpose of this study is to examine the influence of some critical factors on successful implementation of ERP.

1. ASSOCIATIVE RULE MINING

Association rule mining could be a common data processing technique, which may be accustomed turn out fascinating patterns or rules [2]. Association rule mining involves count frequent patterns (or associations) in giant databases, reportage all that exist on top of a minimum frequency threshold referred to as the 'support' e.g. analyzing grocery store basket information, wherever a grocery store would need to ascertain that product area unit oft bought along. Such AN association could be "if a client buys biscuits and patty then they're eightieth doubtless to shop for coffee". Rule support and confidence area unit 2 measures of rule power. Association rules area unit thought of fascinating if they satisfy each a minimum support threshold and a minimum confidence threshold [5].

1.1 Association Rules

Association rules area unit needed to satisfy a user-specified minimum support and a user-specified minimum confidence at a similar time. to realize this, association rule generation could be a ballroom dance method. First, minimum support is applied to seek out all frequent itemsets in a very information. in a very second step, these frequent itemsets and therefore the minimum confidence constraint area unit accustomed kind rules. whereas the second step is simple, the primary step wants a lot of attention. the matter is outlined as: Let I be a group of n binary attributes referred to as things. Let D be a group of transactions referred to as the information. every group action in D incorporates a distinctive group action ID and contains a set of the things in I.

2. MACROMOLECULE AND MACROMOLECULE FUNCTIONS

Proteins area unit giant, advanced molecules that play several important roles within the body. {they do|they area unit doing} most of the add cells and are needed for the structure, perform and regulation of the body's tissues and organs. Proteins area unit created from tons of or thousands of smaller units referred to as amino acids, that area unit connected to 1 another in long chains. There area unit twenty differing kinds of amino acids which will be combined to create a macromolecule. The sequence of amino acids determines every protein's distinctive 3- dimensional structure and its specific perform [6-8]. As amino acids confederate bound to make the things from that our life is born. it is a ballroom dance process: Amino acids get along and kind peptides or polypeptides. it's from these groupings that proteins area unit created. normally recognized amino acids embrace aminoalkanoic acid, glycine, essential amino acid, tryptophan, and valine. 3 of these — essential amino acid, tryptophan, and essential amino acid — area unit essential amino acids for humans; the others area unit essential amino acid, leucine, lysine, methionine, and essential amino acid. The essential amino acids can not be synthesized by the body; instead, they have to be eaten through food. They function accelerator catalysts, area unit used as transport molecules (hemoglobin transports oxygen) and storage molecules (iron is keep within the liver as a fancy

with the macromolecule ferritin), utilized in movement (proteins are the most important part of muscles), they're required for mechanical support (skin and bone contain collagen-a fibrous protein), they mediate cell responses (rhodopsin could be a macromolecule within the eye that is employed for vision), protein proteins area unit required for immune protection; management of growth and cell differentiation uses proteins (hormones) [4].

3. SEQUENCE DERIVED

Options Sequence derived options area unit the varied options of macromolecule that area unit accustomed predict macromolecule category. Sequence derived options area unit vital in macromolecule category prediction as these area unit the input to the HPF predictor. SDF's is derived from a given set of amino-acid (protein) sequences mistreatment numerous web-based bioinformatics tools [12]. the varied sequence derived options area unit as given below:
3.1 Extinction constant (Eprotein) Extinction constant could be a macromolecule parameter that's normally utilized in the laboratory for deciding the macromolecule concentration in a very answer by spectrophotometry. It describes to what extent light-weight is absorbed by the macromolecule and depends upon the macromolecule size and composition similarly because the wavelength of the sunshine. For proteins measured in water at wavelength of 280nm, the worth of the Extinction constant is determined from the composition of aminoalkanoic acid, tryptophane and aminoalkanoic acid.

4. LITERATURE SURVEY

Jensen et al. (2002) planned the human macromolecule perform from post-translational modifications and localization options. The prediction methodology concerned the employment of sequence derived options for human macromolecule perform prediction. The posttranslational modifications (PTMs) area unit the changes that occur to the macromolecule when its production by the method of translation. They extracted the sequence derived options from the various servers like Expasy, PSORT as mentioned in section five. Fourteen options were extracted from the aminoalkanoic acid sequences [6].

Al-Shahib et al. (2007) Calculated the frequency, total range of every aminoalkanoic acid and therefore the set of amino acids for the input macromolecule seqnence. To cipher spacing options, they additionally determined the quantity and size of continuous stretches of every aminoalkanoic acid or aminoalkanoic acid set. They divided each macromolecule into four equally sized fragments and calculated a similar feature values for every fragment and combination of fragments. additionally, the opposite options just like the secondary structure was expected mistreatment academic [10], the position of purported transmembrane helices mistreatment TMHMM [21] and of disordered regions mistreatment DisEMBL [15]. The options were used for macromolecule perform prediction [1].

Kanakubo et al. (2007) declared that association rule mining was one among the foremost necessary problems in data processing. With Apriori ways, the matter becomes inestimable once the full range of things area unit giant. On the opposite hand, bottom-up approaches like artificial life approaches were opposite of the top-down approaches of searches covering all transactions and should offer new ways of breaking faraway from the completeness of searches in standard algorithms.
Gupta et al. (2008) planned a unique feature vector supported chemistry property of amino acids for prediction macromolecule structural categories. They given a wavelet-based time-series technique for extracting options from mapped aminoalkanoic acid sequence and a hard and fast length feature vector for classification is built. moving ridge rework could be a technique that decomposes a symbol into many teams (vectors) of coefficients.
completely different constant vectors contain info regarding characteristics of the sequence at different scales. The planned feature vector contains info regarding the variability of 10 physiochemical properties of macromolecule sequences over totally different scales. The variability of physiochemical properties was painted in terms of moving ridge variance [14].

Jaiswal et al. (2011) Studied that the identification of specific target proteins for any unhealthy condition involves in depth characterization of the doubtless concerned proteins. Members of a macromolecule family demonstrating comparable options could show sure uncommon options once involved in a very pathological condition. They studied the Human matrix metalloproteinase (MMP) family of endopeptidases and discovered their role in numerous pathological conditions like inflammatory disease, coronary artery disease, cancer, liver pathology, cardio-vascular and neurodegenerative disorders, very little is understood regarding the particular involvement of members of the massive MMP family in diseases. They hypothesized that amino acid wealthy and extremely thermostable MMPs could be key players in unhealthy conditions and thus signify the importance of sequence derived options [3].

IPASJ International Journal of Computer Science (IIJCS)
Web Site: http://www.ipasj.org/IIJCS/IIJCS.htm
A Publisher for Research Motivation ..... Email: editoriijcs@ipasj.org
Volume 5, Issue 2, February 2017 ISSN 2321-5992

6. ALGORITHMIC RULE FOR PREDICTING SEQUENCE DERIVED OPTIONS

Among numerous Sequence derived options area unit Extinction constant, open-chain Index, Absorbance, No. of charged residues, No. of charged residues, work out Iso-electric point/molecular weight. These SDF's area unit integrated and computed in one platform mistreatment Associative rule mining. within the existing techniques the varied classes of options weren't computed for a similar input by any single tool rather totally different tools area unit obtainable for various classes as mentioned in section five. The ways for extracting multiple options of relevant to the density of the amino acids were additionally developed [1][14]. this work focuses on developing the one tool for extracting the options of multiple classes from the one input data. The algorithmic rule predicts the Extinction constant, open-chain Index, Absorbance, No. of charged residues, No. of charged residues, work out Iso-electric point/molecular weight from the input aminoalkanoic acid sequence. of these options area unit computed by giving single input sequence file that is style of string made up of doable combos of twenty characters representing the twenty amino acids. it's doable by desegregation of these computations by mistreatment associative rule mining. The aminoalkanoic acid sequences could contain thousands of aminoalkanoic acid i.e characters in a very string. therefore extracting any sequence derived feature from the sequence could be a work out intensive downside. Computing multiple options from single sequence is even a lot of herculean task. every computation has some common half that's referred to as the intersection. The intersection and union operations area unit utilised to offer the integrated results from the one input data. the various algorithms for individual feature prediction area unit shown here for clear interpretation. Fig. one shows the steps concerned in Extinction constant Prediction. Fig. 2 and Fig. three shows the algorithmic rule for prediction of charged and charged residues severally. Fig. four shows the predictor for Iso-electric purpose and mass for the aminoalkanoic acid chain. Fig. five shows the open-chain Index prediction and Fig. vi shows the Absorbance prediction for the input data.

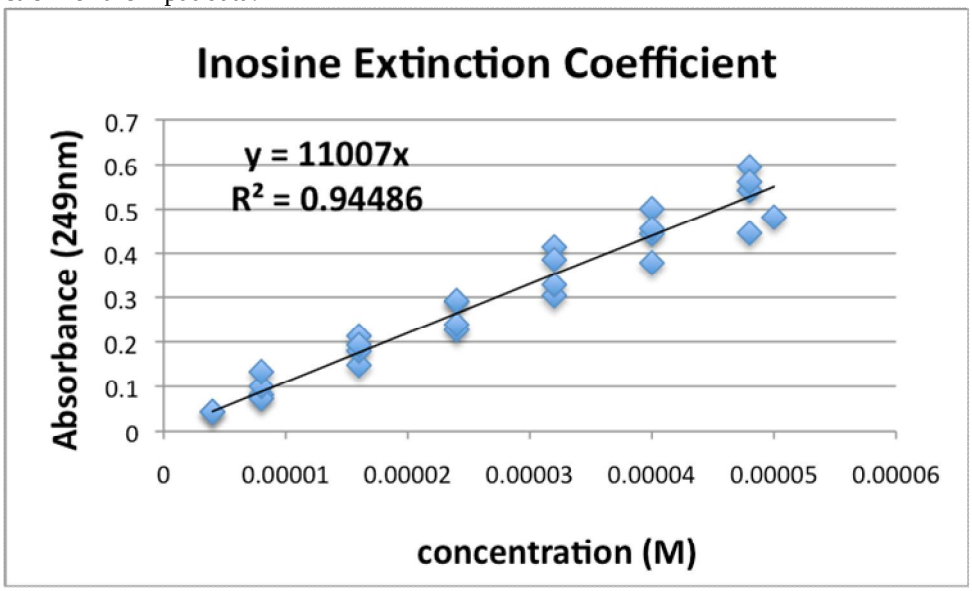

FIGURE 1: Prediction of Extinction Coefficient.

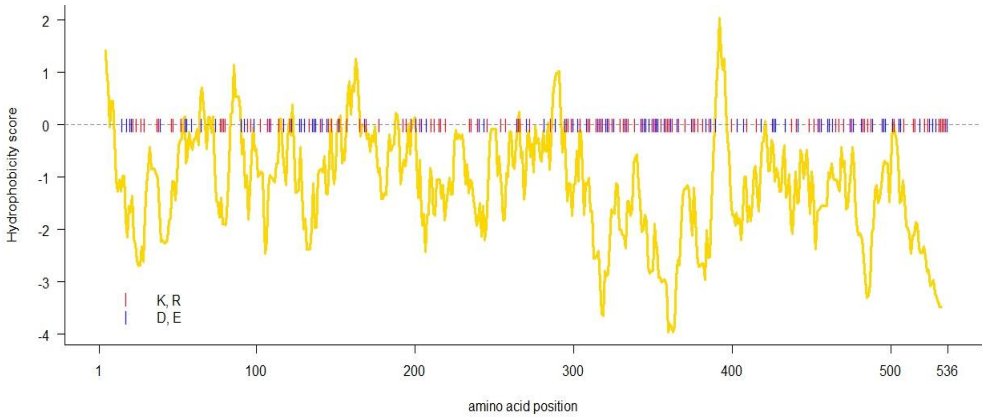

FIGURE 2: Prediction of Negatively charged residues

REFERENCES

[1]. A. Al-Shahib, R. Breitling, and D. R. Gilbert "Predicting macromolecule perform by machine learning on aminoalkanoic acid sequences – a important evaluation" BMC genetics, 8:1-10, 2007
[2]. A. Clare. "Machine learning and data processing for yeast purposeful genomics", Ph.D. thesis, University of Wales, Feb 2003 three. A. Jaiswal, A. Chhabra, U. Malhotra, S. Kohli, V. blue blood "Comparative analysis of human matrix metalloproteinases: rising therapeutic targets in diseases" Bioinformation 6(1): 23-30, 2011
[3]. D. Krane and M. Raymer. "Fundamental ideas of Bioinformatics", Pearson Education, New Delhi, pp.1-314 (2006) five. J. Han and M. Kamber. "Data Mining: ideas and Techniques", Morgan Kaufmann Publishers, pp. 226-229 (2004)
[4]. L. Jensen. "Prediction of macromolecule perform from Sequence Derived macromolecule Features", Ph.D. thesis, Technical University of Danmark, 2002
[5]. L. Jensen, M. Skovgaard and S. Brunak. "Prediction of Novel Archaeal Enzymes from Sequence Derived Features", macromolecule Science, 11: 2894-2898, 2002
[6]. L.J. Jensen, R. Gupta, N. Blom, D. Devos, J. Tamames, C. Kesmir, H. Nielsen, H.H. Starfeldt, K. Rapacki, C. Workman, C.A.F. Andersen, S. Knudsen, A. Krogh, A. Valencia and S. Brunak "Prediction of Human macromolecule perform from Post-Translational Modifications and Localization Features" Journal of biological science, 319(5): 1257-1265, 2002
[7]. M. Kanakubo and M. Hagiwara. "Speed up technique for Associative rule mining supported a man-made Algorithm", GRC book on granular computing, 38(12):318-323, 2007
[8]. M. Ouali, R.D. King "Cascaded multiple classifiers for secondary structure prediction" Prot Sci., 9:1162–1176, 2000