



Received on 13 August, 2016; received in revised form, 22 October, 2016; accepted, 06 December, 2016; published 01 February, 2017

## MULTI DIMENSION PROTEIN IMPACT MATRIX BASED PROTEIN SEQUENCE PREDICTION USING DATA MINING

A. Surendar

School of Electronics, Vignan's University, Guntur, Andhra Pradesh, India.

### Keywords:

Multi Dimension, Protein, Sequence Prediction, Matrix based

### Correspondence to Author:

**A. Surendar**

Assistant Professor, School of Electronics, Vignan's University, Guntur, Andhra Pradesh, India.

**E-mail:** surendararavindhan@gmail.com

**ABSTRACT:** Proteins are the most essential and versatile macromolecules of life, and the knowledge of their functions is a crucial link in the development of new drugs, better crops, and even the development of synthetic biochemicals such as biofuels. Experimental procedures for protein function prediction are inherently low throughput and are thus unable to annotate a non-trivial fraction of proteins that are becoming available due to rapid advances in genome sequencing technology. This has motivated the development of computational techniques that utilize a variety of high-throughput experimental data for protein function prediction, such as protein and genome sequences, gene expression data, protein interaction networks and phylogenetic profiles.


**INTRODUCTION:** The development of information technology has great impact in many domains. The medical industry is focused on developing innovative tools to support medical practitioners. There are many medical solutions has been emerged in this decade. However, there are research gaps where there is no noticeable solutions has been found. The disease prediction is the area what focused by the researchers. The disease prediction is the process of identifying the possibility of the disease from given input details. The disease prediction reads the given input pattern and history of disease pattern. Based on both of them, the method computes the probability of the disease to select a single one.

For any disease found on the human, the genes plays the vital role. There are number of genes present in the human anatomy.

Among them, there are few genes can be named which are supporting the growth of disease. Identifying such influencing genes is a challenging process, because there is only negligible gap between the relationship of the genes. There are number of gene selection approach has been discussed earlier. The popular frequent pattern mining technique can be used in selecting the gene sequence. Because of the size of sequence is higher, identifying the minimum sequence or restricted sequence is essential.

The gene sequence has no formal order and they can present in any form. In order to identify the sequence, it is necessary to generate the various sequences possible. By generating the number of sequences which are possible they can be used to predict the disease. The sequences can be generated in different size from one to many. One size gene has the size of single where the rest are vary from the size 2- any. So the gene sequence can be generated in different size and can be used to perform gene selection.

The growing representation of protein sequence has been used in many medical problems.

<p><b>QUICK RESPONSE CODE</b></p> 	<p><b>DOI:</b> 10.13040/IJPSR.0975-8232.8(2).427-34</p> <hr/> <p>Article can be accessed online on: <a href="http://www.ijpsr.com">www.ijpsr.com</a></p>
<p><b>DOI link:</b> <a href="http://dx.doi.org/10.13040/IJPSR.0975-8232.8(2).427-34">http://dx.doi.org/10.13040/IJPSR.0975-8232.8(2).427-34</a></p>	

Whatever the disease caused in human body is based on certain protein sequence. There are number of research has been ongoing for the detection and prediction of protein sequence. For example, given a long sequence S, there exist number of sub sequences can be generated. Among them you can identify a sequence which is the most reason for the cause of disease. Similarly, the problem of gene prediction has been approached by various researchers.

a	c	d	c	b	d	g	h	k	l
---	---	---	---	---	---	---	---	---	---

Among the small sequence given, we can generate a number of sequences of different size. If the generated sequences are placed in a gene set Gs, among them a sequence S can be identified.

d	G	h
---	---	---

If the selected gene sequence is the above one, then you can say that the above sequence is the reason for the disease. Similarly there are number of sequence can be identified.

The problem is identifying the most impacting gene sequence in many cases. Using the trace of sequence obtained from the diseased people, the method can generate number of sequences. For each sequence Si, the method can compute the impact factor. The impact factor is the measure which represent the fluency of sequence in the gene set.

**Methods Explored:** There are number of methods has been identified for the problem of sequence selection and this section list a set of methods.

To improve the stability of feature selection under varying samples, the author proposed a sample weighting technique in <sup>1</sup>, which uses microarray data. The method improves the performance gene selection and increases the stability of gene selection also. The performance of the method has been evaluated and compared with the performance of support vector machine and relief classifiers.

To perform microarray cancer classification, an Relevant and Significant Supervised Gene Clustering algorithm is discussed in <sup>2</sup>. The method uses mutual information based supervised gene clustering (MSG) algorithm to form the reduced gene clusters for cancer classification.

The efficiency of the method has been evaluated with different micro array cancer data sets and compared with the classifiers like Naïve bayes, K-nearest rule, and SVM.

Cancer subtypes prediction using Gene-Expression using Feature Selection and Transductive SVM has been discussed in <sup>3</sup>. The method adapts both gene selection and transductive support vector machine (TSVM) to predict the gene sets. The method identifies the genes using TSVM to improve the accuracy of prediction. The performance of the method has been compared with standard SVM technique.

To identify uncovered gene pathways which characterize the cancer heterogeneity an efficient method has been proposed in <sup>4</sup>, which uses the sparse statistical method. The method specifies set of pathway activities which are identified from the micro array data using Sparse Probabilistic Principal Component Analysis (SPPCA). The method also generates an association between gene-gene related to the cancer phenotypes.

Predicting metastasis of breast cancer has been discussed in <sup>5</sup> and performs a comparison of classification performed by different methods and analyzes the results. For the prediction of metastasis, the method uses voting approach. The method has produced efficient results than other methods.

The survivability of breast cancer diagnosis has been approached using embedded genetic algorithm in <sup>6</sup>. The shapely value based feature selection technique use include and remove memetic operators. The entire algorithms feature selection has been optimized using the genetic algorithm. The gene selection algorithm selects a subset of genes from the high dimensional data set using the genetic algorithm. The method performs the differentiation based ranking of genes to select them. The method use four different classifiers to improve the quality of gene selection.

In <sup>7</sup>, identifies the pattern of genes present in the breast cancer patients. Using the pattern identified from the gene set available, the method selects the subset of genes in form of pattern. The selected pattern represents the gene selection.

An comparative analysis has been performed with various gene classification approach in <sup>8</sup>. The author presents a comparative study on various classification algorithms and support vector machine.

Identification of a Comprehensive Spectrum of Genetic Factors for Hereditary Breast Cancer in a Chinese Population by Next - Generation Sequencing <sup>9</sup>, discussed to classify a complete spectrum of genetic factors for genetic breast cancer in a Chinese population, we did an analysis of germline alterations in 2,165 coding exons of 152 genes related with genetic growth using next-generation sequencing (NGS) in 99 breast cancer patients from relations of cancer patients irrespective of growth types.

Genomic prediction of disease occurrence using producer-recorded health data: a comparison of methods <sup>10</sup>, discusses of single-trait then two-trait sire models was examined using Bayes A and single-step approaches for mastitis and somatic cell notch. Variance mechanisms were projected. The comprehensive dataset was alienated into exercise and authentication sets to perform perfect comparison. Projected sire upbringing values were used to approximation the amount of daughters probable to develop mastitis. Predictive ability of each model was assessed by the sum of  $\chi^2$  values that associated foretold and observed facts of daughters with mastitis and a number of wrong forecasts.

Gene Change Profiling of Breast Growths for Scientific Decision Creation: Motorists and Travelers in the Cart Beforehand the Mount <sup>11</sup>, discusses topical advances cutting-edge molecular summarizing allow for a rapid and relatively cheap assessment of manifold changed genes or gene crops from small quantities of tumor tissues or gore. The test ahead is how to incorporate these consequences into clinical practice correctly, and in what way to provide patients with the finest possible yet evidence-based care. Herein, the author provides a brief overview of genetic mutation profiling with a focus on next-generation sequencing (NGS) and possible clinical utility.

A single step variance based gene identification method has been discussed in <sup>12</sup>. The method

collects the event data from united state firms. Using the single step gene variance measure the method selects the genes efficiently. FERAL: network-based classifier with submission to breast cancer consequence forecast <sup>13</sup>, has been discussed to improve performance and consistency of discovered markers of the initial molecular classifiers. In malice of the first claims, recent educations exposed that neither presentation nor reliability can be enhanced using these systems. NOPs typically rely on the building of meta-genes by being around the look of several genes linked in a network that encodes protein connections or trail information. In this object, we representation several important issues in NOPs that obstruct on the forecast power, constancy of exposed indicators and confuses organic clarification.

Gene Assortment for the Rebuilding of Stem Cell Differentiation Trees: A Linear Programming Approach <sup>14</sup>, using genetic factor appearance data at both node, we construct a prejudiced Euclidean distance metric such that the smallest spanning tree with admiration to that metric is exactly the given difference hierarchy. We deliver a set of linear restraints that are provably adequate for the wanted building and a linear programming method to classify sparse sets of weights, efficiently identifying genetic factor that is greatest relevant for discerning different shares of the tree.

In <sup>15</sup>, a hybrid gene selection algorithm has been presented. The method selects a sub set of genes from high dimensional data set which are informative. The gene selection is performed using the minimum redundancy score and maximum relevancy score. The method works over support vector machine to evaluate the gene selection performance.

An assortment of a breast cancer subpopulation-specific antibody by means of phage display on tissue sections <sup>16</sup>, prove an approach for phage show selection of recombinant antibody wreckages on cryostat sections of humanoid breast cancer tissue. This technique allows for assortment of recombinant antibodies compulsory to antigens exactly expressed in a small part of the flesh section. In this case, a CD271+ subpopulation of breast tumor cells was targeted, and these may be possible breast cancer stalk cells.

We remote an antibody fragment LH 7, which in resistant histo chemistry trials proves specific binding to breast cancer subpopulations. The selection of antibody rubbles directly in unimportant defined areas within a larger section of malevolent tissue is a novel method by which it is possible to healthier target cellular heterogeneity in proteomic studies.

Intelligent Breast Cancer Diagnosis Using Hybrid GA-ANN, Computational Intelligence <sup>17</sup>, announces an automatic breast cancer judgment technique using a genetic algorithm (GA) for concurrent feature assortment and limit optimization of artificial neural networks (ANN). The recitals of the proposed procedure employing three dissimilar differences of the back propagation method for the fine tuning of the weight of ANN are compared. The algorithm is called the GAANN\_XX where the XX refers to the back-propagation training variation used.

To support the disease prediction o f diabetes, hepatitis and breast cancer an hybrid gene selection approach has been discussed in <sup>18</sup>. The method inherits the features of filter and wrapper feature selection methods. Also the method use the weighted least square twin SVM method for the classification. The method produces efficient results in gene selection and classification.

Comparison of feature selection methods for cross-laboratory microarray analysis <sup>19</sup>, investigate four feature selection methods t-Test, Significance Analysis of Microarrays (SAM), Rank Products (RP) and Random Forest (RF) across breast cancer

and lung cancer microarray data which consists of three cross lab data sets each. Their results show that SAM has the best classification performance. RF also gets high classification accuracy, but it is not as stable as SAM. The Test performance is the worst among the four methods.

In <sup>20</sup>, an binary black hole and random forest ranking algorithms are proposed to improve the performance of gene selection and classification. The method is focused on removing the redundant data from the micro array data sets. The method produces efficient results on gene selection.

**Multi Dimension Protein Impact Matrix Based Protein Sequence Prediction:** To improve the performance of protein structure prediction, a novel protein impact matrix based approach has been discussed. First the method read the protein sequence available from the data set. From the sequence available, a possible combination of sequence within the dimension is generated. For each possible sequence generated, the method computes the number of occurrence in overall sequence set. Using the number of occurrence the method computes the protein impact at each dimension of sequence. This will be iterated for each dimension sequence set and for each dimension sequence of protein, the method computes the protein impact matrix. From the protein impact matrix, the method computes the sequence weight to predict the future sequence. The method improves the performance of sequence prediction and increases the accuracy as well.

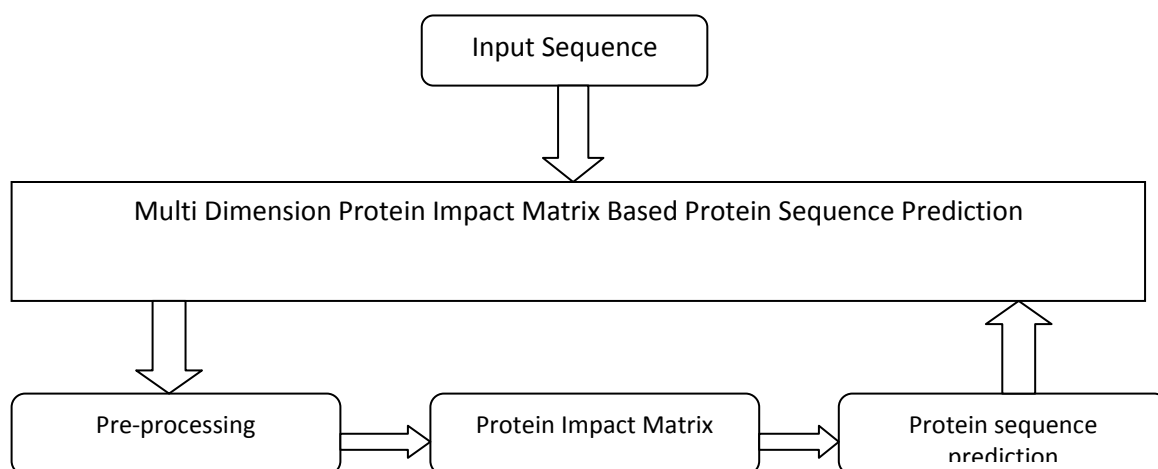


FIG. 1: ARCHITECTURE OF PROTEIN SEQUENCE PREDICTION APPROACH

The Fig. 1 shows the architecture of protein sequence prediction and shows the functional components in detail.

**Pre-processing:** In this stage, the method reads the protein sequences from the data set and for each protein sequence the method verifies the presence

of all dimensions. If there exist any incomplete sequence then that will be removed from the data set. The pre-processing algorithm performs the noise removal operation to enhance the performance of the sequence prediction algorithm.

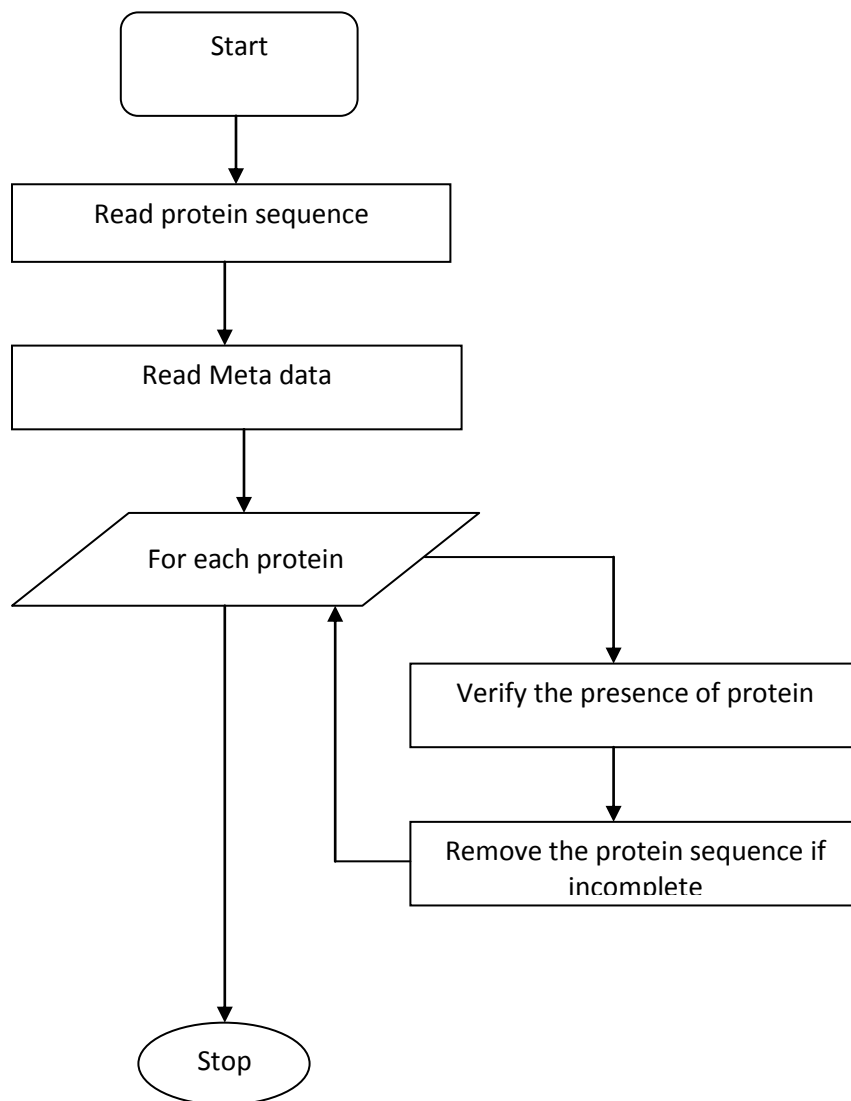


FIG. 2: FLOW CHART OF PRE-PROCESSING

The Fig. 2 shows the flow chart of pre-processing algorithm and shows the details steps.

**Pseudo Code of pre-processing:**

Input: Gene Sequence Set  $G_s$

Output: Gene Set  $G_{es}$

Start

    Read Meta data  $M_d$ .

Identify unique genes  $U_g = \sum_{i=1}^{size(M_d)} G_i(M_d(i)) \neq U_g$

    For reach gene sequence  $G_{si}$

    If  $G_{si}$  contains all genes

    Else

        Remove the sequence.

    End

End

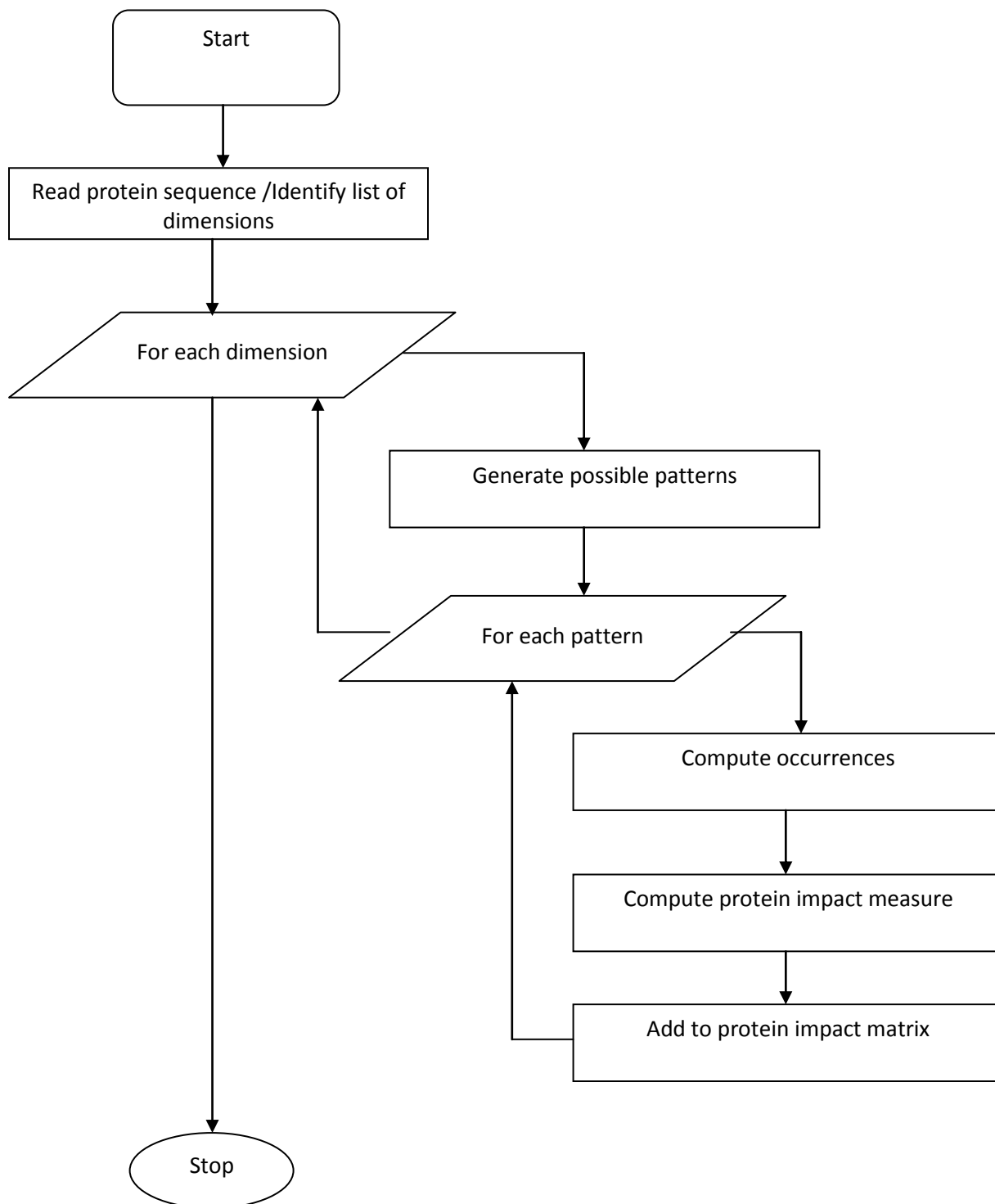
Stop.



The above discussed algorithm identifies the presence of all the genes from the meta data and removes the incomplete genes.

**Protein Impact Matrix Generation:** With the pre-processed sequence set, the method generates number of possible protein sequence. The method generates patterns from 1 to N dimension. For each

pattern generated the method computes the number of occurrence in all the sequence set available in the data set. Using the number of occurrence value computed and the details of the sequence set, the method computes the protein impact value. The computed protein impact value will be stored in the matrix.



**FIG. 3: FLOW CHART OF PROTEIN IMPACT MATRIX GENERATION**

The **Fig. 3** shows the flow chart of protein impact matrix generation and shows the detailed steps.

Pseudo Code of Protein impact Matrix Generation:

Input: sequence S

Output: Impact matrix PIM

Start

    Read sequence S.

Identify the dimension of the sequence  $S_{dim} = \sum Genes \in S$

    For each dimension size Dsize

        Generate possible patterns Ps.

        For each pattern Pi from Ps

            Compute occurrences Oc

            Compute protein impact value iv.

            Add to impact matrix.

        End

    End

Stop.

The above discussed algorithm computes the protein impact value for each of the protein sequence given.

**Protein sequence Prediction:** To perform the future prediction, the method computes the sequence weight for each sequence identified. The method computes the protein weight for each sequence using the others and based on the value the method computes the weight. Based on the weight computed, a single sequence will be identified as the result.

**Pseudo Code:**

Input: Data Set

Output: Protein Sequence ps.

Start

Read data set.

    For each protein sequence

        Verify the presence of all dimensions.

    If incomplete then

        Remove sequence from data set.

End

End

    Compute the dimension D.

    For each dimension from D

        Generate combinatory of possible protein sequence.

        Add to combinations set.

    End

        For each combination from combination set

            Compute number of occurrence.

            Compute protein impact factor.

        Store into the protein impact matrix.

    End

        For each combination

            Compute sequence weight

    End

        Choose the sequence with more weight.

Stop.

The above discussed algorithm performs the prediction of protein sequence from the available sequences.

**RESULT AND DISCUSSION:** In this Paper, an multi dimensional protein impact matrix based gene prediction is presented. The method removes the noisy sequences from the data set and then the method generates the possible sequences. For each sequence the method computes the impact factor and based on that the method computes the weight. Finally based on computed weight, the method selects a single sequence to predict.

**REFERENCES:**

1. Pradipta Maji and Chandra Das, "Relevant and Significant Supervised Gene Clusters for Microarray Cancer Classification", *IEEE Transactions on Nanobioscience*, vol. 11, no. 2, pp. 161-168, 2012.
2. Ujjwal Maulik, Anirban Mukhopadhyay, Debasis Chakraborty, "Gene-Expression-Based Cancer Subtypes Prediction Through Feature Selection and Transductive SVM", *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 4, pp. 1111-1117, 2013.
3. Shuichi Kawano et al., "Identifying Gene Pathways Associated with Cancer Characteristics via Sparse Statistical Methods" *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 9, no. 4, pp. 966-972, 2012.

4. Mark Burton, Mads Thomassen, Qihua Tan and Torben A. Kruse, "Gene Expression Profiles for Predicting Metastasis in Breast Cancer: A Cross-Study Comparison of Classification Methods" *The Scientific World Journal* Volume 2012, Article ID 380495, 11 pages, 2012.
5. Sasikala, A Novel Feature Selection Technique for Improved Survivability Diagnosis of Breast Cancer, *Procedia Computer Science* 50 (2015) 16 – 23.
6. Marina Bessarabova et al., "Bimodal gene expression patterns in breast cancer" *BMC Genomics*, Supplementary 1, 2010.
7. Vitoantonio Bevilacqua et al., "Comparison of data-merging methods with SVM attribute selection and classification in breast cancer gene expression" *BMC Bioinformatics* 2012.
8. Xiaochen Yang, Jiong Wu, Jingsong Lu, Guangyu Liu Identification of a Comprehensive Spectrum of Genetic Factors for Hereditary Breast Cancer in a Chinese Population by Next-Generation Sequencing, *PLoS ONE* 10(4): e0125571. doi:10.1371/journal.pone.0125571.
9. Kristen L Parker Gaddis, Genomic prediction of disease occurrence using producer-recorded health data: a comparison of methods, *Genetics Selection Evolution* 2015, 47:41.
10. Vered Stearns, MD; Ben Ho Park, MD, PhD, Gene Mutation Profiling of Breast Cancers for Clinical Decision Making: Drivers and Passengers in the Cart before the Horse, *JAMA Oncol.* Published online May 14, 2015
11. Parker Gaddis KL, Cole JB, Clay JS, Maltecca C. Genomic selection for producer-recorded health event data in US dairy cattle. *J Dairy Sci.* 2014; 97:3190-9.
12. Amin Allahyar and Jeroen de Ridder, FERAL: network-based classifier with application to breast cancer outcome prediction, *Oxford Journals Science & Mathematics Bioinformatics* Volume 31, Issue 12, Pp. 311-319, 2015
13. Mohamed A. Ghadie, Nathalie Japkowicz1 and Theodore J. Perkins, *Gene Selection for the Reconstruction of Stem Cell Differentiation Trees: A Linear Programming Approach*, Oxford Science and mathematics and Bioinformatics, 2015.
14. Hala Alshamlan, Ghada Badr, and Yousef Alohal, A Hybrid Gene Selection Algorithm for Cancer Classification Using Microarray Gene Expression Profiling, *Hindawi, BioMed Research International* Volume 2015.
15. Simon Asbjørn Larsen, Theresa Meldgaard, Selection of a breast cancer subpopulation-specific antibody using phage display on tissue sections, *Springer, Immunol Res.* 62(3): pages 263–272, 2015.
16. Ahmed F, Intelligent Breast Cancer Diagnosis Using Hybrid GA-ANN, *Computational Intelligence, IEEE, Communication Systems and Networks (CICSyN)*, Page(s):9 – 12, 2013.
17. Divya Tomar and Sonali Agarwal, Hybrid Feature Selection Based Weighted Least Squares Twin Support Vector Machine Approach for Diagnosing Breast Cancer, Hepatitis, and Diabetes, *Hindawi, Advances in Artificial Neural Systems*, Vol. 2015, 2015.
18. Hsi-Che Liu et al., "Comparison of feature selection methods for cross-laboratory microarray analysis" *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2013.
19. Elnaz Pashaei; Gene selection and classification approach for microarray data based on Random Forest Ranking and BBHA, *IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, pp: 308 – 311, 2016
20. Surendar, A., M. Arun, and P. S. Periasamy. "A parallel reconfigurable platform for efficient sequence alignment. *African Journal of Biotechnology* 13.33 (2014): 3344-3351.

**How to cite this article:**

Surendar A: Multi dimension protein impact matrix based protein sequence prediction using data mining. *Int J Pharm Sci Res* 2017; 8(2): 427-34. doi: 10.13040/IJPSR.0975-8232.8(2).427-34.

All © 2013 are reserved by International Journal of Pharmaceutical Sciences and Research. This Journal licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License.

This article can be downloaded to **ANDROID OS** based mobile. Scan QR Code using Code/Bar Scanner from your mobile. (Scanners are available on Google Playstore)