

Distribution based Fuzzy Estimate Spectral Clustering for Cancer Detection with Protein Sequence and Structural Motifs

Thenmozhi K^{1*}, Karthikeyani Visalakshi N², Shanthi S³

Abstract

Objective: In biological data analysis, protein sequence and structural motifs are amino-acid sequence patterns that are widespread and used as tools for detecting cancer at an earlier stage. To improve cancer detection with minimum space and time complexity, Distribution based Fuzzy Estimate Spectral Clustering (DFESC) technique is developed. **Methods:** Initially, protein sequence motifs are taken from dataset to form the cluster. The Distribution based spectral clustering is applied to group the protein sequence by measuring the generalized jaccard similarity between each protein sequence. To develop the clustering accuracy, soft computing technique namely fuzzy logic is applied to calculate membership value of each sequence motif. **Results:** The outcome showed that the presented DFESC technique effectively identifies cancer in terms of clustering accuracy, false positive rate, and cancer detection time and space complexity. **Conclusion:** Based on the observations, evaluation of DFESC technique provides improved result for premature detection of cancer using protein sequence and structural motifs.

Keywords: Protein sequence motifs- cancer detection- distribution based spectral clustering- soft computing

Asian Pac J Cancer Prev, 19 (7), 1935-1940

Introduction

In medical diagnosis, protein sequences determine an amino acid sequence of protein and structure motif is a consecutive residue in polypeptide chain. The protein sequence and structural motif detection is crucial for detecting cancer disease at a former stage. Ranked Neighborhood Comparison (RaNC) method was developed in (Vogt, 2015) for protein structure detection. A multiple sequence alignment tool called GLProbs was designed in (Ye et al., 2015) to improve the precision of protein secondary structure prediction but, it was not accurate.

Systolic arrays (SAs) based Protein sequence alignment technique was presented in (Causapruno et al., 2015) for cancer and hereditary diseases detection. The protein array was established in (Huang and Zhu, 2017) for sensitive cancer biomarker discovery. Co-occurrence-based interaction approaches were introduced in (Zhu et al., 2015) to find out the prostate cancer protein. Feature selection and the taxonomy of protein sequence were performed in (Iqbal et al., 2014) for shrinking the high dimensionality of data for the period of protein structure prediction but, above said methods are failed to detect the different proteomics and genetic diseases.

Supervised and unsupervised clustering algorithms were designed in (Hosseinzadeh et al., 2012) for

classifying the lung cancer tumors. A different mapping and variant-calling methods were introduced in (Atak et al., 2012). A various Machine learning algorithms were designed in (Huang et al., 2015) for predicting the Cancer proteins. Vastly multiplexed proteomic technology (SOMAscan) was introduced in (Mehan et al., 2012) to evaluate the protein expression signatures. But, the complexity of the premature discovery of cancer was high in above mentioned methods.

Materials and Methods

The DFESC technique performs efficient cancer detection at a prior stage using protein sequences and structural motif information's.

Initially, the dataset contains large volume of data points (i.e. protein sequences). These sequences are categorized into different groups using Distribution based Fuzzy Estimate Spectral Clustering. The fuzzy estimated technique used to avoid the overlapping of data between the clusters. Based on the clustering results, the protein sequences are identified either active or inactive. This helps to detect the cancer with minimum time.

Distribution based Fuzzy Estimate Spectral Clustering technique

It is used to detect the cancer by measuring similarity

¹Department of Computer Applications, Selvam College of Technology, Namakkal, ²Department of Computer Science, Government Arts and Science College, Kangayam, ³Department of Computer Applications, Kongu Engineering College, Erode, TamilNadu, India. *For Correspondence: thenmithu@gmail.com

between distributed data points (i.e. protein sequences) in dataset. The DFESC technique uses Eigen values of the similarity matrix of protein sequence motifs for dimensionality reduction before the clustering process. Let us consider the number of protein sequence motifs as input for detecting the cancer disease. Therefore, the data set contains the protein sequences are described as follows,

$$ps = \{ps_1, ps_2, \dots, ps_n\} \in R^d \quad (1)$$

From (1), ps_{ij} denotes a protein sequences which is taken from relational dataset R^d . The similarity matrix is described as the symmetric matrix 'R' which is used to find out the similarity between two sequences ps_i, ps_j defined as follows,

$$D_{ij} = \sum_{i,j}^n R(ps_i, ps_j) \quad (2)$$

From (2), D_{ij} where represent diagonal matrix which includes the degree d_1, d_2, d_3, \dots on the diagonal. The diagonal matrix is described as,

$$D_{ij} = \begin{bmatrix} d_1 & 0 & 0 & 0 \\ 0 & d_2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & d_n \end{bmatrix} \quad (3)$$

From (3) diagonal matrix D_{ij} is constructed. DFESC technique uses Generalized Jaccard similarity measure to discover the similarity between the proteins are measured as follows,

$$R_j(ps_i, ps_j) = \frac{ps_i \cap ps_j}{ps_i + ps_j - (ps_i \cap ps_j)} \quad (4)$$

From (4), $R_s(ps_i, ps_j)$ represents a generalized Jaccard similarity between two protein sequences. A similarity value are used for clustering based on the eigenvector V matching to the second-smallest Eigen value of the symmetric normalized Laplacian function as

$$L_N = D^{-1/2} R(\sqrt{D}) \quad (5)$$

From (5), L_N represents a normalized Laplacian function, D denotes a diagonal, R represents a similarity matrix. Then, the normalized Laplacian matrix (L_{ij}) is constructed through Eigen vectors 'V' and Eigen values 'e'. Therefore, the row of 'V' then normalized is to obtain new matrix is obtained as follows,

$$B_{ij} = \frac{V_{ij}}{\sum_{j=1}^n V_{ij}} \quad (6)$$

From (6), where the rows of 'V' as a collection of 'n' protein sequences in data set, then its applied for cluster the protein sequences. During the clustering process, the soft computing technique namely fuzzy logic is applied to avoid the overlapping of sequences between the clusters. The fuzzy estimation finds the membership of the protein sequence motifs and groups them accordingly. Let us consider the number of sequences, the algorithm returns a list of cluster centers (C) and the partitions matrix is defined as follows,

$$P = \mu_{ij} \in [0,1] \text{ Where } i=1, 2, 3, \dots, n, j=1, 2, 3, \dots, c \quad (7)$$

From (7), where each protein sequences in partition matrix μ_{ij} explains the degree to which the protein sequences $\{ps_1, ps_2, \dots, ps_n\}$ belongs to the cluster C_j and P denotes a partition function. Therefore, the fuzzy technique is used to minimize the following function.

$$\text{argmin}_c \sum_{i=1}^n \sum_{j=1}^c \mu_{ij} \|ps_i - c_j\|^2 \quad (8)$$

From (8), μ_{ij} represents a degree of membership of protein sequence motifs; j denotes a specified number of clusters. The μ_{ij} is described as follows,

$$\mu_{ij} = \frac{1}{\sum_{j=1}^c \left(\frac{D_{ic}}{D_{jc}}\right)^{2/r-1}} \quad (9)$$

From (9), D_{ic} denotes a distance between the protein sequence motif 'i' and the cluster center, D_{jc} represents a distance between the protein sequence motif 'j' and the cluster center. From (9), 'r' denotes the fuzzification factor and it contains the any real number which is greater than one ($r > 1$). Likewise, the spectral clustering group the distribution based protein sequences into two clusters. The algorithmic description of DFESC technique is described as follows,

Input: Number of protein sequences	ps_1, ps_2, \dots, ps_n
Output: Improve clustering accuracy	
Step 1: Begin	
Step 2: For each instances	
Step 3: For each sequence in dataset	
Step 4: Construct the diagonal matrix D_{ij} using (3)	
Step 5: Construct similarity matrix using (4)	
Step 6: Compute Normalized Laplacian function using (5)	
Step 7: Construct the normalized Laplacian matrix using (6)	
Step 8: Minimize the objective function using (8) & (9) to avoid the overlap between cluster	
Step 9: Obtain active or inactive for cancer detection	
Step 10: End for	
Step 11: End for	
Step 12: End	

Algorithm 1 Distribution based Fuzzy Estimate Spectral Clustering

The algorithmic 1 describes process of DFESC is to detect the cancer using protein sequence and structural motifs. For each instances, the similarity is measured using generalized Jaccard similarity to identify active and inactive clusters. Later, the cluster overlapping is avoided by calculating the fuzzy membership value. This in turn improves the clustering accuracy and reduces the cancer detection time.

Results

An experimental evaluation of Distribution based Fuzzy Estimate Spectral Clustering (DFESC) technique is implemented in MATLAB 2015b with Intel i7 processor, 8GB RAM system using p53 Mutants Data Set. P53

Mutants Data Set (<https://archive.ics.uci.edu>) is taken from UCI repository for detecting the cancerous protein sequence. Results and discussion of DFESC technique and existing methods namely RaNC method (Vogt, 2015) and GLProbs (Ye et al., 2015) are described.

The entire simulation task is divided into a set of smaller subtasks with each performed by considering different sets of instances. Hence, the simulation system is carried out as a collection of simultaneous processes, each modeling a different part of the protein sequences and executing on a dedicated MATLAB2015b. The P53 Mutants dataset contains the number of files which defines the instances in the different sets namely K1, K2, K3, K4, K5, K6, K7. The final set K8 defines a full set. P53 Mutants Data Set contains 5409 attributes per instance. The dataset contains 16772 instances and it performs the clustering tasks. DFESC technique is applied to cluster the protein sequences motifs based on the similarity measure. For the simulation consideration, 100 protein sequences are taken as input for first instances. Followed by, different sequences are taken for each instance to detect the cancer using inactive sequence motifs. The dataset contains K1 instance set in which active and inactive sequences are grouped into the sub clusters. Similarly, the other sequences in the instances K2, K3, K4, K5, K6, and K7 are clustered. While considering the K8 full dataset, the sequences are grouped into two clusters (i.e. active or inactive).

Impact of Clustering Accuracy and False Positive Rate

Clustering accuracy is determined by the ratio of number of protein sequences are clustered as active or inactive to the total number of protein sequences in dataset. False positive rate is defined as ratio of number of protein sequences are incorrectly clustered as active or inactive to the total number of sequences in dataset.

Table 1 describes the performance analysis of clustering accuracy and false positive rate based on the number of protein sequences in dataset. The clustering accuracy is considerably increased and reduces a false positive rate than the existing methods. This improvement of proposed DFESC technique is achieved by using fuzzy estimate spectral clustering. The clustering technique is used to distribution based protein sequences

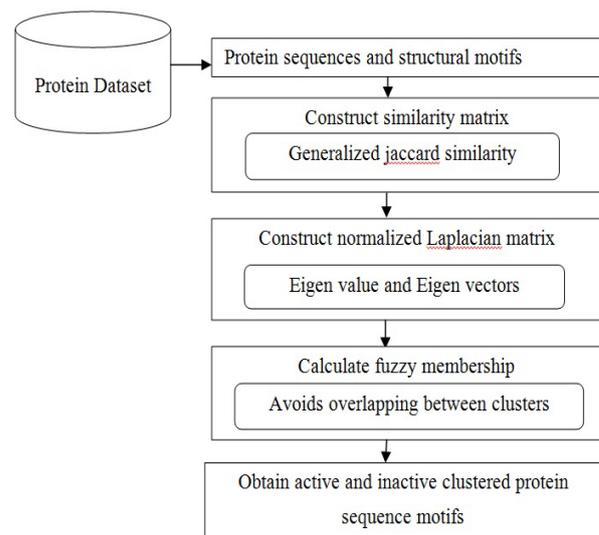


Figure 1. Flow Process of Distribution based Fuzzy Estimate Spectral Clustering

for detecting the cancer at a prior stage. The graphical representation of clustering accuracy and false positive rate is show in Figure 2 and 3.

Figure 2 demonstrates the performance of clustering accuracy with respect to number of protein sequences are varied from 100 to 1000. In order to avoid the overlapping of protein sequences between the clusters, soft computing technique is applied to find the membership value of protein sequences. As a result, the clustering accuracy is significantly increased by 10% and 23% when compared to existing RaNC method (Vogt, 2015) and GLProbs (Ye et al., 2015) respectively.

Figure 3 DFESC groups the sequences into active and inactive clusters for effectively detect the cancer disease. Therefore, the false positive rate is considerably reduced by 31% and 41% when compared to existing RaNC method (Vogt, 2015) and GLProbs (Ye et al., 2015) respectively.

Impact of Cancer detection time and space complexity

Cancer detection time is dogged as the amount of time required for detecting the cancer using clustered protein sequences and structural motifs information's. Space Complexity is discovers by the amount of memory

Table 1. Performance Outcome of Cluster Accuracy and False Positive Rate

No. of protein sequences	Clustering accuracy (%)			False positive rate (%)		
	DFESC	RaNC method	GLProbs	DFESC	RaNC method	GLProbs
100	81.4841	72.3651	60.4192	21.8475	32.1737	41.1950
200	83.4933	73.4398	63.2940	22.6523	35.4816	43.3012
300	84.3392	74.0271	65.4061	23.8295	38.2951	45.4917
400	85.0632	75.3618	67.6270	26.0270	40.3910	47.9037
500	85.7348	77.8263	70.4105	28.3041	43.4916	50.6103
600	88.3847	80.2741	72.1604	30.8363	45.2016	52.7017
700	90.0201	82.0618	75.3817	34.0383	47.1038	53.6192
800	91.2834	82.9371	78.4971	36.1937	50.3910	55.2017
900	93.1865	83.7067	80.2061	38.3391	52.4917	58.3018
1,000	94.7281	86.3731	82.0029	42.4916	55.0173	60.4971

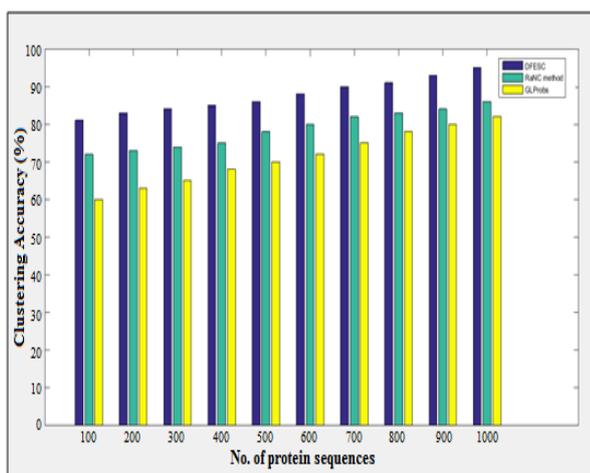


Figure 2. Performance Analysis of Clustering Accuracy

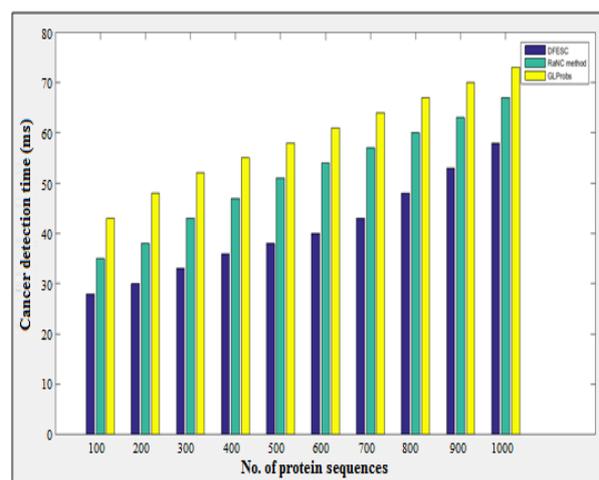


Figure 4. Performance Analysis of Cancer Detection Time

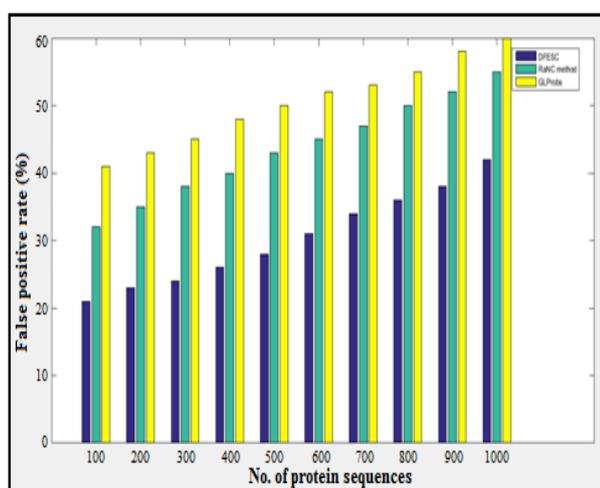


Figure 3. Performance Analysis of False Positive Rate

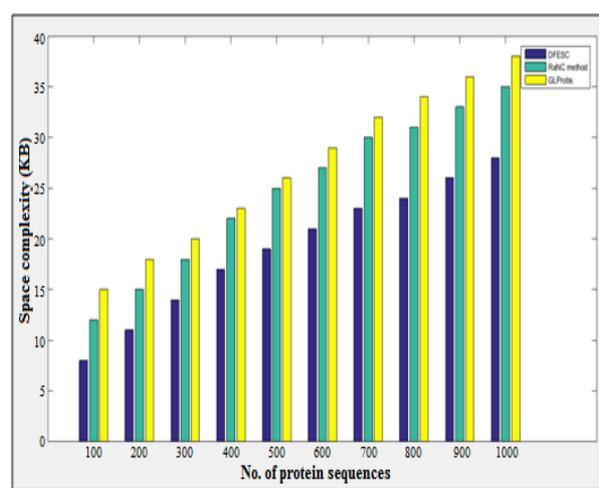


Figure 5. Performance Analysis of Space Complexity

space consumed for storing the clustered protein sequence motifs.

Table 2 shows the effect of cancer detection time and space complexity versus number of protein sequences. Let us consider the number of protein sequences as 100, the proposed DFESC technique requires 28ms for cancer detection whereas RaNC method (Vogt, 2015) and GLProbs (Ye et al., 2015) attains 35ms and 43 ms respectively.

Hence, it is clearly shows that the cancer detection time using proposed DFESC technique is considerably reduced than existing methods (Vogt, 2015; Ye et al., 2015). While considering 100 protein sequences for performing the experiments, the proposed DFESC technique consumes 8KB of storage space whereas RaNC method Vogt, (2015) and GLProbs Ye et al., (2015) required 12KB and 15KB respectively. Therefore, it shows that space complexity using proposed DFESC technique is reduced when

Table 2. Performance Outcome of Cancer Detection Time and Space Complexity

No. of protein sequences	Cancer detection time (ms)			Space complexity (KB)		
	DFESC	RaNC method	GLProbs	DFESC	RaNC method	GLProbs
100	28.2725	35.0381	43.3610	8.5016	12.3917	15.4910
200	30.2743	38.4910	48.1936	11.7105	15.4916	18.3954
300	33.4326	43.3183	51.7285	14.8647	18.4813	20.2913
400	36.2619	47.3195	55.2581	17.4194	22.0200	23.3017
500	38.4914	51.1950	58.3016	19.5285	25.4810	26.4913
600	40.5914	54.3918	61.7159	21.3029	27.3016	29.3071
700	43.2038	57.2910	64.0206	23.5021	30.6104	32.5716
800	48.3017	60.3910	67.5831	36.1937	50.3910	55.2017
900	53.3079	63.3817	70.4916	38.3391	52.4917	58.3018
1,000	58.2018	67.4194	73.4928	42.4916	55.0173	60.4971

compared to other existing works. The Graphical representation of analyzing numerical data of cancer detection time and space is illustrated in Figure 4 and 5.

Figure 4 based on similarity measure, the sequences are grouped into two clusters namely active or inactive cluster. The inactive cluster contains the abnormal gene sequences to detect the cancer with minimum time. The existing ranked neighborhood comparison (RaNC) produces a weighted adjacency matrix for identifying the protein structure using biological data. But it takes more time for detecting the structure of protein using cancer dataset. This issue is addressed by applying DFESC to detect the cancer using protein sequence and structural motifs information. As a result, the cancer detection time is noticeably reduced by 21% and 32% when compared to existing RaNC method Vogt, (2015) and GLProbs Ye et al., (2015), respectively.

Figure 5 illustrate the DFESC technique performs efficient clustering by groups the inactive and active protein sequences. Then these clustering protein sequences are stored and it consumes less storage space than the other existing techniques. As a result, space complexity is considerably abridged by 31% and 21% than existing RaNC method Vogt, (2015) and GLProbs Ye et al., (2015), respectively.

Discussion

A multiscale mutation clustering (M2C) algorithm was developed in (Poole et al., 2017) for discovering changeable length mutation clusters in cancer genes. In (Ye et al., 2010), a mutation in cancer was detected using new statistical approach. Novel serum protein biomarkers were introduced in (Misek and Kim, 2011) for diagnosing breast cancer. The classification of breast cancer protein profiles were carried out in (Velstra et al., 2012). A Hierarchical Clustering was introduced in (Petushkova et al., 2014) for analyzing protein sequence cancer associated liver. But, clustering accuracy was not improved.

A Naive Bayes-based technique was introduced in (Feng et al., 2013) to predict antioxidant proteins an efficient classification algorithm was developed in (Han, 2010) to classify the cancer molecular patterns in microarray data. A protein microarray-based screening method was introduced in (Brezina et al., 2015) for identifying lung cancer. An efficient Incremental Partial Least Squares (IPLS) technique was introduced in (Zeng and Li, 2014). Mutation of specific protein interactions was carried out in (Billur et al., 2016) for tumor detection. But clustering was not used to detect the tumor protein interactions. The above discussion shows that DFESC technique improves the accuracy and cancer detection with minimum time and space complexity.

In conclusion, multiple sequence alignment tool is introduced for arrange the input sequences differently by performing natural measure to calculate the similarity between input sequences. If the input has higher similarity, then the whole sequences align globally. Otherwise, the low similarity input is aligned them locally. Weighted adjacency matrix is used for structure detection

and grouping of data points. But, the model does not implemented in a fixed number of clusters. Therefore, DFESC technique uses Eigen values of the similarity matrix of the protein sequence motifs. Generalized Jaccard similarity measure is popularly used to evaluate the closeness of the data in the process data. Jaccard similarity is a statistical calculation of similarity between sample sets. It is suitable sufficiently to be employed in the protein sequence similarity measurement. Besides, normalized Laplacian matrix is constructed using Eigen vectors and Eigen values in DFESC technique. In addition, soft computing technique of fuzzy logic is applied in the clustering process which computes the membership of each protein sequences by measuring distance between the protein sequence and the cluster center.

DFESC technique is introduced for detecting the cancer with inactive protein sequence and structural motifs information. Initially, the distribution based protein sequences are taken from dataset. The spectral clustering technique is used to distribution based protein sequence and groups the protein sequence according to the similarity between protein sequences. Fuzzy logic is an applied to avoid the overlapping between two clusters. As a result, two clusters are formed to group the active and inactive protein sequences. The cancer is detected using inactive protein sequences with minimum time. Based on the performance results, DFESC technique improves clustering accuracy with minimum space complexity and cancer detection time as well as false positive rate than the state-of-art methods.

References

- Atak ZK, Keersmaecker KD, Gianfelici V, et al (2012). High accuracy mutation detection in leukemia on a selected panel of cancer genes. *PLoS One*, **7**, e38463.
- Billur Engin H, Kreisberg JF, Carter H (2016). Structure-based analysis reveals cancer missense mutations target protein interaction interfaces. *PLoS One*, **11**, e0152929.
- Brezina S, Soldo R, Kreuzhuber R, et al (2015). Immune-signatures for lung cancer diagnostics: Evaluation of protein microarray data normalization strategies. *Microarrays*, **4**, 162-7.
- Causapruno G, Urgese G, Vacca M, et al (2015). Protein alignment systolic array throughput optimization. *IEEE T VLSI Syst*, **23**, 68 -7.
- Feng P, Lin H, Chen W (2013). Identification of antioxidants from sequence information using naive bayes. *Comput Math Methods Med*, **2013**, 1-5.
- Han X (2010). Nonnegative principal component analysis for cancer molecular pattern discovery. *IEEE/ACM Trans Comput Biol Bioinform*, **7**, 537- 9.
- Hosseinzadeh F, Ebrahimi M, Goliaei B, Shamabadi N (2012). Classification of lung cancer tumors based on structural and physicochemical properties of proteins by bioinformatics models. *PLoS One*, **7**, e40017.
- Huang C-H, Peng H-S, Ng K-L (2015). Prediction of cancer proteins by integrating protein interaction, domain frequency, and domain interaction data using machine learning algorithms. *BioMed Res Int*, **2015**, 1-5.
- Huang Y, Zhu H (2017). Protein array-based approaches for biomarker discovery in cancer. *Genomics Proteomics Bioinformatics*, **15**, 73-1.
- Iqbal MJ, Faye I, Samir BB, Said A (2014). Efficient feature selection and classification of protein sequence data in *Asian Pacific Journal of Cancer Prevention*, Vol 19 **1939**

- bioinformatics. *Sci World J*, **2014**, 1-2.
- Mehan M, Ayers D, Thirstrup D, et al (2012). Protein signature of lung cancer tissues. *PLoS One*, **7**, e35157.
- Misek D, Kim E (2011). Protein biomarkers for the early detection of breast cancer. *Int J Proteomics*, **2011**, 1-9.
- Petushkova N, Pyatnitskiy M, Rudenko V, et al (2014). Applying of hierarchical clustering to analysis of protein patterns in the human cancer-associated liver. *PLoS One*, **9**, e103950.
- Poole W, Leinonen K, Shmulevich I, et al (2017). Multiscale mutation clustering algorithm identifies pan-cancer mutational clusters associated with pathway-level changes in gene expression. *PLoS Comput Biol*, **13**, e1005347.
- P53 Mutants Data Set: <https://archive.ics.uci.edu/ml/datasets/p53+Mutants>.
- Velstra B, van der Burgt Y, Mertens B, et al (2012). Improved classification of breast cancer peptide and protein profiles by combining two serum workup procedures. *J Cancer Res Clin Oncol*, **138**, 1983-92.
- Vogt JE (2015). Unsupervised structure detection in biomedical data. *IEEE ACM T Comput BI Journal*, **12**, 753-60.
- Ye J, Pavlicek, Lunney EA, Rejto PA, Teng C (2010). Statistical method on nonrandom clustering with application to somatic mutations in cancer. *BMC Bioinformatics*, **11**, 1-9.
- Ye Y, Cheung DW, Wang Y, et al (2015). GLProbs: Aligning multiple sequences adaptively. *IEEE ACM T Comput BI Journal*, **12**, 67-8.
- Zeng X, Li G (2014). Dimension reduction for p53 protein recognition by using incremental partial least squares. *EEE Trans Nanobioscience*, **13**, 73-9.
- Zhu F, Liu Q, Zhang X, Shen B (2015). Protein interaction network constructing based on text mining and reinforcement learning with application to prostate cancer. *IET Syst Biol*, **9**, 106-2.



This work is licensed under a Creative Commons Attribution-Non Commercial 4.0 International License.