Inverted Indexing for Information Retrieval from Motifs and Domains of Proteins

Kumud Pant, Bhasker Pant, Devvret Verma, Promila Sharma, Vikas Tripathi

Abstract**:** ***The recent advancement in technologies are generating huge amount of data and extracting information from it is being outpaced by data accumulation. The development of hybrid approaches by combining different algorithms for extraction of required from the stock-pile of data is a demand of the hour. One such algorithm is vector space model for inverted indexing that has been used traditionally for search engine indexing in computers. In bioinformatics also it has been used for assembly of DNA fragments generated after sequencing. But it has not been applied for retrieval of relevant protein sequence to the query, based on presence or absence of motifs and domains in it. In this paper the concept of inverted indexing has been applied on small motif/domain data of proteins contained in Motivated Proteins database at http://motif.gla.ac.uk/motif/index.html. The index has been built using 17 small hydrogen bonded motifs present in a dataset of 430 proteins. The entire dataset of 430 proteins has been divided into 19 classes. Seven classes' example cyanovirin, antibiotic and concavalin etc. had very few instances (1 or 2), hence have been omitted from further studies. Rest 12 classes with more than 10 proteins were considered further for testing information retrieval (IR) strategy. The document vector of all the proteins belonging to one class was averaged and 12 queries with averaged vector were prepared for testing. The similarity coefficient (SC) was then compared between query and all the proteins of the dataset. This approach could successfully classify the query as belonging to the class from which it derived. To further validate the importance of document vector as novel attribute for classification, entire dataset of document vector was clustered to ten (10) clusters. Testing was then performed with similarity coefficient (SC) of the query with clusters obtained above. The allocation of cluster to the 12 query sequences followed the same pattern as done with relevant document search using inverted indexing approach. But clustering allocated the queries to only four (4) classes. Maximum number of query proteins (7 proteins or 58%) were found belonging to cluster 5.***

Keywords **:** ***Information Retrieval, Motif/Domain, Clustering, Inverted Indexing***
Computing Classification System: I.4

Revised Manuscript Received on January 30, 2020.
* Correspondence Author

Kumud Pant*, Department of Biotechnology, Graphic Era Deemed to be University, Dehradun, India.

Bhasker Pant, Department of Computer Science & Engineering Graphic Era Deemed to be University, Dehradun, India.

Devvret Verma, Department of Biotechnology, Graphic Era Deemed to be University, Dehradun, India.

Promila Sharma, Department of Biotechnology, Graphic Era Deemed to be University, Dehradun, India.

Vikas Tripathi, Department of Computer Science & Engineering Graphic Era Deemed to be University, Dehradun, India.

I. INTRODUCTION

In a world laden with data, retrieval of useful data or information has been most demanding, as evident from development of new algorithms, models and software. Concepts from various branches have been incorporated for relevant data (Information) retrieval from biological data warehouse. This becomes especially important since all modern day high throughput technologies are outpacing information generation over data accumulation. The realm of information retrieval here involves picking a document (out of bulk) that is most similar to the input query. There may be many documents sharing great amount of similarity with query, but through similarity coefficient the weight or relevance of ranking can be used to ascertain the one most near to the query [1]. Traditionally IR has been used in digital libraries, media search, search engines and many more. Recently it has started to be used in biomedical literature, the information of which has been compiled on the Biomedical Literature Mining Publication (BLIMP) [2].

From a biologist point of view getting the highest match of a documents (protein or DNA or RNA sequence or macromolecular structure) with the query can be of great value. The Entrez search engine of NCBI devised by NCBI's John Wilbur and several bioinformatics including BLAST and FASTA search performs the same task of match finding [1]. However, specialized algorithms have been formulated as a part of information retrieval strategy. The most notable are Vector Space Model, Probabilistic Retrieval Strategies, Inference Networks, Language Models, Neural Networks, Extended Boolean Retrieval, Latent Sementic Indexing, Genetic Algorithms and Fuzzy Set Retrieval.

The above mentioned approaches use document vectors or vector space model for relevant document search. Document vectors represent frequency of occurrence of a term in the document. The inverted document frequency (idf) is then used to assign weight to each term in the query. The greater number of times a term appears in document the lesser is its weight, and the lesser times it appears the more is its weight.

Proteins, as biological entity, are macromolecules composed of multiple motifs and domains with each motif or domain present multiple times. With same motifs and domains present in both query and database an ideal biological information retrieval strategy can be formed. Although IR strategy for DNA alignment has been developed in the past with nucleotide sequence as attribute but no IR search with domain and motif information has been done till now [3].

To understand the role of frequency of occurrence of motif and domain information in relevant protein search, the vector space model has been implemented on protein small motif dataset obtained from http://motif.gla.ac.uk/. The Motivated Protein database contains a list of 430 proteins with various frequencies of 17 small hydrogen bonded motifs.

As a part of IR strategy an attempt have been made at generating inverted index for motif or domain based information of protein dataset obtained from Motivated Protein database [4]. The frequency of motif/domain information of 430 proteins multiplied with their inverse document frequency produced a document vector (DV). The DV constituted the training dataset. For testing the model, twelve (12) queries were made by averaging the document vector of all proteins belonging to a class in the training dataset itself. Hence it was a case of self-validation. The similarity coefficient (SC) between query and 430 proteins in the training dataset was calculated. The queries were said to belong to the class with which the similarity coefficient was the highest. Ten (10) out of twelve (12) proteins were found to belong to the respective classes from where derived but two queries were not able to locate their respective classes on the basis of similarity coefficient.

To analyze the working of inverted indexing and similarity coefficient based approach for protein classification, the entire dataset of 430 proteins with 17 attributes was considered without class label and was used to train an unsupervised clustering algorithm. On testing with document vector of 12 queries the document vector based classifier allocated the 12 query proteins into 4 classes. Maximum number of query proteins (7 proteins or 58%) were found belonging to cluster 5.

The unsupervised learning based clustering algorithm has been used on the above dataset to highlight the importance of document vector in partitioning data. Although SC based approach has been successful in classifying queries into their respective classes, application of unsupervised learning algorithm further strengthened the power of attribute frequency based methods for the same.

II. MATERIAL AND METHODS

1. The motif data set obtained from http://motif.gla.ac.uk/ was used to retrieve the comprehensive list of the small motifs found in 430 proteins obtained from PDB database.

2. Weka suite of software at http://www.cs.waikato.ac.nz, is a software for performing various data mining activities [5]. Clustering was performed in protein motif dataset for assigning groups to previously unsupervised data.

III. RESULTS

Initially an Inverted Index indicating the presence as well as frequency of all the 17 small hydrogen bonded motifs/domains in 430 proteins was built. The 17 motifs taken for the study are Alpha Beta Loop, Asx motif, Asx Turn, Beta Bulge, Beta Bulge Loop, Beta Bulge Turn, Beta Turn, Bulge Loop Motif, Crown Bridge, Crown Bridge Loop, Nest, Niche, Schellman Loop, ST Loop, ST Motif, ST Staple and ST Turn. The proteins have been taken from various classes and have various disease implications like cancer, schizophrenia and TB to name a few. The construction of inverted index avoids lengthy sequential scan through every document to find terms in the query. It is a look up table for every term occurring in the document.

The inverse document frequency of each term was thereafter calculated with the formula idf = log (d/dfi), where d is number of documents (430) and idf is the occurrence frequency of each motif in the entire document set. The idf for all the17 motifs is shown in Table I.

Table- I: The inverse document frequency for all 17 motifs

Table Column Head		
S. No.	***Motif***	***Idf*** [a]
1.	Alpha Beta Loop	0.67
2.	Asx motif	0.106
3.	Asx Turn	0.093
4.	Beta Bulge	0.279
5.	Beta Bulge Loop	0.302
6.	Beta Bulge Turn	0.666
7.	Beta Turn	0.0128
8.	Bulge Loop Motif	0.108
9.	Crown Bridge	1.047
10.	Crown Bridge Loop	1.2
11.	Nest	0.019
12.	Niche	0.0128
13.	Schellman Loop	0.171
14.	ST Loop	0.265
15.	ST Motif	0.156
16.	ST Staple	0.106
17.	ST Turn	0.15

[a] idf : inverse document frequency

A document vector with seventeen terms for few proteins is shown in Table II. It reflects the importance of each term appearing in the document and is obtained by multiplying the idf with frequency of its occurrence in each document.

Retrieval Number: C8044019320/2020©BEIESP
DOI: 10.35940/ijitee.C8044.019320
Journal Website: www.ijitee.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Table- II: Document vector for first 23 proteins

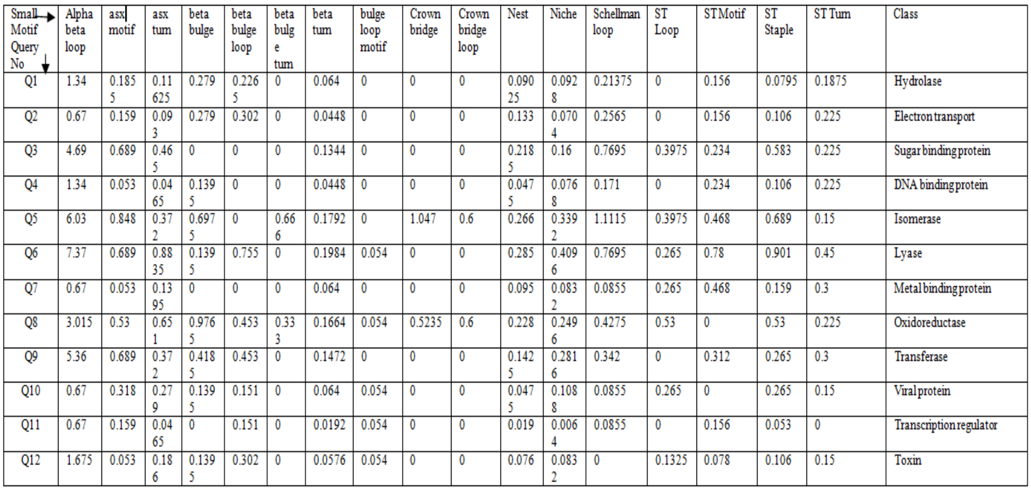
Motif → Protein ID ↓	ABL	AM	AT	BB	BBL	BBT	BT	BLM	CB	CBL	NT	N	SL	STL	STM	STS	STT
1BS9	0.4164	0.308197355	0.343001414	0	0	0	0.084950284	0	0	0	0.051222716	0.139824133	0.210810083	0.644420767	0.292308335	0.442972849	0
1BDO	0	0	0	1.138421981	0	1.294775921	0.050970171	0	0	0	0	0.038133855	0	0	0	0	0.0935
2ACT	0.2776	0.205464903	0	0.487895135	0.178199489	0	0.067960228	0	0	0	0.089639753	0.076267709	0.210810083	0.161105192	0.097436112	0.316409178	0.0935
1BTE	0	0.102732452	0.085750353	0.162631712	0	0	0.025485085	0	0	0	0.025611358	0.038133855	0	0	0	0	0
2ACY	0.1388	0	0	0.162631712	0	0	0.016990057	0	0	0	0.064028395	0.02542257	0.105405041	0.161105192	0	0.189845507	0.0935
1MUN	0.7634	0.102732452	0	0	0	0	0.059465199	0	0	0	0.102445432	0.146179776	0.210810083	0	0.292308335	0.379691013	0
1QB7	0.4164	0.102732452	0.085750353	0	0	0	0.042475142	0	0	0	0.102445432	0.152535418	0.105405041	0.483315575	0.487180559	0.126563671	0.187
2DPM	1.1104	0.308197355	0.343001414	0.162631712	0.178199489	0	0.050970171	0.332648823	0	0	0.102445432	0.088978994	0.105405041	0	0.584616671	0.189845507	0.0935
1BX4	0.9716	0.205464903	0.171500707	0	0	0	0.084950284	0	0.420050399	0.445688336	0.115251111	0.120757206	0.421620166	0	0.194872224	0.316409178	0.0935
1QHV	0	0.61639471	0.514502121	0.162631712	0.178199489	0	0.084950284	0.332648823	0	0	0.038417037	0.101690279	0	0.322210383	0	0.126563671	0.187
1ZIN	0.4164	0.205464903	0.085750353	0	0.356398978	0	0.042475142	0	0	0	0.140862469	0.050845139	0.421620166	0	0.389744447	0.189845507	0
1QS1	0.2776	0.719127161	0.171500707	0.97579027	0	0.971081941	0.152910512	0	0	0	0.153668148	0.190669273	0.316215124	0.96663115	0.292308335	0.316409178	0.374
1OUW	0	0	0.171500707	1.138421981	0.178199489	0	0.033980114	0	0	0	0.025611358	0.02542257	0	0	0	0	0.0935
1BD0	0.4164	0.410929806	0.343001414	0.813158558	0	0.647387961	0.135920455	0	0.420050399	0	0.12805679	0.152535418	0.527025207	0.161105192	0.389744447	0.442972849	0.0935

3BTO	0.6246	0.410929806	0.600252474	0.487895135	0.178199489	0.32369398	0.152910512	0	0.420050399	0.445688336	0.243307902	0.184313631	0.421620166	0.483315575	0	0.506254684	0.2805
1DOS	0.9716	0.205464903	0.171500707	0	0	0	0.067960228	0	0	0	0.166473827	0.076267709	0.421620166	0	0.097436112	0.696100191	0.187
1ADS	0.5552	0.719127161	0.25725106	0	0.178199489	0	0.11043537	0	0	0	0.243307902	0.165246703	0.632430249	0.161105192	0.194872224	0.316409178	0.187
1VJS	0.7634	1.130056967	0.943253888	0.487895135	0.356398978	0.32369398	0.144415484	0.665297646	0	0	0.294530618	0.222447485	0.316215124	0.322210383	0.097436112	0.442972849	0.374
1PEN	0.2082	0	0	0	0	0	0.008495028	0	0	0	0	0.012711285	0	0	0	0	0
1QQ4	0.0694	0	0.171500707	1.626317116	0.534598467	1.294775921	0.11043537	0	0	0	0.064028395	0.076267709	0.105405041	0	0	0	0.0935
1TUD	0	0	0.171500707	0.325263423	0	0.32369398	0.016990057	0	0	0	0	0.044489497	0	0	0	0	0
1MRJ	0.5552	0.513662258	0.514502121	0.162631712	0.534598467	0	0.084950284	0.665297646	0.420050399	0.445688336	0.076834074	0.108045921	0	0.322210383	0.097436112	0.759382027	0
1AAC	0	0	0.085750353	0.325263423	0	0.647387961	0.050970171	0	0	0	0.012805679	0.057200782	0	0	0	0	0

ABL: Alpha beta loop, AM: Asx motif, AT: Asx Turn, BB: Beta Bulge, BBL: Beta Bulge Loop, BBT: Beta Bulge Turn, BT: Beta Turn BLM: Bulge Loop Motif, CB: Crown Bridge, CBL: Crown Bridge Loop, NT: Nest, N: Niche, SL: Schellman Loop, STL: ST Loop, STM :ST Motif, STS: ST Staple, STT: ST TURN

To further implement this vector space model a set of twelve (12) hypothetical queries was prepared that are shown in Table III. The proteins in the Motivated Protein database have been obtained from PDB database at www.rcsb.org [6]. The PDB classified the 430 proteins into more than 19 classes. Few classes had one or two proteins only, therefore, in an alternative strategy only 12 protein classes out of 19, from PDB having more than 10 instances, were considered for similarity coefficient (SC) based analysis. The document vectors of all proteins belonging to one class were averaged. The averaged vector comprised a query vector for testing. Following the similar protocol 12 queries from 12 protein classes were prepared. Now the similarity coefficient of 12 queries was calculated.

Table- III: Twelve (12) queries prepared by averaging document vector of all proteins belonging to one class



Small → Motif Query No ↓	Alpha beta loop	asx motif	asx tum	beta bulge	beta bulge loop	beta bulge tum	beta tum	bulge loop motif	Crown bridge	Crown bridge loop	Nest	Niche	Schellman loop	ST Loop	ST Motif	ST Staple	ST Tum	Class
Q1	1.34	0.1855	0.11625	0.279	0.2265	0	0.064	0	0	0	0.09025	0.0928	0.21375	0	0.156	0.0795	0.1875	Hydrolase
Q2	0.67	0.159	0.093	0.279	0.302	0	0.0448	0	0	0	0.133	0.0704	0.2565	0	0.156	0.106	0.225	Electron transport
Q3	4.69	0.689	0.465	0	0	0	0.1344	0	0	0	0.2185	0.16	0.7695	0.3975	0.234	0.583	0.225	Sugar binding protein
Q4	1.34	0.053	0.0465	0.1395	0	0	0.0448	0	0	0	0.0475	0.0768	0.171	0	0.234	0.106	0.225	DNA binding protein
Q5	6.03	0.848	0.372	0.6975	0	0.666	0.1792	0	1.047	0.6	0.266	0.3392	1.1115	0.3975	0.468	0.689	0.15	Isomerase
Q6	7.37	0.689	0.8835	0.1395	0.755	0	0.1984	0.054	0	0	0.285	0.4096	0.7695	0.265	0.78	0.901	0.45	Lyase
Q7	0.67	0.053	0.1395	0	0	0	0.064	0	0	0	0.095	0.0832	0.0855	0.265	0.468	0.159	0.3	Metal binding protein
Q8	3.015	0.53	0.651	0.9765	0.453	0.333	0.1664	0.054	0.5235	0.6	0.228	0.2496	0.4275	0.53	0	0.53	0.225	Oxidoreductase
Q9	5.36	0.689	0.372	0.4185	0.453	0	0.1472	0	0	0	0.1425	0.2816	0.342	0	0.312	0.265	0.3	Transferase
Q10	0.67	0.318	0.279	0.1395	0.151	0	0.064	0.054	0	0	0.0475	0.1088	0.0855	0.265	0	0.265	0.15	Viral protein
Q11	0.67	0.159	0.0465	0	0.151	0	0.0192	0.054	0	0	0.019	0.0064	0.0855	0	0.156	0.053	0	Transcription regulator
Q12	1.675	0.053	0.186	0.1395	0.302	0	0.0576	0.054	0	0	0.076	0.0832	0	0.1325	0.078	0.106	0.15	Toxin

The SC based approach successfully retrieved relevant proteins to the query. The top three similarity coefficient based relevant documents for the query are shown in Table IV.

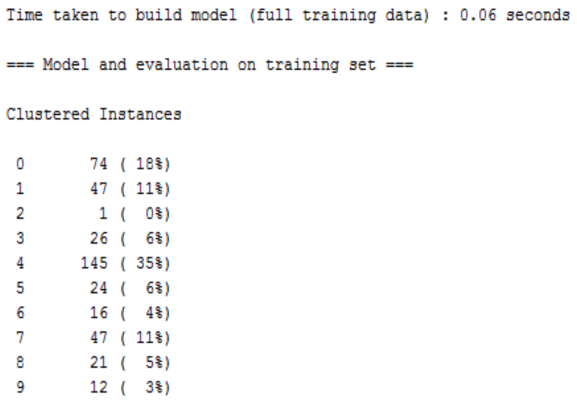
The similarity coefficient (SC) reflecting relevance of the query with each of the 430 sequences in the database was calculated by multiplying document vector of each sequence with query vector shown in table 2. On summing up the similarity coefficient (SC) for every document, the highest SC was given the greatest significance or a particular query was found belonging to the document having the highest SC.

Table- IV: Relevant document allocation on the basis of similarity coefficient

Query (Initial Class)	Similarity coefficient (highest similarity with protein ID) (Class)	Similarity coefficient (second highest similarity)	Similarity coefficient (third highest similarity)
Q1 (Hydrolase)	1SMD (19.42) (Hydrolase)	1VJS (15.30) (Hydrolase)	1CS1 (14.62) (Lyase)
Q2 (Electron transport)	1PLC (19.082) (Electron Transport)	1RIE (15.0058) (Electron Transport)	1JER (14.47) (Electron Transport)
Q3 (Sugar Binding Protein)	1OUW (3.266)(Sugar Binding Protein)	1A7S (3.212) (Ligand Binding Protein)	2IGD (3.198) (Igg Binding Protein)
Q4 (DNA Binding Protein)	1C1K (17.78) (DNA Binding Protein)	3HTS (15.17) (DNA Binding Protein)	1AAC (14.5214) (Electron Transport Protein)
Q5 (Isomerase)	1QRE (19.18) (Lyase)	1HMT (14.97) (Lipid Binding Protein)	1AYL (14.31) (Lipid Binding Protein)
Q6 (Lyase)	4XIS (17.85) (Isomerase)	1QIP (15.02) (Lyase)	3STD (14.50) (Lyase)
Q7 (Metal Binding Protein)	5ICB (19.30) (Metal Binding Protein)	1CYD (15.05) (Oxidoreductase)	1A8E (14.59) (Metal Transport)
Q8 (Oxidoreductase)	1CYD (18.25) (OxRd)	1B4V (15.19) (OxRd)	1ADS (14.57) (OxRd)
Q9 (Transferase)	1QB7 (19.28) (Transferase)	1A8D (15.19) (Toxin)	1BGF (14.57) (Transcription Regulator)
Q10 (Viral Protein)	1EGW (19.19) (Transcription Regulator)	1XWL (15.19) (Tranferase)	1YTB (14.42) (Transcription Rregulator)
Q11 (Transcription regulator)	1MOF (16.2481863) (Viral Protein)	1MFI (15.2345) (Viral Protein)	1BKB (15.05) (Translation Initiation Factor)
Q12 (Toxin)	3SEB (17.234) (Toxin)	1KPT (17.04567) (Toxin)	1PEN (15.0534) (Toxin)

For further validating the importance of document vector as classification attribute the document vector for all 430 proteins were clustered using WEKA 3.8.1. The dataset was considered to be unsupervised, since no class labels were initially provided. The k-means method of clustering was adopted for the study since it produces tighter clusters with large amount of attributes [7]. With document vector for 430 proteins as training dataset, ten (10) clusters were prepared as shown in Figure 1.

The test data set comprised of query vectors for all the 12 queries. The allocation of cluster to the 12 query sequences followed the same pattern as done with relevant document search using inverted indexing approach. But clustering allocated the queries to only four (4) classes. The results for clustering analysis are shown in Figure 2. Maximum number of query proteins (7 proteins or 58%) were found belonging to cluster 5



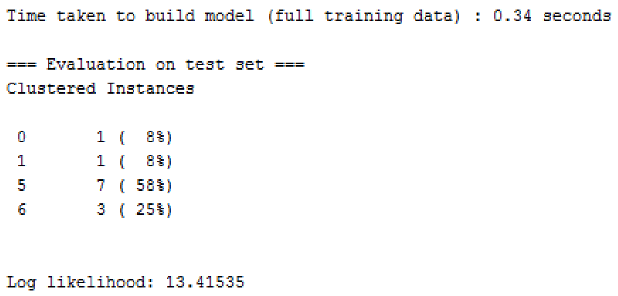
```
Time taken to build model (full training data) : 0.06 seconds

=== Model and evaluation on training set ===

Clustered Instances

 0       74 ( 18%)
 1       47 ( 11%)
 2        1 (  0%)
 3       26 (  6%)
 4      145 ( 35%)
 5       24 (  6%)
 6       16 (  4%)
 7       47 ( 11%)
 8       21 (  5%)
 9       12 (  3%)
```

Fig. 1. Cluster information obtained from Weka 3.8.1 (Eibe Frank et.al.; 2013)



```
Time taken to build model (full training data) : 0.34 seconds

=== Evaluation on test set ===
Clustered Instances

 0        1 (  8%)
 1        1 (  8%)
 5        7 ( 58%)
 6        3 ( 25%)


Log likelihood: 13.41535
```

Fig. 2. Cluster allocation for 12 query proteins

IV. DISCUSSION AND CONCLUSION

Initially the 413 proteins downloaded from http://motif.gla.ac.uk/ on careful observation were found to belong to more than 19 different classes. Few classes had very few instances example class Cyanovirin had only 1 instance. Class antibiotic and Concavalin had 2. Therefore, only those classes were considered for analysis that had more than 10 protein instances. There were twelve (12) classes that fulfilled this criterion, as shown in table 3. The document vector of proteins belonging to one class was averaged. This was repeated for all 12 classes to generate 12 queries. The similarity coefficient of all the queries with all the proteins was generated. 10 protein queries namely were able to locate their respective class of origin. On careful observation it was found that query 5 and query 6 obtained from class Isomerase and Lyase shared document vector at many places. Similarly document vector for query 10 and query 11 of viral protein and transcription regulator shared many values. Hence queries on the basis of similarity coefficient could not be distinguished into two classes.

To analyze the importance of document vector as an attribute for classification, the dataset was made to undergo unsupervised learning in the form of clustering. The training of k-means algorithm for clustering was performed with document vectors as attributes. The classifier clustered the dataset into 10 instances.

On testing Weka clustering algorithm with vectors of query, the k-means algorithm allocated the query into 4 groups as shown in figure 2. Although clustering based classifier could not very well distinguish the query into respective groups, it could partly be due to less demarcation between document vectors. Hence document vector can prove to be an effective attribute for protein classification. This approach can also help in classifying a new protein with motif or domain information, as belonging to a particular class.

The implementation of vector space model using inverted index can be very well applied to big datasets. It can avoid lengthy sequential search through every document to find the most relevant one to the query.

ACKNOWLEDGMENT

The authors express the deep sense of gratitude to the Department of Biotechnology, Department of Computer Science & Engineering of Graphic Era Deemed to be University for all the support, assistance, and constant encouragements to carry out this work.

REFERENCES

1. Nadkarni, P. M., 2002, An introduction to information retrieval: applications in genomics. The Pharmacogenomics Journal. 2(2), 96–102.
2. Biomedical Literature Mining Publication (BLIMP) at http://blimp.cs.queensu.ca/.
3. Ling, W., Zhao KaiYong., 2013, A new DNA alignment method based on inverted index. At https://arxiv.org/abs/1307.0194v1.
4. Motivated Protein database at http://motif.gla.ac.uk/.
5. Frank, E., Mark A. Hall., Ian H. Witten., 2016, The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016. The Weka suite of software at http://www.cs.waikato.ac.nz.
6. RCSB Protein Data Bank at www.rcsb.org.
7. Technology for you at http://playwidtech.blogspot.in/2013/02/k-means-clustering-advantages-and.html.

AUTHORS PROFILE

Dr. Kumud Pant is working as an Assistant Professor in the Department of Biotechnology, Graphic Era University, Dehradun, Uttarakhand, India. She has completed her Masters in Biotechnology from Jiwaji University, Gwalior and PhD in Bioinformatics from MANIT, Bhopal in 2014. She got air fare Grant from CSIR for attending IC2IT held at Bangkok, Thailand in 2009. She is a supervisor for two (2) PhD Scholar. She has successfully completed a project entitled "Reverse vaccinology approach for detecting major virulent proteins of Encephalitis virus", granted by UCOST, Dehradun. Her area of research is Bioinformatics, Computational Biology and Data Mining. She has more than 25 research publication in reputed journals.

Dr. Bhasker Pant completed his graduate from M.B. Khalsa College, Indore, post graduate from Barkatullaha University, Bhopal, and doctoral studies at Maulana Azad National Institute of Technology, Bhopal, India in 2000, 2003 and 2012 respectively. He has done significant research work on Data Mining, Machine Learning, and Bioinformatics. His research interest includes Internet-of-Things and soft-computing too. He has 7 doctoral students and supervised more than 20 masters students. He has delivered invited talks, guest lectures and presented his research results in various counties like China, Singapore and India. He has more than 40 publications in top quality international conferences and journals. He is holding senior level academic positions the university time to time. He is currently the Dean (Research and Development) in Graphic Era University. He is a recipient of the Eminent Research Award (Computer Science & Engineering Engineering) of the Graphic Era University.

Devvret Verma is a research scholar in the Department of Biotechnology, Graphic Era Deemed to be University Dehradun. He did B.Tech in Bioinformatics from Amity University, Uttar Pradesh and M.Tech. in Biotechnology from Graphic Era Deemed to be University, Dehradun. He has written 2 book chapters and published 5 Research Papers in peer reviewed journals. He has got a Young Scientist Award through UCOST in 11th Science Congress. He had secured second Best Oral Presentation award in conference organized by Uttaranchal University Dehradun. His area of research is Bioinformatics, Drug Designing and Data Mining.

Dr. Promila is working as an Associate Professor in Department of Life Sciences, Graphic Era University, Dehradun, Uttarakhand, India. She did her B.Sc (H) Human Biology from AIIMS, New Delhi. She completed her postgraduation in Biochemistry from Jamia Hamdard, New Delhi and Ph.D from Department of Biotechnology, Jamia Hamdard, New Delhi only. Her area of research is Cell and Molecular Biology. Currently she is working on NO and nitric oxide synthase. She has written three book chapters and published 13 research papers in peer reviewed journals. She is also member of editorial board of International Journal of Plant Biotechnology.

Dr. Vikas Tripathi has done BE in information technology from Technocrats institute of technology, Bhopal, M. Tech in Software engineering from Indian institute of information technology Gwalior and PhD from Uttarakhand technical university, Dehradun. He is actively involved in research related to Software engineering, Computer Vision, Machine learning and Video Analytics. He has published many papers in reputed international conferences and journals. Currently he is working as an associate professor in Graphic era deemed to be university Dehradun, India.