PROTEIN SUB-CELLULAR LOCALIZATION PREDICTION USING THRESHOLD ML-DT

A Pre-Dissertation

Proposal Submitted

To

Department of Computer Science and Engineering

In partial fulfilment of the Requirement for the

Award of the Degree of

MASTER OF TECHNOLOGY

in

COMPUTER SCIENCE & ENGG.

by

JASPREET KAUR

Reg. No. 41500009

Under the guidance of

Ms. Komal Arora

School of Computer Science and Engineering

Lovely Professional University

Phagwara, Punjab (India)

December, 2017.

December, 2017.

TOPIC APPROVAL PERFORMA

School of Computer Science and Engineering

Program : P172-T::M.Tech. (Computer Science and Engineering) [Part Time]

COURSE CODE : CSEP548 **REGULAR/BACKLOG :** Regular **GROUP NUMBER :** CSERGD0362

Supervisor Name : Komal Arora **UID :** 17783 **Designation :** Assistant Professor

Qualification : _____ **Research Experience :** _____

SR.NO.	NAME OF STUDENT	REGISTRATION NO	BATCH	SECTION	CONTACT NUMBER
1	Jaspreet Kaur	41500009	2015	K1521	9463408953

SPECIALIZATION AREA : Database Systems **Supervisor Signature:** _____

PROPOSED TOPIC : Algorithmic bioinformatics

Qualitative Assessment of Proposed Topic by PAC		
Sr.No.	**Parameter**	**Rating (out of 10)**
1	Project Novelty: Potential of the project to create new knowledge	5.33
2	Project Feasibility: Project can be timely carried out in-house with low-cost and available resources in the University by the students.	5.67
3	Project Academic Inputs: Project topic is relevant and makes extensive use of academic inputs in UG program and serves as a culminating effort for core study area of the degree program.	5.33
4	Project Supervision: Project supervisor's is technically competent to guide students, resolve any issues, and impart necessary skills.	5.67
5	Social Applicability: Project work intends to solve a practical problem.	5.67
6	Future Scope: Project has potential to become basis of future research work, publication or patent.	5.33

PAC Committee Members		
PAC Member 1 Name: Kewal Krishan	UID: 11179	Recommended (Y/N): Yes
PAC Member 2 Name: Raj Karan Singh	UID: 14307	Recommended (Y/N): NA
PAC Member 3 Name: Sawal Tandon	UID: 14770	Recommended (Y/N): NA
PAC Member 4 Name: Dr. Pooja Gupta	UID: 19580	Recommended (Y/N): NO
PAC Member 5 Name: Kamlesh Lakhwani	UID: 20980	Recommended (Y/N): Yes
PAC Member 6 Name: Dr.Priyanka Chawla	UID: 22046	Recommended (Y/N): NA
DAA Nominee Name: Kuldeep Kumar Kushwaha	UID: 17118	Recommended (Y/N): NA

Final Topic Approved by PAC: Algorithmic bioinformatics

Overall Remarks: Approved **PAC CHAIRPERSON Name:** 11024::Amandeep Nagpal **Approval Date:** 04 Nov 2017

ABSTRACT

Predicting the appropriate protein subcellular localization has attracted much attention in the field of bioinformatics for determining the cellular function of proteins. Several traditional biochemical experimental methods have been developed to determine the protein subcellular localization is expensive and time-consuming and during the last decade, many computational based methods have been developed to predict the protein subcellular localization in different organisms. However, most of the methods succeeded to predict proteins in only one subcellular location but there are many proteins, which have two or more subcellular locations. To predict subcellular localization of protein with multiple sites, Threshold ML-DT (Multi-label Decision Tree algorithm) will be used. Threshold-MLDT performs much better than ML-DT and produces better prediction accuracy in much lesser time.

DECLARATION

I hereby declare that the pre-dissertation proposal entitled, '**PROTEIN SUB-CELLULAR LOCALIZATION PREDICTION USING THRESHOLD ML-DT'** submitted for M.Tech. degree is entirely my original work and all ideas and references have duly acknowledged. It does not contain any work for the award of any other degree or diploma.

Date: _____ **Investigator**

Reg. No. 41500009

CERTIFICATE

This is certifying that Jaspreet Kaur has completed M.Tech. Pre-dissertation proposal titled **'PROTEIN SUB-CELLULAR LOCALIZATION PREDICTION USING THRESHOLD ML-DT'** under my guidance and supervision. To the best of my knowledge, the present work is the result of her original investigation and study. No part of the dissertation has ever been submitted for any diploma and degree.

The dissertation is fit for the submission and partial fulfilment the conditions for award of the M.Tech. Computer Science & Engg.

Signature of Advisor

Ms. Komal Arora

Date:----------

Counter Signed by:

1) **Concerned HOD:**
   HoD's Signature: _____
   HoD Name: _____
   Date: _____

2) **Neutral Examiners:**

   External Examiner

   Signature: _____

   Name: _____

   Affiliation: _____

ACKNOWLEDEGEMENT

I express my deep gratitude to my honourable guide Ms. Komal Arora, Department of Computer Science and Engineering, Lovely Professional University, Phagwara for providing me stimulating guidance, continuous encouragement and support throughout the preparation of thesis.

I would also like to thank all the faculty members, LPU Phagwara and all my friends who were always there at the need of the hour and provided all the help and support, which I required for the completion of the thesis.

Last but not the least; I would like to thank God for not letting me down at the time of crisis and showing me the silver lining in the dark clouds.

JASPREET KAUR

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER – 1

INTRODUCTION

1.1 PROTEIN SUB-CELLULAR LOCATION PREDICTION

In the modern era, researchers studies proteomics to understand the biological processes at the cellular level. The main goal in proteomics is protein subcellular localization and that has received a lot of attention recently because it is the main functional attribute of a protein. Information about the sub-cellular location of proteins is important to determine how they interact with each other and with other molecules and knowing the information about protein subcellular localization is crucial to understand not only the function of proteins but also the organization of the whole cell [1]. However, it is time-consuming, expensive, and laborious process.

Nowadays, many more proteins are found, and it makes the existing methods more unpractical to implement [3]. Recent advancements in proteomics and genomics methods bought a scientific revolution and resulted in large accumulation of proteins whose functions are unknown. Due to this advancement, many powerful tools for predicting the subcellular location of proteins and number of methods have been proposed to predict the protein subcellular localization [2]. There is a need to use an algorithm, which can identify multi sites subcellular localization fast and reliable.

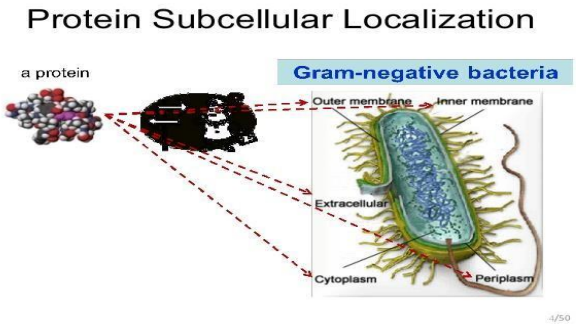

Figure 1.1 Protein Sub-cellular Localization [10]

1.2 MULTI-LABEL LEARNING

Multi-label learning discusses the problem in which each example is depicted by a single feature space instance, which is concerned with a set of labels [13]. Multi-label Learning is an advanced learning in which the classification algorithm learns different class labels from a set of instances, in which single instance can be taken from multiple classes and results to predict a set of class labels for a new feature instance [6].

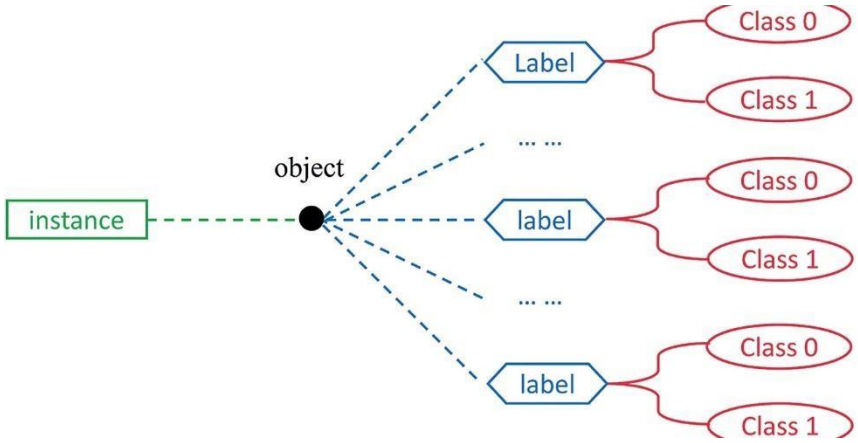

Figure 1.2 Multi-label Classification [15]

In other words, we can say that multi-label classification or multi-label learning is the problem of finding a model that maps inputs **x** to some other value **y** and results into the creation of new instance.

1.3 MULTI-LABEL CLASSIFICATION ALGORITHMS

There are many multi-label classification algorithms exist [12],[13]. Multi-label classification algorithms can be broadly classified into following categories:

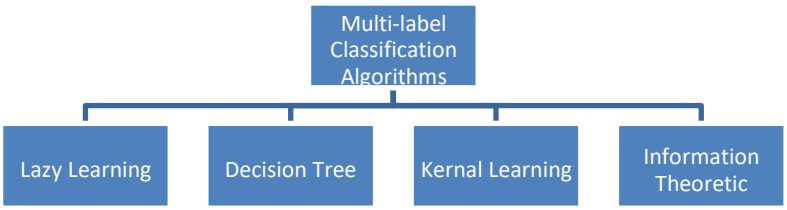

Figure 1.3 Algorithms for Multi-label Classification

These algorithms are described as under:

1.3.1 LAZY LEARNING: It is a study of classification which delays the training of data until there is a need to classify a particular instance i.e. protein. There are many lazy learning approaches e.g. KNN (k-nearest neighbour). The core approach of these different approaches are same in the sense that each of them are using KNN algorithm as a lazy learning method.

1.3.2 DECISION TREE: The idea behind decision tree approach designed with the aim to reduce the computational complexity at each level and thus making the learning algorithm efficient especially with many labels. There are two categories of decision tree models, these are:

a) Tree based boosting

b) Random decision tree

1.3.3 KERNAL LEARNING: Kernel based learning are a well-defined tool to analyse the relationship between input data and the resultant output of a given function. Kernels allow algorithms to swap functions of differnt complexity.

1.3.4 INFORMATION THEORETIC: Statistical Classification is the problem of identifying to which of a set of categories a new observation belongs, based on a training set of data containing observations (or instances) who is category membership or a value is known.

1.4 PROTEIN FEATURE EXTRACTION METHODS

1.4.1 Protein Feature Extraction Methods

In order to analyze the protein sequences and to use them in ML-DT (Multi-label Learning Decision Tree Algorithm) and Threshold multi-label learning in Decision Tree (Threshold-MLDT) algorithm, one of the main challenge is to extract the proteins feature. The three feature extraction methods that are used in this process are [1]:

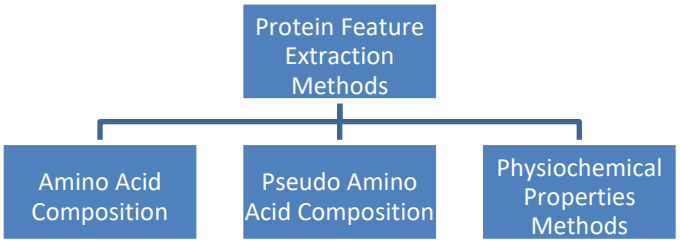

Figure 1.4 Categories of Protein Feature Extraction Methods

1.4.1.1 Amino Acid Composition (AAC): In this method, the protein sequence is encapsulated into 20 vectors and the vector composition can be calculated as:

$$AAC(i) = \frac{N_i}{N} \qquad i=1,2,..20 \qquad (1)$$

Where N_i is the number of amino acids of type i, N is the protein sequence length [1]

1.4.1.2 Pseudo Amino Acid Composition (PseAA): In this method, different amino acid sequences are encapsulated to create a new proteomics instance different from the Amino Acid Composition (AAC). The PseAA is more accurate than the traditional AAC. It can be calculated as

$$P=[p_1,p_2,\ldots,p_{20},\ldots,p_{20+n}] \text{ , } (n<N) \quad (2)$$

Where P is the protein and p_1, $p_2\ldots$ are the components [1].

1.4.1.3 Physiochemical Properties (PC): By using the physiochemical properties of proteins, we can achieve more accuracy in predicting the features of proteins. This method is somehow similar to Amino Acid Composition methods and the difference lies in how the protein features are extracted in this PC method. So, this can be formulated as,

$$f_{i,polar} = n_{polar}/N \; [i=1,2,\ldots,7] \qquad (3)\ [1]$$

$$f_{i,nuetral} = n_{neutral}/N \; [i=1,2,\ldots,7] \qquad (4)\ [1]$$

$$f_{i,hydrophobic} = n_{hydrophobic}/N \; [i=1,2,\ldots,7] \qquad (5)\ [1]$$

CHAPTER – 2

LITERATURE REVIEW

U.Subhashini et al. (2017) puts forward a new algorithm named as Threshold ML-KNN (MultiLabel Learning using k Nearest Neighbor) to predict sub-cellular localization of protein with multiple sites. This algorithm has been derived from the existing classification k Nearest Neighbor algorithm. The Threshold value considered in KNN has improved the prediction accuracy of the ML-KNN algorithm and that improved accuracy can be achieved in lesser time [1].

Shalini Kaushik et al. (2017) This research article, put stress on the protein extraction using the machine learning. The research shown that characteristics such as geonomics, protein functional can improve the prediction accuracy [2].

Kirtan Dave et al. (2017) In this paper, Dave, has suggested a method to predict locations of human proteins using SVM. The tool used in this paper, predicts locations with high accuracy and low false positives. The online interface of the tool is under development, but the present version of the tool can be used to predict the proteins location [3].

Xumi Qu et al. (2016) In this paper, the author discussed feature extraction methods like N-Terminal Signals, PAAC (Pseudo Amino Acid Composition) etc. The author took the best characteristics of different multi-label classification algorithms and combined them to devise a better prediction method with the implementation of ML-KNN algorithm. [4].

Jing Chen et al. (2014) By considering the limitations of protein sub-cellular localization prediction methods Chen suggested a tree based model FHML (Fuzzy Hypergraph Multi-Label Learning) for both one and multi-location proteins. Chen created a fuzzy hypergraph to create relations between different set of prediction values, and then find location of proteins and label the non-labelled protein sequences [5].

Xiao Wang et al. (2013) proposed an algorithm and named it as, Random Label Selection (RALS) derived from multi-label learning, that expand Binary Relevance algorithm used to predict protein subcellular locations. Experimental results are then compared on different data sets and high prediction accuracy has been achieved in multi-site location prediction values [6].

Guo-Zheng Li et al. (2012) In this paper, the author, presented and then compared two multi-label learning approaches, which identify co-relations between different class labels to get high prediction accuracy. The method used by the author achieves better performance [7].

Qian Xu et al. (2011) In this paper, Qian formulated an algorithm for protein subcellular localization problem to implement on multi-task learning approaches to apply across different organisms like humans etc. Qian adapted and compared two areas of the multi-task learning approaches on 20 various organisms and their experimental results show that multi-task learning performs much better as compared to conventional methods [8].

Tien-ho Lin, Robert F. Murphy and Ziv Bar-Joseph et al. (2011) In this paper, the author describes the discriminative motif finding which is used to predict protein sub-cellular localization and then also fused discriminative motif finding with hierarchical structure that imitate the protein sorting process. With this approach the author, find the potential bugs in online databases for the location of proteins [9].

Kuo-Chen Chou, Hong Bin Shen et al. (2010) In this paper, the author, uses Euk-mPLoc 2.0 tool to predict the sub-cellular localization to find the protein locations with both one and multiple location sites. The tool Euk-mPLoc is freely available on the internet and it is a novel approach to get highly accurate results [10].

Hong-Bin Chen and Kuo-Chen Chou et al. (2010) In this paper, the author, finds the protein sub-cellular location in multiple sites using a fusion classifier to find viral proteins. The protein sub-cellular location found using the Virus-mPLoc tool [11].

Trias Thireou and Martin Reczko et al. (2007) proposed an efficient algorithm called Bidirectional Long Short-Term Memory Networks (BLSTM) for transforming linear data. In this algorithm, a neural network was trained to use long-range same sequences using collaboration of non-linear processing of elements and feedback cycle for storing different context. The proposed algorithm applied to the sequence-based prediction of protein localization and predicts 93.3% non-plant proteins and 88.4% plant proteins accurately and results in an improvement on feed forward and networks solving the same problem [12].

Grigorios Tsoumakas, Ioannis Katakis et al. (2005) This paper introduces the different classification algorithms used for multi-label learning. This also provides comparison between different classification algorithms and list out the strength and weaknesses of various algorithms [13].

CHAPTER – 3

PROBLEM FORMULATION

3.1 SCOPE

Protein sub-cellular localization is an important study on proteomics, geonomics, cytobiology, drug analysis etc. There are so many fields of biology where the use of computational methods results in drastic advancements in biomedical sciences and results in great benefits to medical field [4].

Prediction of sub-cellular localization of proteins is very important for understanding the biological functions and processes as well as discovering new drugs [10]. When the proteins locate into the specific sub-cellular locations, they can perform their appropriate functionality. There are so many conventional methods exits, which can be used for protein sub-cellular localization prediction like MLKNN, SVM etc [9]. The Decision Tree approach is one of the simple, accurate and efficient approach that can be used to locate the sub-cellular locations in proteins [11].

3.2 PROBLEM DEFINITION

The concept of multi-label learning can be represented as, let us consider X represents the protein sequences for protein subcellular localization and let Y={λ1, λ2,…, λ m} be the well-defined set of feasible class labels. Given a training set T={(x1,Y1), (x2,Y2),…,(xi,Yi)}(xi belongs to X and Yi is a proper sub-set of Y). The goal of the learning system is to learn a multi-label classifier h:X->2^y which identifies some labeled classifier proteins . The function can be represented using relation of X and Y. the real valued function can be formulated as, f: X x Y -> R [7].

3.3 OBJECTIVES

1) To predict sub-cellular localization of protein with multiple sites.
2) To achieve better performance than the existing algorithms.
3) To get better prediction accuracy in less time.

3.4 EXPECTED OUTCOMES

Protein sub-cellular localization prediction is an area of interest where location prediction accuracy is highly important as if the proteins are located into the specific locations then they can result into better functionality and biological processes can be studied efficiently. So, this work is expected to result with an algorithm, which results into the better performance with improved accuracy of existing methodology.

CHAPTER – 4

METHODOLOGY

1) Conventional Decision Tree algorithm will be applied on proteins with single site and on the proteins with multiple sites inputted into algorithm.
2) The algorithm will be fused with the threshold value to predict the protein sub-cellular locations.
3) The accuracy can be checked by the equation,

   Accuracy (x) = p(x) / expr (x) [3] (6)

   Where, x can be any sub-cellular location, expr (x) is the total no. of sequences observed, and p (x) is the no. of correctly predicted sequences.
4) Analyse the performance parameters like, prediction accuracy, prediction time and efficiency and then compare the results with existing methodologies for protein location prediction.

4.1 FLOWCHART OF PROPOSED TECHNIQUE

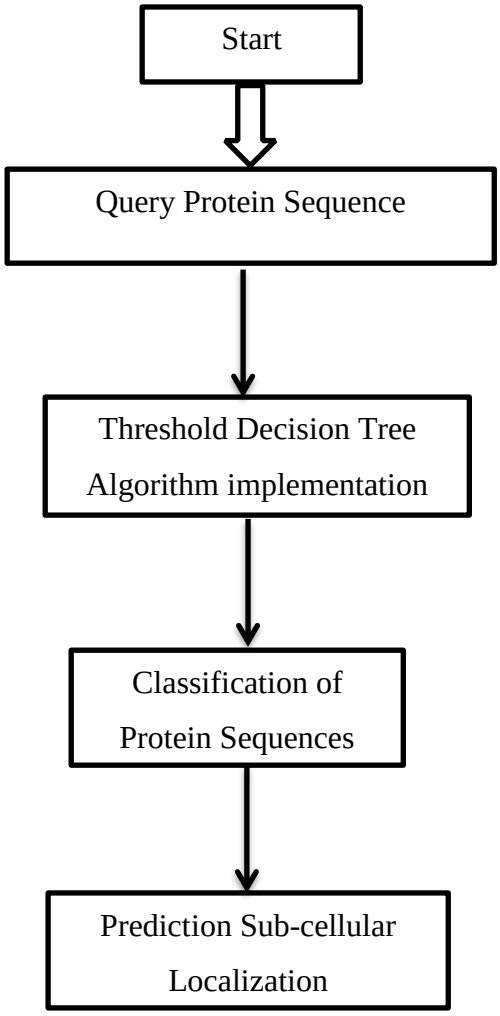

Figure 1.5 Flowchart of the proposed technique

CHAPTER – 5

CONCLUSION

Experimental techniques for the prediction and study of protein sub-cellular localization are hot research area of this modern era. Prediction of the multi-sites proteins locations is a challenging problem. Transforming different facts from the biological operations to computational version and over coming from the problems is a challenging task.

The main objective of this study is to find a method for efficient subcellular localization with more accuracy in lesser time. The Decision Tree (DT) classification algorithm can be used to classify proteins sub-cellular locations and its fusion with threshold value will result in improved accuracy and better performance. Prediction of multi-sites protein sub-cellular localization would help to study biology in a better way.

REFERENCES

[1] U. Subhashini, P. Bhargavi, S. Jyothi, D.M. Mamatha, "Predicting Sub-cellular Localization of Proteins with Multiple Sites using Threshold ML-KNN", Int J Pharma Bio Sci 2017 July; 8(3): (B) 278-285

[2] Shalini Kaushik, Usha Chouhan, Ashok Diwedi "Study of Protein Sub-cellular Localization Prediction-A Review", International Journal of Life Science and Pharma Research, Vol 7/Issue 3/ 2017, ISSN 2250-0480

[3] Kirtan Dave, Nikita Patel, Dr. Hetal Panchal, "Prediction of Sub-cellular Localization of Human Protein by Support Vector Machine", under consideration for Young Scientist of India award.

[4] Xumi Qu, Dong Wang, Yuehui Chen, Shanping Qiao, Qing Zhao, "Predicting the Subcellular Localization of Proteins with Multiple Sites Based on Multiple Features Fusion", IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol 13, No. 1, January/February 2016.

[5] Jing Chen, Yuan Yan Tang, C.L. Philip Chen, Bin Fang, Yuewei Lin, Zhaowei Shang, "Multi-label Learning With Fuzzy Hypregraph Regularization for Protein Sub-cellular Location Prediction", IEEE Transactions of Nano Bio Science, Vol 13, No. 4, December 2014.

[6] Xiao Wang, Guo-Zheng Li, "Multi-label Learning via Random Label Selection for Protein Sub-cellular Multi-locations Prediction", IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol 10, No. 2, March/April 2013.

[7]Guo-Zheng Li, Xiao Wang, Xiaohua Hu, Jia-Ming Liu, Rui-Wei Zhao, "Multi-label Learning for Protein Sub-cellular Location Prediction", IEEE Transactions on Nano Bio Science, Vol 11, No. 3, September 2012.

[8] Qian Xu, Sinno Jialin Pan, Hannah Hong Xue, Qiang Yang, "Multi-task Learning for Protein Sub-cellular Location Prediction", IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol 8, No. 3, May/June 2011.

[9] Tien-ho Lin, Robert F. Murphy and Ziv Bar-Joseph, "Discriminative Motif Finding for Predicting Protein Subcellular Localization", IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol 8, No. 2, March-April, 2011.

[10] Kuo-Chen Chou, Hong-Bin Shen, “A New Method for Predicting the Subcellular Localization of Eukaryotic Proteins with Both Single and Multiple Sites: Euk-mPLoc 2.0”, IEEE/ACM Transactions on Computational Biology and Bioinformatics, February, 2010.

[11] Hong-Bin Shen and Kuo-Chen Chou, “Virus-mPLocc: A Fusion Classifier for Viral Protein Subcellular Location Prediction by Incorporating Multiple Sites”, Journal of Biomolecular Structure & Dynamics, ISSN 0739-1102 Volume 28, Issue Number 2, (2010).

[12] Min-Ling Zhang and Zhi-Hua Zhou, “A Review on Multi-Label Learning Algorithms”, IEEE/ACM Transactions on Computational Biology September, 2007.

[13] Grigorios Tsoumakas, Ioannis Katakis, “Multi-Label Classification: An Overview”, IEEE/ACM Transactions on Computational Biology , 2006.

[14] https://www.slideshare.net/JIAMINGCHANG/2008-0117-psldocslides

[15] http://pubs.rsc.org/en/content/articlelanding/2014/ay/c4ay01240b/unauth#!divAbstract