ANALYZING DNA SEQUENCES USING CLUSTERING ALGORITHM

TAHA TALEB RAGHEB ALHERSH

UNIVERSITY UTARA MALAYSIA
2009

ANALYZING DNA SEQUENCES USING CLUSTERING ALGORITHM

A thesis submitted to college Arts & Sciences
in partial fulfillment of the requirement for the degree
Master of Science (Intelligent Systems)
University of Utara Malaysia

By
Taha Alhersh

PERMISSION TO USE

In presenting this thesis in partial fulfillment of the requirements for a Master of Science in IT degree from University Utara Malaysia, I agree that the University Library may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purpose may be granted by my supervisor or, in their absence by the Academic Dean College of Arts and Sciences. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to University Utara Malaysia for any scholarly use which may be made of any material from my thesis.

Requests for permission to copy or to make other use of materials in this thesis, in whole or in part, should be addressed to

Dean (Academic) College of Art and Sciences
University Utara Malaysia
06010 UUM Sintok
Kedah Darul Aman.

ABSTRACT

Data mining gives a bright prospective in DNA sequences analysis through its concepts and techniques. This study carries out exploratory data analysis method to cluster DNA sequences. Feature vectors have been developed to map the DNA sequences to a twelve-dimensional vector in the space. Lysozyme, Myoglobin and Rhodopsin protein families have been tested in this space. The results of DNA sequences comparison among homologous sequences give close distances between their characterization vectors which are easily distinguishable from non-homologous in experiment it with a fixed DNA sequence size that does not exceed the maximum length of the shortest DNA sequence. Global comparison for multiple DNA sequences simultaneously presented in the genomic space is the main advantage of this work by applying direct comparison of the corresponding characteristic vectors distances. The novelty of this work is that for the new DNA sequence, there is no need to compare the new DNA sequence with the whole DNA sequences length, just the comparison focused on a fixed number of all the sequences in a way that does not exceed the maximum length of the new DNA sequence. In other words, parts of the DNA sequence can identify the functionality of the DNA sequence, and make it clustered with its family members.

ACKNOWLEDGEMENT

I would like to start with the words that any job will not be complete without, so I will say: “In the Name of Allah, the Beneficent, the Most Merciful”. All the thanks to Allah that pave the way for me to obtain my master degree.

My sincere appreciation goes to my supervisor AP Fadzilah Bt Siraj for her supervision, guidance, advice, knowledge and word of encouragement during this study, I’ll always be thankful to you, Terimah kasih!

Special thanks to all my lecturers Dr. Ahmad Otoom, Dr. Yuhanis binti Yusof, Dr. Norita Md Norwawi, Dr. Fawziah Ahmad, Mrs. Nur Azzah Abu Bakar, Dr. Shaidah Jusoh, Mr. Zhamri Che Ani and other UUM staff.

I must acknowledge the immeasurable contributions of my friends and collegues who have shown great love and care during my study especially Qais, Abdullateef, Hossam, Ali, big Thank you!

This acknowledgment won’t be complete without my family. Profound gratitude goes to my parents Taleb and Shifa who have never failed to give me the best in life. May Allah reward your efforts! To my brothers Faisal, Firas, Fadi, Abdullrahman and Abdullraheem, you never stopped in encouraging me, special thanks and appreciation to my brother Fadi for his continuous support. My family, you will be always in my heart, love you all.

DEDICATION

To my parents Taleb and Shifa, and to my brothers.

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATION

A	Adenine
BLAST	Basic Local Alignment Search Tool
bp	Base pairs
C	Cytosine
COI	Cytochrome “C” Oxidase I
D_A	The distribution of A in the DNA sequence
D_C	The distribution of C in the DNA sequence
D_G	The distribution of G in the DNA sequence
DNA	Deoxyribonucleic Acid
D_T	The distribution of T in the DNA sequence
EM	Expectation-Maximization
FCM	5 Fuzzy C-Means
G	Guanine
GBS	Global Bio-identification System
ILP	Inductive Logic Programming
IR	Information Retrieval
KDD	Knowledge Discovery in Database
KNIES	Kohonen Incorporating Explicit Statistics
LCC	Library of Congress Classification
LVQ	Learning Vector Quantization
MST	Minimal Spanning Tree

n_A	Number of instances A in the DNA sequence
n_C	Number of instances C in the DNA sequence
NCBI	National Center for Biotechnology Information
n_G	Number of instances G in the DNA sequence
NLP	Natural Language Processing
n_T	Number of instances T in the DNA sequence
PEs	Processing Elements
RNA	Ribonucleic Acid
SOM	Self Organizing Map
SVMs	Support Vector Machines
T	Thymine
T_A	The total distances of A from the origin of DNA sequence
T_C	The total distances of C from the origin of DNA sequence
T_G	The total distances of G from the origin of DNA sequence
TIS	Translation Initiation Sites
TSP	Travelling Salesman Problem
TSPLIB	Travelling Salesman Problem Library
TSS	Transition Split Site
T_T	The total distances of T from the origin of DNA sequence
VQ	Vector Quantization

LIST OF APPENDICES

Appendix

CHAPTER ONE

INTRODUCTION

This chapter introduces a brief description of this study. A general overview of the field of this work, problem statement, the objective and the scope of this study has been presented.

In the last few decades the rapid development of technology reflects to the number of biological data which has been growing in an exponential curve, from Gene Bank (www.ncbi.nlm.nih.gov) site the growth falls down in Fig.1.1. GenBank in 1982 had only 606 sequences with 680,338 bp (base pairs). In year 1992, GenBank contained 78,608 sequences with 101,008,486 bp. By the end of year 2002, GenBank had 22,318,883 sequences with 28,507,990,166 bp. This number had almost doubled in only two years. By the end of year 2008, GenBank had 98,868,465 sequences with 99,116,431,942 bp. Efficient and highly computational tools are needed to analyze the massive amount of data that contains rich information.

The contents of the thesis is for internal user only

REFERENCES

Abonyi, J., & Feil, B. (2005). Computational Intelligence in Data Mining. *Informatica, 29,* 3-12.

Aksoy, S., & Haralick, R. M. (1999). Graph–Theoretic Clustering for Image Grouping and Retrieval. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'99), 1*, 1063.

Anastassiou, D. (2000). Frequency-domain analysis of biomolecular sequences. *Bioinformatics, 16*(4), 1073-1081.

Ansari, A., & Viswanathan, R. (1992). Application of Expectation-Maximization Algorithm to the Detection of Direct-Sequence Signal in pulsed Noise Jamming. *IEEE Military Communications Conference, 3*, 811-815.

Apon, A., Mache, J., Buyya, R., & Jin, H. (2004). Cluster Computing in the Classroom and Integration with Computing Curricula 2001. *IEEE Transactions on Education, 47*(2), 188-195.

Arasa, N., Oommenb, B. J., & Altınelc, I. K. (1999). The Kohonen network incorporating explicit statistics and its application to the travelling salesman problem. *Neural Networks, 12*(9), 1273-1284.

Ayre, L. B. (2006). *Data Mining for Information Professionals*.

Bach, F. R., & Jordan, M. I. (2003). Learning Spectral Clustering. *Learning graphical models with Mercer kernels in Advances Neural Inform, 1,* 1009-1016.

Bolshoy, A., & Volkovich, Z. (2008). Whole-genome prokaryotic clustering based on gene lengths. *Discrete Applied Mathematics, 157*(10), 2370-2377.

Borman, S. (2009). *The Expectation Maximization Algorithm A short tutorial*.

Carvalho, F. A. T. (2006). *Fuzzy clustering algorithms for symbolic interval data based on adaptive and non-adaptive Euclidean distances*.

Draghici S., Graziano, F., Kettoola, S., Sethi, I., & Towfic, G. (2003). Mining HIV dynamics using independent component analysis. *Bioinformatics, 19*(8), 981-986.

Erban, G., & Moldovan, G. S. (2006). A Comparison of Clustering Techniques in Aspect Mining. *Informatica, 1*, 69-78.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). *From Data Mining to Knowledge Discovery in Databases*.

FitzGerald, P. C., Shlyakhtenko, A., Mir, A. A., & Vinson, C. (2004). Clustering of DNA Sequences in Human Promoters. *Genome Res, 14*, 1562-1574.

Gates, M. A. (1985). Simpler DNA sequence representations. *Nature, 31*, 219.

Ghanem M., Chortaras, A., Guo, Y., Rowe, A., & Ratcliffe, J. (2005). *A Grid Infrastructure for Mixed Bioinformatics Data and Text Mining.*

Graham, J., Page, C. D., & Kamal, A. (2003). *Accelerating the Drug Design Process through Parallel Inductive Logic Programming Data Mining*.

Grammalidis, N., Bleris, L., & Strintzis, M. G. (2002). *Using the Expectation-Maximization Algorithm for Depth Estimation and Segmentation of Multi-view Images*.

Guinepain, S., & Gruenwald, L. (2006). *Automatic Database Clustering Using Data Mining*.

Guo, X., & Nandy, A. (2002). *Numerical characterization of DNA sequences in a 2-D graphical representation scheme of low degeneracy.*

Hebert, P. D. N., Cywinska, A., Ball, S. L., & deWaard, J. R. (2003). *Biological identifications through DNA barcodes*.

Hu, X. O., & Pan, Y. (Eds.). (2007). *Knowledge Discovery in Bioinformatics Techniques, Methods, and Applications.* Hoboken: Wiley.

Huang, G., Liao, B., Li, Y., & Yu, Y. (2009). *Similarity studies of DNA sequences based on a new 2D graphical representation.*

Irene, M. M. (1999). *Hierarchical Clustering*. Retrieved September 29, 2009, from http://www.cse.iitb.ac.in/dbms/Data/Courses/CS632/1999/clustering/node3.html

Jain, A. K., & Dubes, R. C. (1988). *Algorithms for Clustering Data*. Upper Saddle River: Prentice-Hall.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data Clustering: A Review. *ACM Computing Surveys, 31*(3).

Jenssen, R., Hild, K. E., Erdogmus, D., Principe, J. C., & Eltoft, T. (n.d.). *Clustering using Renyi's Entropy*.

Kauer, G., & Blocker, H. (2003). Applying signal theory to the analysis of biomolecules. *Bioinformatics, 19*(16), 2016-2021.

Kozobay-Avrahama, L., Hosid, S., Volkovich, Z., & Bolshoy, A. (2008). *Prokaryote clustering based on DNA curvature distributions.*

Liu, L., Ho, Y., & Yau, S. (2006). *Clustering DNA sequences by feature vectors*.

Lv, T., Huang, S., Zhang, X., & Wang, Z. (2006). *A Robust Hierarchical Clustering Algorithm and its Application in 3D Model Retrieval*.

Myller, N., Suhonen, J., & Sutinen, E. (2002). *Using Data Mining for Improving Web-Based Course Design*.

Ng, H. P., Ong, S. H., Foong, K. W. C., Goh, P. S., & Nowinski, W. L. (2006). *Medical Image Segmentation Using K-Means Clustering and Improved Watershed Algorithm*.

Paccanaro, A., Casbon, J. A., & Saqi, M. A. S. (2006). *Spectral clustering of protein sequences*.

Palace, B. (1996). *Data Mining*. Retrieved September 29, 2009, from http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/index.htm

Qi, Z., & Qi, X., (2009). *Numerical characterization of DNA sequences based on digital signal method*.

Randi, M., Vracko, M., Ler, N., & Plavsi, D. (2002). *Novel 2-D graphical representation of DNA sequences and their numerical characterization.*

Schenker, A. (2003). *Graph-Theoretic Techniques for Web Content Mining*.

Silverman, B. D., & Linsker, R. (1986). *A measure of DNA periodicity*.

Silverman, J. F., & Cooper, D. B. (1988). *Bayesian Clustering for Unsupervised Estimation of Surface and Texture Models.*

Song, J., & Tang, H. (2005). *A new 2-D graphical representation of DNA sequences and their numerical characterization.*

Stoeckle, M. (2003). Taxonomy, DNA, and the Bar Code of Life. *BioScience, 3*(9), 796-797.

Tan, P., Steinbach, M., & Kumar, V. (2006). *Introduction to Data Mining.* Boston: Pearson Education.

Valgren, C., Duckett, T., & Lilienthal, A. (2007). Incremental Spectral Clustering and Its Application To Topological Mapping. *IEEE International Conference on Robotics and Automation*.

Vinod, V. V., Chaudhury, S., Mukherjee, J., & Ghose, S. (1994). *A Connectionist Approach for Clustering with Applications in Image Analysis.*

Visnick, L. (2003). *Clustering Techniques*.

Voss, R. (1992). Evolution of long-range fractal correlation and 1/f noise in DNA base sequences. *Physical Review Letters*, 68, 3805-3808.

Wang, W., & Johnson, D. H. (2002). Computing linear transforms of symbolic signals Signal Processing. *IEEE Trans. Sig. Proc., 50*(3), 628-634.

Weiming, H. X. L., & Zhang, Z. (2007). *Corner Detection of Contour Images Using Spectral Clustering*.

XL Miner (n.d.). *Hierarchical Clustering*. Retrieved September 29, 2009, from http://www.resample.com/xlminer/help/HClst/HClst_intro.htm

Zhang, H., Ho, T., & Linz, M. (2004). *An Evolutionary K-Means Algorithm for Clustering Time Series Data*.

Zhang, Q., Peng, Q., & Xu, T. (2008). *DNA splice site sequences clustering method for conservativeness analysis.*

Zien, A. , Ratsch, G., Mika, S., Scholkopf, B., Lemmen, C., Smola, A., Lengauer, T., & Muller, K. R. (n.d.). *Engineering Support Vector Machine Kernels That Recognize Translation Initiation Sites.*