Feb, 2014. Vol. 4, No. 9 ISSN2305-8269
International Journal of Engineering and Applied Sciences

www.eaas-journal.org

GENETIC ANALYSIS: FROM CLASSICAL APPROACHES TO NEXT-GENERATION APPROACHES

Nour E. Oweis[1], Mohammad A. Alrababah[2], Khaled Sadeq Al-Shredei[3], Mohammad Al-ansari[4], Shady G. Oweis[5]

VSB - Faculty of Electrical Engineering and Computer Science[1]
Northern Border University[2]
King AbdulAziz University[3]
Northern Border University[4]
National Mining University of Ukraine[5]
oweis.nour@gmail.com[1], hamzehamerah@yahoo.com[2], ansare55@hotmail.com[4]
shadi.oweis@yahoo.com[5]

Abstract

Recently bio-informatics has become a central focus area in research in the life sciences due to the accumulation of biological data as a result of deferent organism projects such as the Human-Genome project and other studies [1][2]. The interaction between molecular biology and informatics is rejected in the word bio-informatics [3]. Bio-informatics is a new, developing field, and researchers, especially from computer science, have a good opportunity to contribute by engineering solutions to the open problems existing in this field.

Keywords: Bioinformatics, Human genome, Molecular biology

INTRODUCTION

Most of the datasets in this field can be categorized as sequence-based datasets. Consequently, most of the problems related to these datasets can be categorized as sequence-based problems. These problems are common in other fields, such as speech recognition and natural language processing.

Thus, the main entity in these biological databases is the "sequence". There are three types of bio-sequences, these types are: DNA, RNA and protein sequences. Each type is drawn from a different finite alphabet. The cardinality of the alphabet is either 4 for DNA and RNA sequences or 20 for protein sequences. Bio-sequences carry useful information about organisms (e.g. human) or microorganisms (e.g. bacteria or viruses). Extracting and analyzing the information from bio-sequences is not an easy task, and different types of methods or approaches need to be employed to achieve this goal [4] [5] [6] [7] [8].

The data revolution in biological field encourage researchers from different research fields to produce sequence analysis tools with the capability of extract hidden information from sequence-based datasets, and have the capacity of analyzing and inferring decisions based on the extracted hidden information. The validity of the inferred decisions is mainly based on the validity of extracted information, which consequently has an impact on decisions of specialists in medical sector. Now the question is: Are specialists in the medical sectors seeking for tools that offer: (i) automated decisions generated by sequence analysis tools, or (ii) semi-automated decisions augmented with human decisions (i.e. specialist decisions). The goal of the next generation sequence analysis tools is to infer plausible decisions under the constraint: the extracted informations from sequence-based datasets are probably approximately correct. The term probably approximately correct is a vague term and its definition must be given in details. The term probably approximately correct has three levels, and those levels are: (i) correct information (CI), (ii) approximately correct information (ACI), and (iii) probably approximately correct information (PACI). The highest level is CI and the lowest level is PACI. To test the correctness of the extracted information from sequence-based datasets is a complicated task due to uncertainty of the data nature under consideration, and the following example explains the concept. Any bio-sequence has different structures. The primitive structure of any bio-sequence is its primarily structure. The primary structure is defined as a linear structure of bases (or symbols). Sequences with the same biological function may have different primary structures. The submitted datasets are sequence-based datasets, and the submitted sequences are either labeled sequences or unlabeled sequences. The process of labeling sequences is a complicated task, for example, in cancer problems, the process of classify (i.e. assign labels to sequences) sequences is a vague process in the sense of defining the cancer stages or levels, for example, a sequence may belongs to one of the following classes: (i) anomaly sequence (or abnormal sequence), (ii) approximately anomaly sequence, or (iii) probably approximately anomaly sequence.

Without loss of generality, assume that we have an alphabet $\Sigma=\{\sigma_1,\sigma_2, ..., \sigma_n\}$, and assume that the sequences $\sigma_1\sigma_1\sigma_2\sigma_1\sigma_6\sigma_{10}$and $\sigma_1\sigma_1\sigma_1\sigma_6\sigma_{10}$aredrawn from Σ. The number of bases (i.e. symbols) that constitute the primary structure of the first sequence is greater than the number of bases that constitute the primary structure of the second sequence. Assume that both sequences have the same biological function. Now the question is: Is the amount of information contained in the primarily structure of first sequence is the same as the amount of information contained in the primarily structure of the second sequence, which may lead to the same biological function. Suppose that we extract more information from those sequences using 1-gram model. The 1-gram model of the first sequence is: 3 σ_1, 1 σ_2, 1 σ_6, 1 σ_{10}, while the 1-gram model of the second sequence is: 3 σ_1, 1 σ_6, 1 σ_{10}. It is clear that the 1-gram model of the first sequence is different from the 1-gram model of the second sequence, but both sequences have the same biological function. Therefore, as a conclusion, is the extracted information can be classified as: CI, or ACI, or PACI?

For the above example, it is clear that the 1-gram model (i.e. stochastic language model) is either approximately correct or probably approximately correct, because both sequences have the same biological function, but the resultant 1-gram models are different. Moreover, the 2-gram models of the above sequences are different. Hence, the n-gram model is either approximately correct stochastic language model or probably approximately correct stochastic language model. Consequently, it is necessary to propose new feature extraction techniques that have the capability to produce new representations for bio-sequences in order to increase the validity of the extracted information, and consequently to increase the validity of the decision making process.

To provide the medical sectors with next-generation algorithms (or techniques) that have the capability to deal with serious problems such as: Detecting and Analyzing Cancer, and advanced bio-sequence-based analysis in evolutionary biology (e.g. analyzing biodiversity) and virology fields, a new paradigm shift in the area of sequence analysis, and specifically in genetic analysis, is required to provide specialists in the medical sector with cutting-edge algorithms for analyzing bio-sequences regardless of the problem domain. The implementation of the above research vision requires efforts by researchers from different research fields to achieve this valuable goal. In the next section, we present the research objectives.

The Research Objectives

To achieve well defined research goals of any research vision, first we have to establish a research platform. The research platform is a research environment provided with latest research tools and technology. The platform should link different scientific research fields with biology. The main scientific fields are: Computer Science, Mathematics, Statistics, Engineering, and Medical Sciences (see Figure 2). The modern societies face a number of serious health problems, hence building a Multi-disciplinary research platform is a necessity. Biological systems of organisms or microorganisms are complicated systems, hence, analyzing such systems is the key point in building knowledge about those systems, and consequently the extracted knowledge can be used to solve the current ill-defined health problems. Ill-defined health problems are complicated problems, for example, cancer, drug-resistance, evolutionary of viruses, and waves of virus infection (i.e. spread of viruses). Analyzing complicated health problems using bio-sequence based datasets requires new generations of sequence analysis algorithms with capability of extracting and analyzing most of the hidden information in sequence-based datasets compared with conventional sequence analysis algorithms, where the term hidden information is defined as the amount of information that has a significant impact on the process of analyzing sequence-based datasets. In other words, to move from the level probably approximately correct information to the level approximately correct, or from the level approximately correct information to the level the level correct information, we have to proceed in improving the existing sequence analysis algorithms, and proposing new algorithms.

The process of analyzing sequence-based datasets can be divided into two stages. The first stage is to extract information from sequence-based datasets, and the second stage is to analyze the extracted information. Hence, there are two types of algorithms must be proposed and applied in this context: (i) feature extraction algorithms, and (ii) information analysis algorithms. Therefore, to develop new sequence analysis algorithms, we have to develop feature extraction algorithms and we have to develop information analysis algorithms. Precisely, the development of new sequence analysis algorithms can be achieved through two research directions (RD): (RD_1) Feature Extraction Approach and (RD_2) Information Analysis Approach.

As we mentioned, the datasets under consideration are sequence based datasets, and the main entity is the bio-sequence. Bio-sequences are expected to have different structures. The main structure is the primary structure, and there are other structures, for example, the secondary structure. Now, the critical questions are: Is the primary structure fully dominant all other higher structures?, Is the primary structure partially dominant all other higher structures? (see Figure 1). Now, to analyze any biological phenomenon, first, we have to gather information. The extracted information must be highly related to the core of the biological phenomenon under study. For example, extract highly related information from sequence-based datasets to analyze cancer problems is subject to data representations in feature space. Now, the questions is: Can we measure the degree of association between the biological phenomenon (i.e. biological side) and the extracted information (i.e. information theory side)?

Therefore, the way of representing the extracted information in feature space needs to be improved to increase the degree of association the degree of association between the biological phenomenon and the extracted information (Research

Approach RD_{21}).

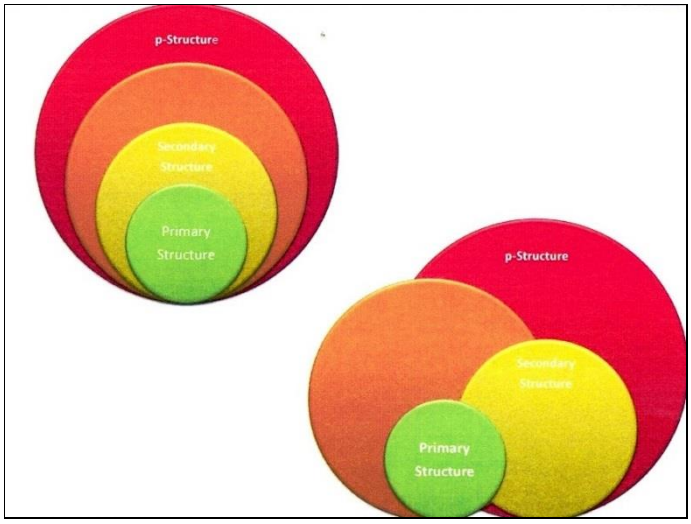

Figure 1 structure of bio – sequences

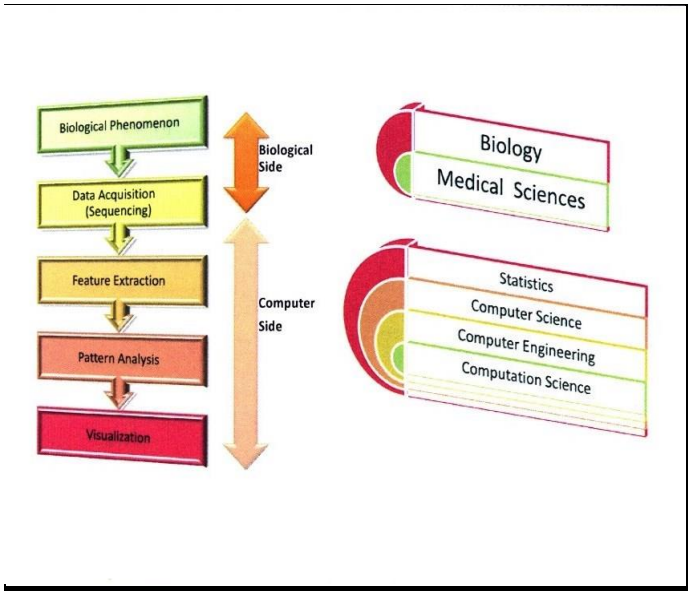

Figure 2 analysis of bio – sequences

General Algorithm

The general algorithm for local computation can be seen as having two phases. During the first phase, a suitable computational structure is established. In the second phase the computations themselves are executed.

The first phase is sometimes referred to as *Compilation*, the latter as *propagation*. In the genetics Literature, it is more usual to describe the peeling algorithm with these two phases executed simultaneously

SOME SPECIFIC APPLICATIONS

In this section we will show how some specific problems in pedigree analysis can be formulated using the Bayesian network representations of Section 2.4.

In the first of these, we will consider the problem of detecting a rare recessive disease by linkage to a marker locus where both loci are assumed to be discrete. A graphical modeling approach to this particular problem has been taken by several authors (Kong, 1991; Jensen and Kong, 1999; Thomas, Gutin, Abkevich and Bansal, 2000). Our aim here is to emphasize that the *entire* model can be specified by the graph without the need for complicated equations and derivation of the relevant joint and full conditional distributions.

The second example is taken from Sheehan et al. (2002) and shows how to incorporate a continuous quantitative trait into this setting In addition, software tool boxes are required. The required software tool boxes developed by **MATHWORKS** (academic version) are listed below:

1. Neural Network toolbox.
2. Statistics toolbox.
3. Bio-informatics toolbox.
4. Pattern Recognition toolbox.
5. Signal Processing toolbox.
6. Image Processing and Analysis toolbox.
7. Fuzzy-Logic toolbox.
8. Genetic Algorithms toolbox.
9. Data Acquisition toolbox.
10. Graph Theory toolbox.
11. Discrete Event Simulation toolbox.
12. Symbolic Math Toolbox.
13. Parallel Processing toolbox.

The process of studying any biological phenomenon can be divided into two sides: (i) Biological Side and (ii) Computer Side. Biologically speaking, biological processes are complicated processes and extracting sequence-based data requires advanced technology. The biological labs expected to be equipped with latest tools, and the extracted data submitted to databases through defined procedures. Offering unlimited access to digital libraries is an essential requirement for the research team. Moreover, attending workshops and top-ranked international professional conferences is mandatory, implicitly, this step allows the research team to communicate and establish scientific research networking with the international research community. To maintain the power of computing, high performance computers are required to be the core of the first high performance computing lab in the Middle East.

The Proposed Research Topics

In this section, we present potential research topics. The first research track is to solve the problem of extracting real-valued features from bio-sequences using recurrent neural networks. The second research track is to analyze the cancer problem using occurrences and recurrences of words in bio-sequences. The third research track is to analyze genomes of pathogens (e.g. viruses) using dispersion maps.

Extracting Features from Bio-sequences Using Neural Networks

The core feature of any bio-sequence is its sequential nature. The bases that constitute the primary structure of any bio-sequence are sequentially related. The main feature extraction model used to extract real-valued features from a sequence is the n-grams model. The n-grams model is a well-known statistical language model, which is frequently used to extract features from bi-sequence. The main concern in implementing the n-grams model is the fragmentation process performed on bio-sequences. Fragmentation process demolishes the sequential nature of bio-sequences. The question is: Can we map a given bio-sequence into the real-valued feature space without demolish its sequential nature?

One of the frequently implemented algorithms used for sequential data is the recurrent neural network. A recurrent neural network is a stochastic approximation algorithm used to classify unlabeled data items with sequential nature after training the network with positive exemplars (i.e. labeled data items).In this research, we aim to find another implementation for recurrent neural network, which is by convert it to a feature extraction machine with the capability of extract features from a sequential data item without demolish its sequential nature. The feature representation of a sequential data item is expected to be embedded in a real-valued high dimensional feature space $\Re^p \times p$. The new representation of sequential data can be used in performing well know pattern analysis tasks in feature space instead of data space, for example, (i) visualization, (ii) supervised classification, and (iii) unsupervised classification.

Analyzing Cancer Using Occurrences and Recurrences of Words in Bio-sequences

One of the complicated problems exists in medical fields is the cancer problem. Classifying a cell as either abnormal cell or normal cell is implicitly defined as a binary classification task of a data item. A data item is defined as a bit of information extracted from a cell. A number of classification algorithms exists in the literature, which can be implemented, but the main concern is the degree of success achieved by the existing algorithms to detect cancer (i.e. to detect abnormal cells). Different classification algorithms exist in the literature, and those algorithms are mainly classified as: (i) machine learning algorithms, (ii) statistical pattern recognition algorithms, and (iii) signal processing algorithms. In this research, we aim to study the impact of occurrences and recurrences of sets of words in bio-sequences extracted from cells on the task of detecting cancer. In addition, in this research, we aim to implement the concept of PAC learning theory (i.e. Probably Approximately Correct Learning), and to test the applicability of the PAC concept on the problem under consideration.

Genetic Analysis of Viruses Using Dispersion Maps

Analyzing coding regions of different types of viruses can be achieved using different approaches. The approaches can classified as either stochastic approaches or deterministic approaches. In this research, we focus on proposing a new approach to analyze genomes of different viruses using the dispersion maps. The dispersion map is defined as a matrix in $\Re^p \times p$, where each element belongs to the matrix is stochastically (approximately) estimated. The dispersion map is non-classical visualization tool that can be used to visualize genomes of viruses in high dimensional spaces, and to visualize different hidden structures of genomes that are undetectable in low dimensional spaces. The proposed visualization tool is expect to be a new shifted visualization tool compared with the existing classical visualization tools that are currently used by specialists in medical sector to visualize extracted features (information) from genomes of viruses.

- Project Statement: We define the project statement as follows:

Definition 1: We aim to propose new algorithms/techniques to analyze Bio-sequence based datasets, and genetically speaking, we aim to tackle serious medical problems, for example, cancer problem, and to analyze the impact of biogenetics-diversity of pathogens on health issues.

From the research statement, we can infer the main research topics that should be considered by the research team, and consequently, we can define the scope of scientific problems under study. Moreover, we can setup the main goals of project under consideration.

In section 2, we presented the main research objectives to be accomplished by the research team. As we mentioned, the datasets under consideration are bio-sequence-based datasets. The datasets are collected from organisms (or microorganisms) to study a biological phenomenon. Hence, the main objective is to propose advanced analysis algorithms/techniques for analyzing bio-sequence-based datasets. Therefore, the expected research topics are: (i) Feature Extraction Approach(s) and Information Analysis Approach(s).

- Literature Review: The literature review is an essential step to explore the exiting bio-sequence analysis algorithms, and summarize their strengths and weaknesses.

- Data Collection: The data collection is mandatory step in building the basic knowledge about the scientific problem under consideration .The design of experiments is mainly based on the data availability. There are various ways to handle

the lack of data, for example, simulated data and augmented data (i.e. generate simulated data using real data).

- Methodologies and Approaches: First, we have to define the scope of the methodologies and approaches that can be used to solve the scientific problem under consideration. In this context, we can define two road-maps. The principle of the first road-map is: the solution of the scientific problems under consideration can be achieved by implementing and adapting the existing methods and algorithms, which can be called application-based road-map, and the principle of the second road-map is: the solution of the scientific problems under consideration cannot be achieved using the exiting methods and algorithms, and hence, new generations of methods and algorithms should be proposed and implemented, which can be called theory-application based road-map. In this project, we will focus on the second road-map to tackle the current serious scientific problems under study.

- Coding and Implementation: Coding the proposed ideas can be accomplished using various software tool boxes and coding languages. The coding stage is an essential stage to implement, evaluate, and upgrade the proposed ideas.

- Research Work Writing- Type-I Paper(s): The critical step is to select the way of presenting the proposed research work to others. The research writing is the necessary step to advertise the proposed research work, and to receive feedback from local and global research communities. Type-I research writing is defined as the process of presenting the achieved research work that focus on the implementation of the exiting methods and algorithms without focusing on theory, and conducting scientific comparison with other existing methods, algorithms and models. The type-I research writing must focus on criticizing the implementation of the proposed research work instead of criticizing theory.

- Research Work Writing- Type-II Paper(s): Type-II research writing is defined as the process of presenting the achieved research work that focus on the implementation and theorization (i.e. theoretical justification) of the proposed new method(s) and algorithm(s), and conducting scientific comparison(s) with other existing method(s), algorithm(s) and model(s) (if the comparison is applicable). Type-II research writing must focus on criticizing the implementation and theorization of the proposed research work.

- Presenting the final version of the proposed project: Series of keynote pre- sentation(s) should be presented to show the main achievements and results of the proposed project.

The sketch of the executive summary is given in Figures 3 and 4.

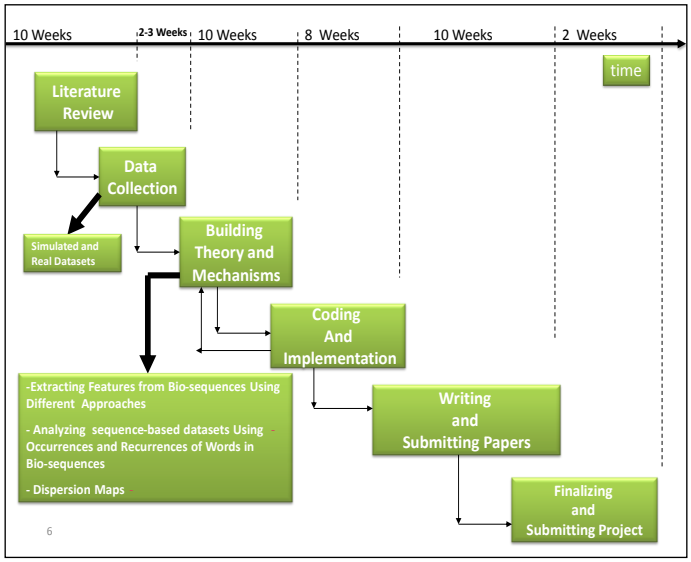

Figure 3 : Executive Summary - Time Table

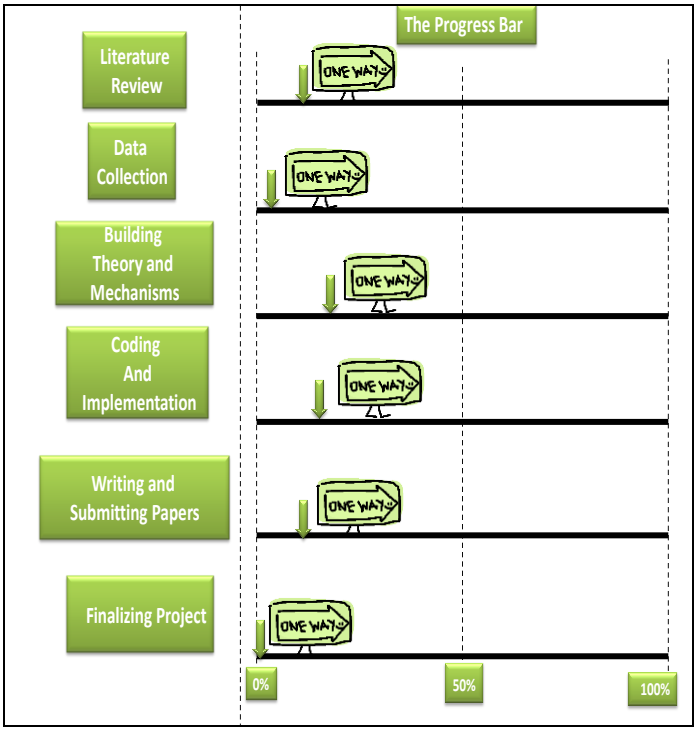

Figure 4 : Executive Summary - Progress Bar

Summary of the Article

The executive summary of this article is divided into two parts: (i) time table and (ii) progress bar. We notice that to perform any ambitious project, we have to setup the main stages of project under consideration. The first stage is to setup project statement, and the project statement is simply defined as: the scope of the project and questions related to scientific problem

under consideration. Precisely, and also we answers the questions are hypothetical and/or practical scientific questions. The second stage in this project is to setup number of goals based on project statement, and then we sketch methodologies and approaches that it should be used to achieve those goals, implicitly, we aim to criticize information content in the project statement, and find scientific answers for the specified scientific questions.

Results

From the research statement, we infer the main research topics that should be considered by the research team, and consequently, we define the scope of scientific problems under study. Moreover, we setup the main goals of project under consideration.

In section 2, we presented the main research objectives to be accomplished by the research team. As we mentioned, the datasets under consideration are bio-sequence-based datasets. The datasets are collected from organisms (or microorganisms) to study a biological phenomenon. Hence, the main objective is to propose advanced analysis algorithms/techniques for analyzing bio-sequence based datasets.

Bibliography

1. [1] E. Birney, “Hidden Markov models in biological sequence analysis,” IBM Journal of Research and Development, vol. 45, no. 3/4, pp. 449–454, 2001.
2. [2] J. Wang, Q. Ma, D. Shasha, and C.H.Wu, “New techniques for extracting features from protein sequences,” IBM Systems Journal, vol. 40, no. 2, pp. 426–441, 2001.
3. [3] R. Backofen and D. Gilbert, “Bioinformatics and constraints,” Constraints, vol. 6, no. 2/3, pp. 141–156, 2001.
4. [4] M. Daoud and S. Kremer, “Neural and statistical classification to families of bio- sequences,” International Joint Conference on Neural Networks, 2006 (IJCNN ’06), pp. 699–704, 2006.
5. [5] M. Daoud and S. C. Kremer, “Detecting similarities between families of bio-sequences using the steady-state of a pca-neural network,” In IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB06), 2006, pp. 179–185.
6. [6] M. Daoud and S. Kremer, “A new distance distribution paradigm to detect the variability of the influenza-a virus in high dimensional spaces,” in Bioinformatics and Biomedicine Workshop, 2009. BIBMW 2009. IEEE International Conference on, nov. 2009, pp. 32 –37.
7. [7] M. Daoud, “A new variance-covariance structure-based statistical pattern recognition system for solving the sequence-set proximity problem under the homology-free assumption,” Ph.D. dissertation, 2010.
8. [8] M. Daoud and S. C. Kremer, “A new variance-covariance structure-based statistical pattern recognition system for solving the sequence-set proximity problem under the homology-free assumption,” 2012, paper in Preparation.