International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 05 | May-2018 www.irjet.net p-ISSN: 2395-0072

Disease Detection Using Gene Files

Junie Sanam K M[1], Malavika A B[2], Najla Ajmal Puthan Veettil[3], Joby George[4]

[1,2,3] Dept. of Computer Science and Engineering, MACE, Kerala, India
[4] Assistant Professor, Dept. of Computer Science and Engineering, MACE, Kerala, India

---

Abstract - *The large amounts of biological information generated using advanced high-throughput experimental techniques continue to motivate the design of suitable methods for valuable knowledge mining. Finding proper means to examine and analyze such information allows better understanding of normal biological processes as well as uncovering malfunctions that trigger various diseases. Several computational approaches were developed to complement the experimental work which is often restricted by high time and cost requirements. In this paper, we consider the problem of disease- gene association and we propose a methodology based on a classification approach which integrates protein-protein interaction network topology features and biological information collected from various data sources. When applied on a dataset of multiple disease types and using a learning model which classifies genes as disease-related or not, based on both topological and biological features. We also consider three case studies of Type II Mellitus, Breast Cancer and Alzheimer's. The experimental results greatly favor our approach. Given a list of genes, the goal is to maximize the contrast between disease and non-disease classes. We make use of three dynamic programming algorithms-Smith Waterman, Fasta and Blast algorithm to perform local sequence alignment and thereby calculate the similarity score. Given a list of genes, the goal is to maximize the contrast between disease and non-disease classes.*

***Key Words*: Gene Disease, Gene Files, Nucleic Acid Sequence, Sequence Alignment.**

1. INTRODUCTION

A gene is a linear sequence which is discrete consisting of nucleotide bases (molecular units) of the DNA that codes for, or directs, the synthesis of a protein; there are an estimated 20,000 to 25,000 genes in the human genome. Proteins, of which are mostly enzymes, carry out all cellular functions [1]. Alteration of the DNA may result in the defective synthesis and subsequent malfunctioning of one or more proteins. If the formed mutated protein is a key enzyme in normal metabolism, the error may have serious or fatal consequences. There are more than 5,000 distinct diseases that have been ascribed to mutations that result in deficiencies of critical enzymes. There is a large amount of biological data that are generated nowadays using high-throughput experimental techniques. The analysis of similar information and particularly the identification of the genes and the inter-molecular events leading to the formation of diseases remain essential towards the development of effective medical therapies. Human genetic diseases, any of the diseases and disorders are caused by mutations in one or more genes. With the increasing ability to control infectious and nutritional diseases in several developed countries, there has come the realization that genetic diseases are a major cause of disability, death, and human tragedy. Rare are families that is entirely free of any known genetic disorder [2]. Many hundreds and thousands of different genetic disorders with defined clinical symptoms have been identified. Of the 3 to 6 percent of newborns are born with a recognized birth defect, at least half involve a predominantly genetic contribution. About 30 percent of all postnatal infant mortality in developed countries is due to genetic disease; 30 percent of pediatric and 10 percent of adult hospital admissions can be traced to a predominantly genetic cause. Medical investigators estimate that genetic defect however minor are present in at least 10 percent of all adults. Through our paper, we make use of gene files to find out the probability of occurrence of diseases such as diabetes mellitus, Alzheimer's and breast cancer in an individual [3]. From the computed value, even early disease diagnosis is possible for a person by following very simple procedures. Your symptoms might be reversible. The symptoms you are concerned about might be caused by a condition that is reversible. Even if there is also an underlying dementia such as Alzheimer's disease, diagnosis and treatment of reversible conditions. It can improve brain function and reduce symptoms. It may be treatable. Some causes of cognitive decline are not reversible, but might be treatable. Appropriate treatment can stop or slow the rate of further decline. With treatments, the sooner the better. Treatment of Alzheimer's and other dementia causing diseases is typically most effective when they are started early in the disease process. Once more effective treatments are available, obtaining an early and accurate diagnosis will be even more crucial. It's empowering, an earlier diagnosis enables the person to participate in their own legal, financial, and long-term care planning and to make their wishes known to family members. We make use of the resources available to us. Individuals diagnosed early in the disease process can take advantage of early-stage support groups and learn tips and strategies to better manage and cope with the symptoms of the detected disease. It will help families, an earlier diagnosis and more opportunity to learn about the disease, develop realistic expectations, and plan for their future together which can result in reduced stress and feelings of burden and regret later in the disease process.

2. SYSTEM DESIGN

2.1 Fasta

Fasta helps to locate regions of the database sequence and matching regions in the query sequences that have high densities of exact word matches (without gaps). The length of

the matched word is termed as the ktup parameter. The highest scoring ten regions of the sequence are rescored using a scoring matrix. The score for such a pair of regions is saved as the init1 score. It determines if any of the initial regions from different diagonals may be joined together to form an approximate alignment with gaps. Only non-overlapping regions can be joined. The score for the joined regions is the sum of the scores of the initial regions minus a joining penalty for each of the gap. The score of the highest scoring region, at the end of this step, is saved as the initn score. After computing the initial scores, Fasta helps to determine the best segment of similarity between the query sequence and the search set sequence, using a variation of the Smith-Waterman algorithm. The score for the alignment is the opt score. In order to evaluate the significance of such alignment. Fasta empirically estimates the score distribution from the alignment of many pairs of random sequences. More precisely, the characters of the query sequences are reshuffled (to maintain bias due to length and character composition) and searched against a subset of the database randomly. This empirical distribution is extrapolated, assuming that it is an extreme value distribution, and each alignment to the real query is assigned a Z-score and an E-score.

2.2 Blast

BLAST first searches for short regions of a given length (W) called 'Words" (or substrings) that score at least "T" when compared to the query sequence that align with sequences In the database ("target sequences"), using a substitution matrix. For every pair of sequences (query and target) that have a word or words in common, BLAST extends the alignment in both directions to And alignments that score greater (are more similar) than a certain score threshold (S). These alignments are called high scoring pairs or HSPs; the maximal scoring HSPs are called MSF's. Query words are compared to the database (target sequences) and exact matches identified. For each word match, alignment Is extended in both directions to find alignments that score greater than some threshold (maximal segment pairs, or MSPs).

2.3 Smith Waterman

Determine the substitution matrix and the gap penalty scheme. A substitution matrix assigns each pair of bases or amino acids a score for match or mismatch. Usually matches get positive scores, whereas mismatches get relatively lower scores. A gap penalty function determines the score cost for opening or extending gaps. It is suggested that users choose the appropriate scoring system based on the goals. In addition, it is also a good practice to try different combinations of substitution matrices and gap penalties. Initialize the scoring matrix. The dimensions of the scoring matrix are 1+length of each sequence respectively. All the elements of the first row and the first column are set to 0. The extra first row and first column make it possible to align one sequence to another at any position, and setting them to 0 makes the terminal gap free from penalty. Scoring. Score each element from left to right, top to bottom in the matrix, considering the outcomes of substitutions (diagonal scores) or adding gaps (horizontal and vertical scores). If none of the scores are positive, this element gets a 0. Otherwise the highest score is used and the source of that score is recorded. Starting at the element with the highest score, trace back based on the source of each score recursively, until 0 is encountered. The segments that have the highest similarity score based on the given scoring system is generated in this process. To obtain the second best local alignment, apply the trace back process starting at the second highest score outside the trace of the best alignment.

3. RESULTS AND EVALUATIONS

Finding suitable methods for the identification of disease genes remains essential towards understanding how various disorders are formed. It can ultimately allow finding appropriate medical therapies. Our solution integrates topological features calculated based on protein interaction network analysis with various biological information / features of genes stored in multiple databases. Our approach is tested on a large dataset including genes associated with multiple disorders. In addition, two case studies are considered; Type II Diabetes Mellitus and breast cancer. The experimental work greatly supports our initial hypothesis. Combining computationally conveyed network study and experimentally-generated biological information can greatly enhance the gene-disease association process.

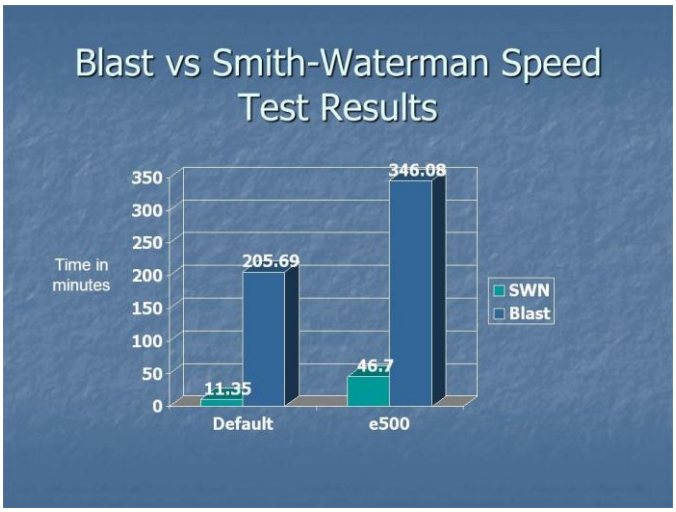

Fig -1: Comparison between Blast and Smith Waterman

4. CONCLUSION

Through this paper, we were able to develop a customer interactive website that will help many individuals in finding the probability of occurrence of disease using their gene files. Gene files of different diseases like diabetes mellitus, breast cancer and Alzheimer's are stored in the database by the admin (doctor). Each patient is given a login id. Our website provides them the privilege to view results and

prescribed medicines. Lab technician has the provision to add gene files of different patients. Doctor is also given a login id through which he can view the result of different patients, his medical history and also he can prescribe medicines to the corresponding patient. With the help of our website the time lag and enormous effort caused by the conventional methods of disease diagnosis can be reduced to a great extent. It helped the patients for early disease diagnosis also which is an advancement over the present Method. The new system is very much user friendly and is very much useful in reducing the time and effort of the patient. Thus we can conclude that the website can help the medical field in building a reliable platform for user interactions.

REFERENCES

[1] Smith, Temple F. Waterman, Michael S. (1981). "Identification of Common Molecular Subsequences" (PDF). Journal of Molecular Biology.

[2] Osamu Gotoh (1982). "An improved algorithm for matching biological sequences". Journal of Molecular Biology..

[3] Stephen F.Altschul Bruce W. Erickson (1986). "Optimal sequence alignment using affine gap costs". Bulletin of Mathematical Biology.