MYCOBACTERIUM TUBERCULOSIS ON GRID COMPUTATIONAL ALGORITHM

Ms. R. Geetha*[1] and Dr. D. Ramyachitra*[2]

[#1] M.Phil Research Scholar, Department of Computer Science. [#2]Assistant Professor, Department of Computer, Bharathiar University, Coimbatore, Tamilnadu.

geethachitra90@gmail.com

jaichitra1@yahoo.co.in

***Abstract*— Grid computing as a new computing generation that uses the resources of numerous split up computers linked by a mesh and is utilized for explaining large computation problems by making use of the underutilized resources or grid distributed resources. The major study in Grid computing agreements with the protein sequences in Bioinformatics. For example, the research mostly focuses on finding the disulfide bonds and space groups that may happen in the proteins. Protein analysis which includes protein classifications, searching, alignment, etc requires a coordination of very large databases, tools and methods. Since thousands and lakhs of proteins are deposited into protein related databases by the researchers, it may result in a broad variety of database search. The protein classification is a processing field of research since there arises a need for classifying the newly discovered proteins in to diverse families or in infection proposition. This paper provides an overview for the analysis of mycobacterium tuberculosis proteins in the Grid natural environment.**

Keywords- Grid computing, Protein Analysis, Mycobacterium Tuberculosis, disulphide bond, NMR/XRD method.

I. INTRODUCTION

A computational grid is a hardware and programs structure that provides a reliable, responsible, continual and economical access to high-end computational capability. Though grid computing has become the buzzword in both industry and informative community, it is not a technology which has been evolved from scrape [1]. Grid computing is a supercomputer architecture that blends computer assets from diverse domains to reach a main goal. In grid computing, the computers on the mesh can work on a task simultaneously, thus the presentation as a supercomputer.

Normally, a grid works on diverse jobs inside a network, but it is furthermore capable of employed on specialized applications. it is conceived to explain troubles that are too big for a supercomputer while maintaining the flexibility to process numerous smaller troubles. Computing grids consign a multiuser infrastructure that accommodates the discontinuous claims of large information processing. grid computing is emerging as a undertaking expertise for three reasons: (i) its capability to make more cost-efficient utilization of a given amount of computing assets (ii) as a way to explain large scale troubles that cannot be solved without a gigantic amount of computing power (iii) it suggests that the assets of numerous computers can be controlled and organized in the direction of a common target [2].

Bioinformatics is an interdisciplinary field that develops and advances upon procedures for saving, retrieving, organizing and investigating biological data [19]. Bioinformatics is a conceptualizing biological science in terms of macromolecules and then engages the application of "informatics" methods [18]. It engages the retrieval and investigation of biochemical and biological data. It values methods and concepts from informatics, statistics, numbers, computer science, chemistry, biochemistry, physics, and linguistics [17].

The aspires of bioinformatics are threefold. First at its simplest, bioinformatics organizes facts and figures in a way that permits investigators to get access to living information and to submit new entries as they are made, e.g. the Protein Data Bank for 3D macromolecular structures. While data-acuration is an essential task, the data stored in these databases is vitally ineffective until analyzed. Thus the reason of bioinformatics expands much farther. The second aim is to develop devices and assets that help in the investigation of data. For demonstration, having sequenced a particular protein, it is of interest to contrast it with previously distinguished sequences. This desires more than just an easy text-based seek and programs such as FASTA and PSI-BLAST should address what comprises a biologically significant match. Development of such assets dictates know-how in computational idea as well as a methodical comprehending of biological science.

The third aim is to use these tools to investigate the facts and figures and understand the outcomes in a biologically meaningful manner. Normally, biological investigations examined individual schemes in minutia, and often contrasted those with a couple of that are associated. In bioinformatics, international investigates of all the accessible data can be undertook with the aim of uncovering widespread values that apply across numerous schemes and highlight innovative features [8].

The rest of the paper is organized as follows: Section 2 describes bioinformatics in grid computing. Section 3 describes the protein analysis. Section 4 describes the mycobacterium tuberculosis. Section 5 describes the mycobacterium tuberculosis protein analysis in grid and conclusion is given in section 6.

II. BIOINFORMATICS ON GRID

Bioinformatics is the blend of biological science and information expertise. This control and respect embraces computational devices and methods that are utilized to organize, investigate and manipulate large groups of biological facts and figures. Bioinformatics is essential for accomplishing so numerous convoluted jobs such as use of genomic information in comprehending human infections, identification of new molecular goals for drug discovery and in unravelling human evolution secrets.

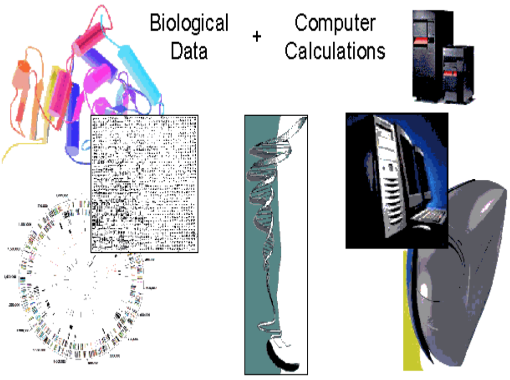

Figure 1: Bioinformatics

Bioinformatics has three significant components namely [8],

- The creation of databases, permitting the storage and management of large biological facts and data sets.
- The growth of algorithms and information to determine relationships among members of large data sets.
- The use of these devices for the analysis and interpretation of various kinds of biological facts and data including DNA, RNA and protein sequences, protein organizations, gene expression profiles, and biochemical pathways.

Grid technologies enable the distributing of bioinformatics data from different sites by conceiving a virtual association of the facts and data. The present grid-enabling programs expertise such as Globus toolkit [4-6], allows the distributing of geographically distributed facts and data. Thus the grid is adept to decrease the single point of malfunction inherited in a centralized database scheme. New research outcomes can be stored on a localized scheme and shared with the research community directly. Users no longer need to understand the position of their target information, but are adept to get access to and retrieve in a clear manner [5]. This paradigm is exceedingly appropriate for large-scale genomic and proteomic undertakings. Study of the evolution of distinct protein sequence and procedure is significant to biologists since it has functional submissions including pharmaceutical breakthrough, space group and disulfide bond development and finding the function of the proteins.

III. PROTEIN ANALYSIS

Proteins are made of 20 or so building blocks called amino acids. Entire proteins comprise the 9 absolutely vital amino acids that the body needs to construct other new proteins [10]. The secondary structure corresponding to a protein is forecast to be a mixture of α-helices and β –strands. Though the function is not renowned, proteins with this function exact to mycobacterial species may be associated with a widespread function [6].

Protein structures are recounted at different levels. Protein Databank (PDB) is a worldwide repository for the processing and circulation of 3D biological macromolecular structure data and it is sustained by the RCSB. If the contents of the PDB are thought of as primary facts and data, then there are hundreds of derived (i.e., secondary) facts and databases that categorize the facts and data differently [12]. Several parameters are also derived for an entire protein or protein chains, protein steadiness, folding and unfolding rates [9].

The secondary structure elements such as helix, sheet, turn and coil constitute the construction blocks of the folding proteins. Regularly doing again localized organizations are stabilized by hydrogen bonds. The most widespread examples are the α-helix, β-sheet and turns. Because secondary organizations are localized, numerous districts of distinct secondary structure can be present in the identical protein molecule [16]. Structural classification designs, as implemented for demonstration in the SCOP, CATH, and FSSP databases, elucidate the relationship between protein bends and function [17].

IV. MYCOBACTERIUM TUBERCULOSIS

Tuberculosis (TB) is a possibly mortal contagious disease that can affect nearly any part of the body but it mainly affects the lungs. It is initiated by a bacterial microorganism, the Mycobacterium tuberculosis. Although TB can be treated, healed and can be stopped, researchers have not ever come close to swabbing it out. couple of diseases have initiated so much causing anguish illness for centuries and asserted so numerous lives [3].Around 40% of persons who have hardworking TB disease have the infection in another part of their body (e.g., lymph glands, brain, spine, kidneys, or other organs). This happens when the bacteria disperse out-of-doors of the lungs. In these cases, TB is more complicated to identify since the persevering does not have the usual signals and symptoms associated with pulmonary TB.

A.The Disease

Tuberculosis (TB) is a disease initiated by the pollution from the pathogens M. tuberculosis. If not treated properly, TB can be fatal. Actually, the World wellbeing association approximates that over 13 million persons have TB and about 1.5 million die every year from the infection. Tuberculosis most usually affects the lungs (pulmonary TB). Patients with hardworking pulmonary TB generally have a hack, an abnormal barrel x-ray and it is infectious. TB can furthermore happen outside of the lungs (extra pulmonary), most routinely in the centered nervous, lymphatic, or genitourinary systems or in the skeletal parts and joints. Tuberculosis which happens scattered all through the body is referred to as miliary TB. Extra pulmonary TB is more common in immune stifled individuals and in young children.

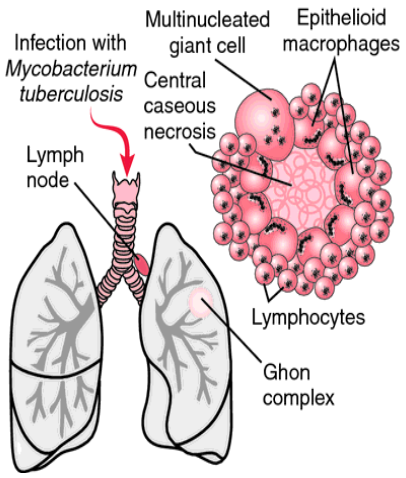

Fig.2 Mycobacterium Tuberculosis.

B. The Symptoms

Pulmonary tuberculosis affects the lungs. Its initial symptoms are easily bewildered with those of other diseases. An infected individual may at first feel vaguely unwell. Symptoms of hardworking pulmonary TB can encompass heaviness decrease, high temperature, evening worries, and decrease of appetite [11]. The disease can either proceed into remission or become graver with the onset of barrel pain and hacking up bloody sputum [3]. The accurate symptoms of extra pulmonary TB alter according to the location of disease in the body.

IV. ANALYSIS ON MYCOBACTERIUM TUBERCULOSIS PROTEIN

Tuberculosis is a contagious bacterial infection initiated by Mycobacterium tuberculosis, which most routinely affects the lungs. It is transmitted from individual to individual via droplets from the throat and lungs. Tuberculosis is treatable with a six-month course of antibiotics [20].

In this paper, the mycobacterium tuberculosis proteins are analyzed for finding their methods, disulfide bond and its space group. The protein chain and the FASTA sequences of these proteins are furthermore analyzed. The growing demand for automated investigation of large and circulated protein facts and figures impersonates new challenges to the available computational power. Since the most of bioinformatics submissions exhibit a high degree of parallelism, they can benefit from the increased accessibility, reliability and effectiveness of computing resources in a Grid natural environment. Therefore the protein analysis techniques can be applied in a grid natural environment which in turn decreases the execution time.

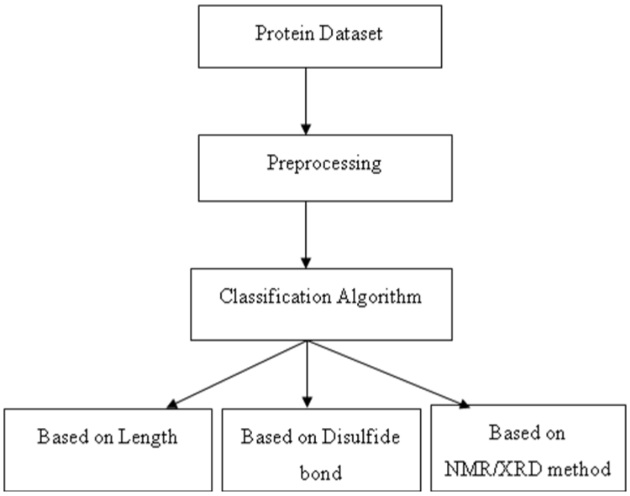

Fig: 3 Flow chart for Classification of protein based on various parameters

The above flowchart shows the classification of protein datasets based on the length, disulphide bond and NMR/XRD methods.

A. Classification

Classification engages decision directions that partition the facts and data into disjoint assemblies. The classification module in the mining system is usually called classifier. There are two types of classifiers, the parametric classifier and non-parametric classifier.

There are some classification discovery models. Some of them are the decision tree, neural systems, genetic algorithms and some statistical models. The procedures for searching a database of known sequences for one most similar to the protein sequence and assign the classification with the best-scoring known sequence. The likeness seek is done by performing a pair-wise sequence alignment between the protein sequences. The search is also presented against a database of renowned sequences, but instead of matching the protein sequence against the multiple sequences directly, these procedures first align multiple sequences from the same protein.

B. Disulfide bond

Disulfide bond is a single covalent bond derived from the coupling of thiol (SH) groups. The linkage is furthermore called as an "S-S-bond" or "disulfide bridge". Disulfide bonds play a significant role in the folding and steadiness of some proteins generally secreted to the additional cellular medium.

Disulfide bonds are formed by the oxidation of thiol (-SH) assemblies in cysteine residues. Disulfide bonds happen intramolecularly (i.e. inside a single polypeptide chain) and intermolecular (i.e. between two polypeptide chains). Intermolecular disulfide bonds stabilize the tertiary organizations of proteins while those that occur intermolecular are involved in stabilizing quaternary structure. Not all proteins comprise disulfide bonds [22].

The fig:4 shows the structure of disulphide bond that exists in proteins.

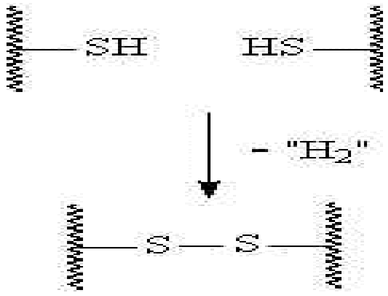

Fig: 4 Structure of Disulfide Bond

C. NMR/XRD method

Atomic magnetic resonance, NMR, and X-ray crystallography are the two procedures that can be applied to the study of three-dimensional molecular organizations of proteins at atomic resolution. NMR spectroscopy is the only procedure that allows the conclusion of three-dimensional structures of proteins substances in the solution stage. A major advantage of NMR spectroscopy procedure is that it presents data on proteins in answer as are against to those locked in a crystal or compelled to a microscope grid and thus NMR spectroscopy is the premier procedure for revising the atomic structures of flexible proteins.

VI. CONCLUSION

Grid computing is a good solution for the challenges faced in bioinformatics field. The analysis of protein sequence is a kind of computation driven science which rapidly increases the size of biological data. Grid environment can help to reduce the execution time for the analysis of Mycobacterium tuberculosis protein. Further research involves the improvement of the grid infrastructure for the development of the space group and disulfide bond for the Mycobacterium tuberculosis protein. Future developments within the grid will certainly overcome certain drawbacks and present a whole new way in how grids are used in life sciences.

REFERENCES

[1]. I. Foster, C. Kesselman, and S. Tuecke," The Anatomy of the Grid - Enabling Scalable Virtual Organizations", International Journal of Supercomputer Applications, 2001.

[2]. R. Al-Khannak, B. Bitzer, "South Modifying Modern Power Systems Quality by Integrating Grid Computing Technology", 2008.

[3]. http://medical-dictionary.thefreedictionary.com/Tuberculosis.

[4]. Swathi Adindla and Lalitha Guruprasad , "Sequence analysis corresponding to the PPE and PE proteins in Mycobacterium tuberculosis and other genomes", School of Chemistry, University of Hyderabad, Hyderabad 500 046, India.

[5]. Grid computing and bioinformatics development. A case study on the Oryza sativa (rice) genome, Pure Appl. Chem., Vol. 74, No. 6, pp. 891–897, 2002. © 2002 IUPAC.

[6]. B. Fran, F. Geoffrey and H. Anthony, "Grid Computing: Making the Global Infrastructure a Reality", Chichester: Wiley, 2003.

[7]. J. H. Kaufman, T. J. Lehman, and J. Thomas, "Grid computing made simple", The Industrial Physicist, pp 32-33, Aug- Sept 2003.

[8]. R. Gupta, "Bio-Informatics, Bonding genes with IT", in INDIACOM 2010, Jaipur, Feb 2010.

[9]. M.Michael Gromiha, "Protein Bio-Informatics, from sequence to function".

[10]. http://www.doctoroz.com/videos/power-protein

[11]. http://medical-dictionary.thefreedictionary.com/Tuberculosis

[12]. http://symptomchecker.about.com/od/Diagnoses/tuberculosis.htm

[13]. http://en.wikipedia.org/wiki/Protein_Data_Bank

[14].http://en.wikipedia.org/wiki/Nuclear_magnetic_resonance_spectroscopy_of_proteins

[15].http://groups.molbiosci.northwestern.edu/holmgren/Glossary/Definitions/Def-/protein.html

[16]. https://en.wikipedia.org/wiki/Protein

[17].http://www.pasteur.fr/recherche/unites/Binfs/definition/bioinformatics_definition.html

[18]. http://www.ncbi.nlm.nih.gov/pubmed/11552348

[19]. http://en.wikipedia.org/wiki/Bioinformatics

[20]. http://www.who.int/topics/tuberculosis/en/

[21]. www.nhn.ou.edu/~bumm/NanoLab/ppt/X-ray_Diffraction.ppt

[22]. *http://webhost.bridgew.edu/fgorga/proteins/disulfide.htm*