epcALGO: a home-grown algorithm for entire proteome comparison

Abstract

Due to the advancement of bioinformatics and genome sequencing project, entire genome and proteome sequences of different organisms are available in the public domain. These vast data are repeatedly compared and explored to find out identical and similar sequence patterns. In this paper we employed NCBI's Standalone BLAST program for entire proteome comparison of any two strains / species and illustrate a simple algorithm for the same. The implementation of this epcALGO algorithm is to identify systematically conserved proteins that are missing in a given proteome and also identify proteins unique to a particular species. This algorithm is simple and quick to apply for revealing the species / strain variation among any two closely related species / strains by identifying identical and non-identical proteins in their proteomes and also identifying where there is mutation in the protein sequence. We implemented this algorithm for proteome comparison of two strains of *Mycobacterium tuberculosis* $H_{37}Rv$ and $H_{37}Ra$ and elucidated the methodology for finding out their proteomic variation.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.210v1 | CC-BY 4.0 Open Access | received: 19 Jan 2014, published: 19 Jan 2014

Satish Kumar[1*], Lingaraja Jena[1,2]

[1] Bioinformatics Centre & Biochemistry, Mahatma Gandhi Institute of Medical Sciences, Sevagram, Maharashtra, India
[2]Department of Bioinformatics, Shri JJT University, Jhunjhunu, Rajasthan

Email address:
SK: satishangral@gmail.com
LJ: lingaraj.jena@gmail.com

*Corresponding Author:

Dr. Satish Kumar
Professor, Biochemistry &
Dy Coordinator, Bioinformatics Centre
Mahatma Gandhi Institute of Medical Sciences
Sevagram-442 102 (Wardha) Maharashtra, India
Tel +91 7152 284679
Fax +91 7152 284038
Email: satishangral@gmail.com

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.210v1 | CC-BY 4.0 Open Access | received: 19 Jan 2014, published: 19 Jan 2014

Introduction

Due to the rapid development of new sequencing technologies (Li and Homer, 2010) and hasty progress in bioinformatics, the complete genome and proteome sequences of numerous organisms have become available in the public databases. As of September 2012, Genomes Online Database (GOLD) version 4.0 (Pagani *et al.*, 2012), contains information on 3699 complete sequencing projects of different organisms (http://www.genomesonline.org). In order to retrieve and analyze those huge amount of sequence data, development of novel algorithms and techniques are becoming increasingly important.

Comparison is an essential feature of all biological research and in early days, the comparisons were revolved around morphological and physiological level of research due to unavailability of DNA and proteins sequences (Bachhawat, 2006). In recent era of Genomics & Proteomics the comparison paradigm has been shifted to entire genome & proteome level since complete genome & proteome sequences of several organisms have become freely available in the public databases. Comparative genomics / proteomics among different pathogenic organisms can not only reveal evolutionary history among them but also identify the similarity and variation in their DNA / protein sequences, which accounts for their morphological & physiological changes as well. Besides, such analysis is of great value in phylogenetic reconstruction, drug discovery and functional annotation of hypothetical proteins (Bachhawat, 2006).

Many tools have been developed for complete determination of genome sequence of a huge number of bacteria, but still, their proteomes remain relatively poorly defined. In the post genomic era, proteomics is a rapidly growing field of research for studying proteins involved in carcinogenesis as well as novel biomarker discovery for clinical use such as screening, diagnosis, prognosis, detection of recurrent disease etc (Cho, 2007). Since proteins are specifically directly involved in both normal and disease related biochemical processes, a more comprehensive understanding of disease may be achieved by looking directly into the proteins within a disease cell or tissue (Cho, 2007). Proteomics has much promise in novel drug discovery by targeting proteins of pathogenic organisms causing different diseases in host, whereas comparative proteomics is very significant in studying the proteomic variations among different pathogens.

The simultaneous development of rapid sequence comparison algorithms has revolutionized the role of biological sequence comparison in molecular biology. The Basic Local Alignment

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.210v1 | CC-BY 4.0 Open Access | received: 19 Jan 2014, published: 19 Jan 2014

Search Tool (BLAST) (Altschul *et al.*, 1990) is now prevails over as the fastest and most widely-used tool for sequence similarity searches (McGinnis and Madden, 2004). The stand-alone executable BLAST from the NCBI BLAST site (ftp://ftp.ncbi.nlm.nih.gov/blast/executables/release/LATEST/) provide easy ways for a user to perform blast searches via command line or a local web server. However, the standalone version requires each user to install and configure the program and customize the databases for performing specialized research (Deng *et al.*, 2007).

In this study we have developed a simple algorithm for entire proteome comparison of any two organisms (strains / species) using NCBI standalone BLAST and discussed on how to retrieve and analyze those comparison data. We also implemented the algorithm for comparison of virulent (MTB $H_{37}Rv$) and avirulent (MTB $H_{37}Ra$) strains of *Mycobacterium tuberculosis* (MTB). Our observations provide a unique platform for discovery of proteomic variation in different strains / species.

Materials and Methods

The methodology for entire proteome comparison and analysis involves following simple steps using EpcALGO algorithm (Fig. 1).

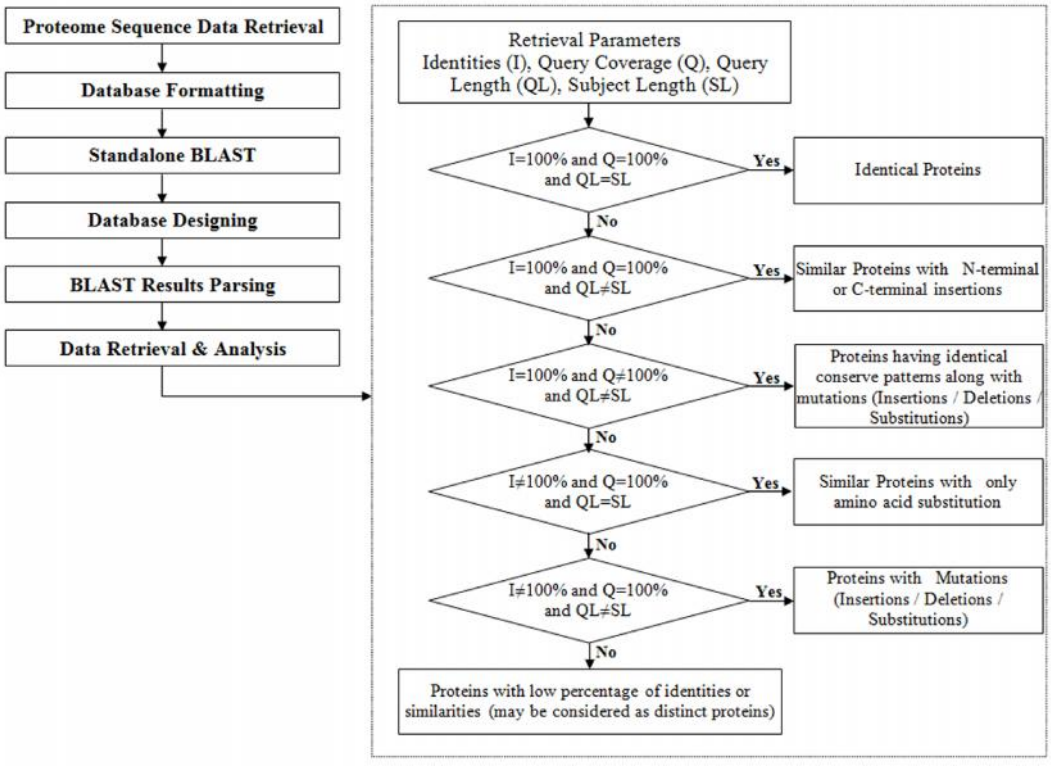

Figure 1: epcALGO Algorithm.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.210v1 | CC-BY 4.0 Open Access | received: 19 Jan 2014, published: 19 Jan 2014

Data Retrieval

Dataset was prepared by retrieving the entire proteome of MTB $H_{37}Rv$ (NCBI RefSeq: NC_000962.3) & MTB $H_{37}Ra$ (NCBI RefSeq: NC_009525.1) from NCBI FTP site (ftp://ftp.ncbi.nih.gov/genomes/).

Database formatting

Protein sequence data of $H_{37}Rv$ (NC_000962.faa) & $H_{37}Ra$ (NC_009525.faa) were formatted using formatdb application of NCBI Standalone BLAST-2.2.26.

Perform Standalone BLAST

Blastall application was used to perform protein BLAST (Altschul *et al.,* 1990) between the proteome of MTB $H_{37}Rv$ against MTB $H_{37}Ra$ to carry out proteomic comparison.

Database Designing

Database table was created using Micosoft SQL (MS SQL) Server for storing proteome comparison information of MTB $H_{37}Rv$ vs MTB $H_{37}Ra$. For database designing other database server such as MySQL can also be used.

Parsing BLAST results

The output of the BLAST result was parsed and stored in MS SQL relational database tables using in-house developed PERL script using Bio::SearchIO module of Bioperl (Stajich *et al.,* 2002). While parsing BLAST output results, percentage identities, positivities, number of gaps, identical residues, bits, bits score, e-value, query length, subject length, query sequence, subject sequence, consensus sequence etc of the first hit obtained were taken into consideration for each protein comparison.

Data retrieval & analysis

Different SQL queries were written to retrieve and analyze comparison data from MS SQL database tables.

Results and Discussion

In spite of several studies in the past the potential causes for variation in virulence between MTB $H_{37}Rv$ and MTB $H_{37}Ra$ have remained unclear. A single amino acid mutation in protein sequence may cause alteration in protein structure and function that may account for

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.210v1 | CC-BY 4.0 Open Access | received: 19 Jan 2014, published: 19 Jan 2014

virulence and drug resistance properties of pathogenic organisms. Therefore, the development of an *in silico* technology to study the proteomic variations of different strains of genetically intractable pathogens such as MTB will enhance the analysis of virulence and drug resistance properties and significantly advance the understanding of the mechanisms of disease.

In our previous study we have implemented this algorithm and developed *Mycobacterium tuberculosis* proteome comparison database (MTB-PCDB) which provides integrated access to protein sequence comparison with identical and non identical protein data for five strains of MTB ($H_{37}Rv$, $H_{37}Ra$, CDC 1551, F11 and KZN 1435) (Jena et al., 2011). In this study, we implemented epcALGO algorithm to performed comparative proteomic analysis of MTB $H_{37}Rv$ and MTB $H_{37}Ra$. While comparison, protein sequence of MTB $H_{37}Rv$ was taken as query and sequences of MTB $H_{37}Ra$ were taken as database sequences (subject).

A total of 4018 protein-coding sequences (CDS) are identified amongst 4111 genes in the MTB $H_{37}Rv$ genome while there are 4084 genes with 4034 protein coding sequences in the genome of MTB $H_{37}Ra$. There were seven categories (Table 1) obtained depending on the percentage identities, query coverage, query and subject length upon entire proteome comparison between these two strains.

Table 1. Comparison of proteomic variations between MTB $H_{37}Rv$ and MTB $H_{37}Ra$

Category	Features	Total Number
1	Identical Proteins in both the strains	3804
2	Proteins having 100% identities and query coverage but with varying sequence length	20
3	Proteins having 100% identities but with variation in query coverage	21
4	Proteins having same sequence length and 100% query coverage but with variation in identities	36
5	Proteins having 100% query coverage but variation in length and identities	31
6	Proteins having variation in length, identities and query coverage	101
7	Proteins with no significant similarities	5

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.210v1 | CC-BY 4.0 Open Access | received: 19 Jan 2014, published: 19 Jan 2014

Category 1: Identical Proteins in both the strains (MTB $H_{37}Rv$ and MTB $H_{37}Ra$)

Proteomic comparison of *M. tuberculosis* $H_{37}Rv$ and *M. tuberculosis* $H_{37}Ra$ in our study revealed 3804 identical proteins between these two strains. A protein of MTB $H_{37}Rv$ is said to be identical compared to the corresponding protein of MTB $H_{37}Ra$ when both identities and query coverage are 100% and the query length is equal to the subject length. So, 214 proteins were identified as non identical proteins between these two strains.

Category 2: Proteins having 100% identities and query coverage but with varying sequence length

There were 20 proteins identified in this category. In this case the length of subject (MTB $H_{37}Ra$) sequences was found to be greater in comparison to sequences of MTB $H_{37}Rv$. This observation revealed that there were insertions in the respective proteins of MTB $H_{37}Ra$.

Category 3: Proteins having 100% identities but with variation in query coverage

21 proteins of MTB $H_{37}Rv$ were found where identities was 100% but with variation in query coverage (less than 100%) when compared with proteins of MTB $H_{37}Ra$. This revealed that there were mutations (insertions / deletions / substitutions) in these sequences.

Category 4: Proteins having same sequence length and 100% query coverage but with variation in identities

Out of 214 non identical proteins, 36 proteins of MTB $H_{37}Ra$ were identified with only amino acid substitution compared to corresponding proteins of MTB $H_{37}Rv$, as there was only variation in identities with same sequence length and 100% query coverage.

Category 5: Proteins having 100% query coverage but variation in length and identities

31 proteins were identified in this category. In this case BLAST results showing 100% query coverage with varying length and identities revealed that there were insertions / deletions / substitutions in these sequences.

Category 6: Proteins having variation in length, identities and query coverage

There were 101 proteins observed with variation in length, identities and query coverage. So, most of the dissimilar proteins were found in this category. As many proteins in this group were found to have lower percentage of similarities and identities, these proteins may be considered as distinct proteins in MTB $H_{37}Ra$, which needs further study.

Category 7: Proteins with no significant similarities

There were five proteins of MTB $H_{37}Rv$ identified with no significant similarities compared to the proteins of MTB $H_{37}Ra$. So, these proteins were unique to MTB $H_{37}Rv$.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.210v1 | CC-BY 4.0 Open Access | received: 19 Jan 2014, published: 19 Jan 2014

Overall comparative analysis provide proteomic differences between MTB $H_{37}Rv$ and $H_{37}Ra$, which may be useful for better understanding of the basis of pathogenesis of *Mycobacterium tuberculosis* and virulence attenuation in MTB $H_{37}Ra$ (Jena *et al.*, 2013). Further studies on functional characterization of non-identical proteins identified in MTB $H_{37}Ra$ and MTB $H_{37}Rv$ may help in understanding the important role of variation among these two strains.

Conclusion

epcALGO algorithm has been successfully implemented in designing MTB-PCDB (Jena *et al.*, 2011). This algorithm is designed especially for protein sequence comparison using 'blastall' application of standalone BLAST, however the same method can also be applicable for DNA sequence comparison using nucleotide blast. As genome and proteome sequences of different organisms are available and easily accessible to researchers, this simple and proficient algorithm developed in this study, would be helpful in finding proteomic variations in different strains / species and subsequently reveals the morphological dissimilarities amongst them.

Acknowledgments

Authors express gratitude to Dr. B.C. Harinath, Director, JBTDRC & Coordinator, Bioinformatics Centre for providing Bioinformatics Laboratory wherein this study has been carried out. Grateful thanks to Shri D.S. Mehta, President, Kasturba Health Society, Dr. (Mrs.) P. Narang, Secretary, Kasturba Health Society, Dr. B.S. Garg, Dean, MGIMS & Dr. S.P. Kalantri, Medical Superintendent, Kasturba Hospital, MGIMS, Sevagram for their encouragement & unconditional support.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.210v1 | CC-BY 4.0 Open Access | received: 19 Jan 2014, published: 19 Jan 2014

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403-410.

Bachhawat AK. 2006. Comparative genomics. *Resonance*, 11(8):22-40.

Cho WC. 2007. Contribution of oncoproteomics to cancer biomarker discovery. *Molecular Cancer*, 6:25.

Cho WC. 2007. Proteomics technologies and challenges. Genomics Proteomics. *Bioinformatics*, 5(2):77-85.

Deng W, Nickle DC, Learn GH, Maust B, Mullins JI. 2007. ViroBLAST: a stand-alone BLAST web server for flexible queries of multiple databases and user's datasets. *Bioinformatics*, 23(17):2334-2336.

Jena L, Kashikar S, Kumar S, Harinath BC. 2013. Comparative proteomic analysis of *Mycobacterium tuberculosis* strain $H_{37}Rv$ versus $H_{37}Ra$. *International Journal of Mycobacteriology*, 2(4):220-226.

Jena L, Wankhade G, Kumar S, Harinath BC. 2011. MTB-PCDB: *Mycobacterium tuberculosis* proteome comparison database. *Bioinformation*, 6(3):131-133.

Li H, Homer N. 2010. A survey of sequence alignment algorithms for next-generation sequencing. *Briefings in Bioinformatics*, 11(5):473-483.

McGinnis S, Madden TL. 2004. BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Research*, 32:W20-W25.

Pagani I, Liolios K, Jansson J, Chen IM, Smirnova T, Nosrat B, Markowitz VM, Kyrpides NC. 2012. The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Research*, 40:D571-579.

Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JG, Korf I, Lapp H, Lehväslaiho H, Matsalla C, Mungall CJ, Osborne BI, Pocock MR, Schattner P, Senger M, Stein LD, Stupka E, Wilkinson MD, Birney E. 2002. The Bioperl toolkit: Perl modules for the life sciences. *Genome Research*, 12(10):1611-1618.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.210v1 | CC-BY 4.0 Open Access | received: 19 Jan 2014, published: 19 Jan 2014