A Database for Mycobacterium Secretome Analysis: 'MycoSec' to Accelerate Global Health Research

Article *in* Omics: a Journal of Integrative Biology · August 2013

DOI: 10.1089/omi.2013.0015 · Source: PubMed

4 authors, including:

Ayan Roy
Lovely Professional University
47 PUBLICATIONS **342** CITATIONS

SEE PROFILE

Asim Kumar Bothra
Raiganj University
79 PUBLICATIONS **297** CITATIONS

SEE PROFILE

OMICS A Journal of Integrative Biology
Volume 17, Number 10, 2013

DOI: 10.1089/omi.2013.0015

A Database for Mycobacterium Secretome Analysis: 'MycoSec' to Accelerate Global Health Research

Ayan Roy,[1] Sanghati Bhattacharya,[1] Asim K Bothra,[2] and Arnab Sen[1]

Abstract

Members of the genus *Mycobacterium* are notorious for their pathogenesis. Investigations from various perspectives have identified the pathogenic strategies employed by these lethal pathogens. Secretomes are believed to play crucial roles in host cell recognition and cross-talks, in cellular attachment, and in triggering other functions related to host pathogen interactions. However, a proper idea of the mycobacterial secretomes and their mechanism of functionality still remains elusive. In the present study, we have developed a comprehensive database of potential mycobacterial secretomes (MycoSec) using pre-existing algorithms for secretome prediction for researchers interested in this particular field. The database provides a platform for retrieval and analysis of identified secretomes in all finished genomes of the family *Mycobacteriaceae*. The database contains valuable information regarding secretory signal peptides (Sec type), lipoprotein signal peptides (Lipo type), and Twin arginine (RR/KR) signal peptides (TAT type), prevalent in mycobacteria. Information pertaining to COG analysis, codon usage, and gene expression of the predicted secretomes has also been incorporated in the database. MycoSec promises to be a useful repertoire providing a plethora of information regarding mycobacterial secretomes and may well be a platform to speed global health research. MycoSec is freely accessible at http://www.bicnbu.in/mycosec.

Introduction

Mycobacterium is one of the oldest known disease-causing microorganisms associated with human and bovine pathogenesis. *Mycobacterium tuberculosis* and *Mycobacterium leprae* are notorious obligate pathogens (Cole et al., 2001; De Voss et al., 2000) that have posed a serious menace to human health from antiquity. Opportunistic pathogens such as *Mycobacterium abscessus* and *Mycobacterium ulcerans* also have a significant impact on mycobacterial pathogenesis (Zumla and Grange, 2002). However, several members of the genus are also nonpathogenic, saprophytic, and eco-friendly strains such as *Mycobacterium vanbaalenii* and *Mycobacterium* sp. strains JLS, KMS, and MCS help in bioremediation process by degrading environmentally toxic polycyclic aromatic hydrocarbons (Miller et al., 2004). Thus, the genus *Mycobacterium,* which includes lethal pathogens such as *M. tuberculosis and M. leprae* and also biofriendly strains like *M. vanbaalenii,* generates a thrill among researchers not only from the pathogenic perspective but also from the eco-friendly angles.

Bacterial pathogenesis and its impact on human health has always been a sensitive field of biomedical research. Bacterial communities exhibit a variety of pathogenic strategies to infect the human host. However, every mode of infection has a common scenario of bacterial adhesion to the host receptor, secretion of toxins, and thus, paving way for successful insertion of the virulence factors (Lee and Schneewind, 2001). Secretory proteins hold the key to interaction with the host and inception of the disease (Bonin-Debs et al., 2004). Secretomes are often found to be linked crucially with virulence and thus promise to be striking drug targets for proper remedy of bacterial infections (Niederweis et al., 2010).

Secretomes have been defined as the complete set of proteins secreted by a cell (Ranganathan and Garg, 2009) and are associated with a broad range of functions and critical biological processes, such as cell-to-cell communication and cross talks, cell migration, and most inevitably virulence and potential infective strategies in disease mechanism (Tjalsma et al., 2004).

The signal peptide part of the secreted protein, which is generally composed of around thirty amino acid residues, transports the newly synthesized protein to the protein-conducting SecE and SecY channels associated with the plasma membrane (Leversen et al., 2009). Signal peptides in most cases are reported to possess three domains: a positively charged n-terminus (n-region), a stretch of hydrophobic

[1]Bioinformatics Facility, Department of Botany, University of North Bengal, Siliguri, India.
[2]Cheminformatics Bioinformatics Laboratory, Department of Chemistry, Raiganj College (University College), Raiganj, India.

residues (H-region), and a region of mostly small uncharged residues containing a characteristic cleavage site recognized by a specific signal peptidase (SPase) (von Heijne, 1984, 1989, 1990a,b). It is this characteristic site that holds the key in cleavage of a secretory protein by either of the two SPases, Type I or Type II.

Various types of signal peptides are reported in bacterial systems among which secretory signal peptides (Sec type), Twin arginine signal peptides (TAT type), lipoprotein signal peptides (Lipo type), pseudopilin-like signal peptides, and bactericin and pheromone type signal peptides are most prevalent (Tjalsma et al., 2004). However, mainly the first three types of signal peptides (i.e., Sec type, TAT type, and Lipo type) are common in gram-positive bacteria (Pallen et al., 2003). Sec type and Tat type signal peptides are cleaved by Type I SPase, whereas Lipo types are cleaved by Type II SPase (Storf et al., 2010).

The tremendous advancement in genome sequencing technology has yielded complete genome sequences of a broad range of bacterial population. Automated prediction of the secretome has generated a lot of interest. Prediction of the signal peptide-containing genes, along with their cleavage sites in the finished bacterial genomes, have been achieved by employing various algorithms such as Hidden Markov Model (HMM), Neural Network (NN) (Bendtsen et al., 2004), and Support Vector Machines (SVM) (Vert, 2002).

There have been various web-based servers that employ these algorithms and use perl scripts to predict the secretomes accurately in a given genome such as Signal P, Signal-CF, SIGCLEAVE, Predisi, SPEPLip, SecretomeP, and Phobius. Bioinformatics-based analysis and comparison of secretomes have been performed in a few cases; however, extensive information pertaining to the features and behavior of secretomes in various bacterial genomes remains to be plowed from the depths of secretomic study. The study of codon usage patterns, expressional behavior, and functional classification of the predicted secretomes and also the evolutionary constraints on these secretomes in many bacterial genomes still remain elusive.

Significant progress has been made in the field of mycobacterial secretome analysis and their possible role in infections. Secretory proteins were reported to be crucial for the efficiency of BCG vaccines (Heimbeck, 1948). Various other novel findings by Gomez et al., (2000), Rosenkrands et al., (2000), McDonough et al., (2005, 2008) have conferred considerable information about mycobacterial secretory systems and their varying types. However, predicting secretomes in mycobacterial genomes appears to be a difficult chore due to the unusual nature of mycobacterial cell membranes (Leversen et al., 2009). Recently Leversen et al. (2009) reported a set of confirmed signal peptides in the *Mycobacterium tuberculosis* H37Rv strain by validating the putative signal peptides that they and previous researchers had analyzed, employing various algorithms and finally matching them with high accuracy MS data. However, a complete schema of signal peptides in *Mycobacterium* is yet to be reported.

Keeping this in mind, we have developed the database, MycoSec, a repository of potential mycobacterial secretomes. The database has a plethora of information pertaining to the secretome analysis in mycobacterial strains and presents information in an organized manner. The putative secretomes that have been included in the database promise to be strong candidates for being confirmed signal peptides once experimental validation is accomplished.

Materials and Methods

Software

The database has been designed on a HTML platform using the Macromedia Dreamweaver database development software version 8.

Strategies for identification of Mycobacterial secretomes

Identifying the initial pool of secretomes: Complete genome sequences of *Mycobacterium* strains were retrieved from the IMG website (http://img.jgi.doe.gov/cgi-bin/w/main.cgi) (Markowitz et al., 2006). Initially, forty-one strains representing twenty-five species were taken for analysis. However, more species will be added as and when available. Predictions of signal peptides were done with SignalP (version 3.0). Although several other algorithms, including a newer version of SignalP, (SignalP 4.0) are available, we used SignalP 3.0 because as per Leversen et al., (2009) and Leversen and Wiker, (2012), SignalP 3.0 is more suitable for accurate prediction of signal peptides in *Mycobacteria* than any other web-server, including Signal P 4.0. Gram-positive bacteria have been found to secrete proteins in the external environment by virtue of three important pathways (Pallen et al. 2003). These include Sec (general secretion) pathway, Twin arginine transporter (TAT) pathway, ESAT-6 pathway (Champion, 2007), Type VII secretion system (Abdallah et al., 2007), and most importantly Lipo signaling pathway (Rezwan et al., 2007). Since mycobacteria are gram-positive bacteria, we use these three types of signaling systems for identification and analysis. The primary pool was processed in three different ways for the classification of signal peptides.

Identification of Sec Type signal peptides. The initial pool of secretomes was fed to TMHMM (version 2.0) server in order to fish out the sec type of signal peptides from the transmembrane proteins. We have considered protein sequences, with 0 to 2 transmembrane helices, as potential sec type of signal peptides as per Mastronunzio et al. (2008) and Gore (2011).

Filtering lipoprotein-type signal peptides. For filtering lipoprotein-type signal peptides, two algorithms (Pred-Lipo and LipoP) are widely used. However, we used Pred-Lipo, which operates on the Hidden Markov Model, and has been reported to be the most efficient in terms of prediction accuracy and reports the lowest false positives (Bagos et al., 2008). The SignalP predicted data set was fed to Pred-Lipo server (http://www.compgen.org/tools/PRED-LIPO) for lipoprotein prediction.

TAT-type signal peptide prediction. Among the three widely used prediction servers (Pred-Tat, TatP, and TatFind), TatFind has a slight edge over others as it executes on a combined approach of regular expression search (searching twin arginine-RR/KR pattern) and hydrophobicity analysis (Rose et al., 2002). Moreover, TatFind results are more specific while matching with experimentally validated set of proteins. Similar to previous section, the SignalP predicted data set was

fed to the TatFind server (http://signalfind.org/tatfind.html) for TAT-type signal peptide prediction.

A complete flowchart depicting our method of *in silico* identification of signal peptides is illustrated in Figure 1.

Comparison with experimentally validated data

In silico prediction of any kind always demands an experimental validation. Scarcity of experimental wet lab data is a major bottleneck in the field of mycobacterial secretomic research. However, Leversen et al. (2009) identified fifty-seven signal peptides and confirmed them by experimental validation in *Mycobacterium tuberculosis* H37Rv. We have matched our results with those of Leversen et al. (2009) and found that around eighty-one percent of the signal peptides identified by Leversen and co-workers were present in our identified dataset of H37Rv strain. Leversen et al. (2009) also identified around sixty-one proteins that had the potential to be signal peptides, but were experimentally validated to be nonsignal peptides. All of these proteins were not screened by our identification method. This shows that our method is quite accurate. However, eleven proteins identified by Leversen et al. (2009) as signal peptides in strain H37Rv were not filtered by our method of identification (*Locus tags: Rv0519c, Rv0744c, Rv0999, Rv1845c, Rv2693c, Rv3484, Rv3717, Rv0129c, Rv0285, Rv2576c, Rv2878c*). This may be accounted by the stringency of our identification schema. These particular proteins have also been incorporated in our database marked with an asterisk (*).

Database description

The database main page consists of the following interfaces:

HOME. The homepage provides a general introduction to the genus *Mycobacterium* and pathogenesis of different mycobacterial strains. It also discusses the characteristics of secretomes and utility of research in the area of secretomics. The page also provides an idea as to why the database has been developed, thus sketching the advantages of the database in

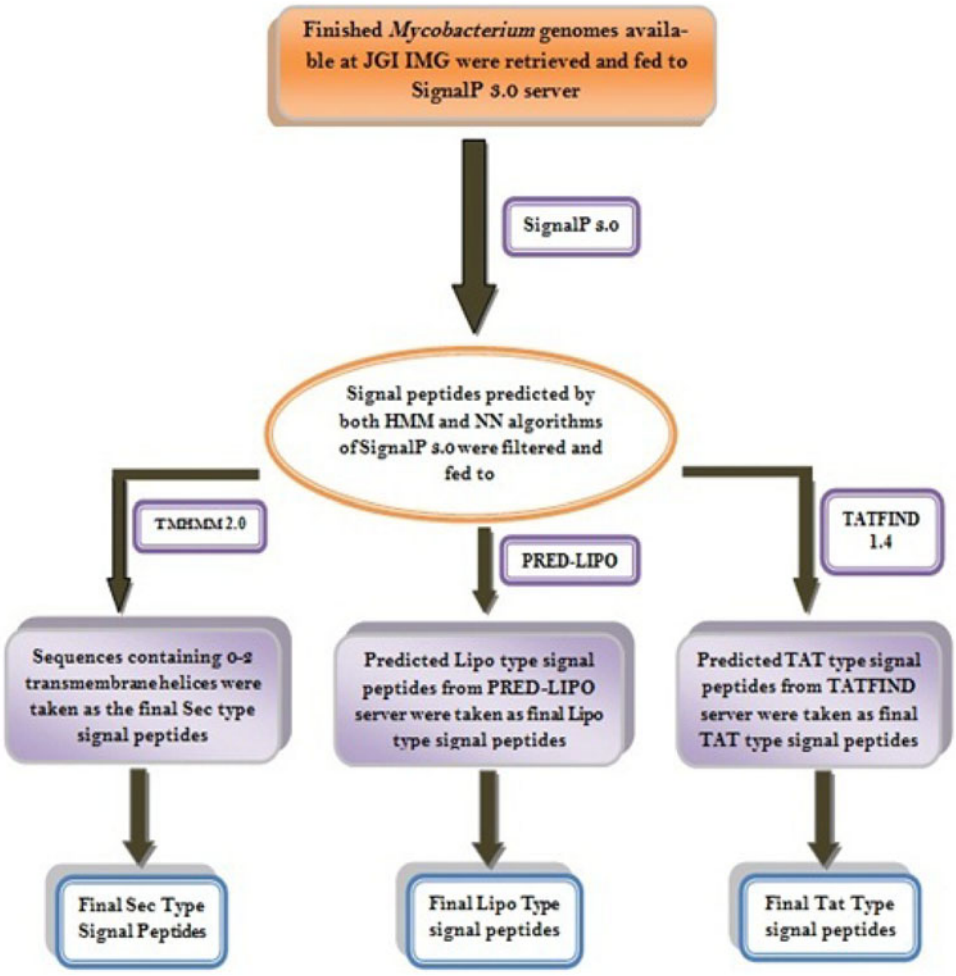

FIG. 1. Flowchart displaying the method of in silico identification of signal peptides. Completely sequenced Mycobacterium genomes were used as input data. The genomes were fed to SignalP 3.0 web-server for initial prediction of secretomes. Sequences that were predicted by both Hidden Markov Model and Neural Network algorithms of SignalP 3.0 were screened and used as the initial set of secretomes. The initial pool was fed to three different web based servers (i.e., TMHMM 2.0, PRED LIPO, and TATFIND 1.4) for the prediction of secretory signal peptides (Sec type), lipoprotein signal peptides (Lipo type), and Twin arginine (RR/KR) signal peptides (TAT type) respectively, allowing us to identify the final set of signal peptides of each type.

mycobacterial research. There is a "QUICK SEARCH" drop down menu comprising the list of mycobacterial species/strains analyzed in all major pages. Clicking on a species/strain will take the user to the specific page displaying all pertinent information regarding the secretome classification and properties of that particular strain.

ORGANISMS. The ORGANISMS page displays the list of all the mycobacterial species/strains on which we have executed our analysis.

ANALYSIS. The analysis page describes the general scheme adopted for prediction and analysis of secretomes.

USEFUL LINKS. For the benefit of the users, this page has links to all major web-servers and tools used in MycoSec.

FUTURE PLANS. This includes our future plans as how to improve the database and update the contents with the availability of more finished sequences of *Mycobacterium* at publicly available domains.

ABOUT US. Provides an insight into the field of research being carried out by our group, recent developments and activities.

CONTACT. Contact information of corresponding author and our group members who have been instrumental in developing the database.

ORGANISM specific page/analysis page for a specific strain. These species/strain specific pages contain general information about the respective species/strain. These pages also contain icons which lead the users to *TAT-TYPE, LIPO-TYPE, and SEC-TYPE* specific analysis.

Each specific analysis page consists of all the general information regarding the predicted secretomes in tabular form in various columns: GenBank Accession, Locus Tag, COG Categories, GC3%, and Nc & CAI values. The indexes *Twin Arginine* and *Hydrophobic Region* are specific for the TAT-type pages, as the twin arginine region is a specific characteristic pattern of TAT-type signal peptides. Similarly, the LIPO-TYPE page contains the parameters: *Most likely cleavage site*--the predicted cleavage site, *Cleavage at*–signifying the position of cleavage by Type II SPase, and the *Reliability score*–the reliability score for cleavage prediction by PRED-LIPO server.

Each specific secretome type page contains five icons on the top:

GC3/CAI-Nc plot. GC (frequency of guanine and cytosine), GC3s (frequency of guanine or cytosine in the third position of the codon), and Nc (effective number of codons) of the mycobacterial genomes and secretomes were calculated using CodonW (Ver. 1.4.2) software (http://www.molbiol.ox.ac.uk/cu) (Peden, 1999). The effective number of codons has always been an important index in understanding the extent of codon preference in codon usage of a genome (Wright, 1990). It is a quantitative measure reflecting the frequency of a subset of codons used by a gene and its value ranges from 20 (on usage of one codon per amino acid) to 61 (on usage of all the codons with equal frequency excluding the termination codons).

Codon adaptation index (CAI) has been a well-established parameter in determining the extent of codon usage bias for a gene concerned relative to a reference set of genes (usually ribosomal proteins) (Sharp and Li, 1987). CAI values have been employed extensively as measure of gene expression level (Ikemura, 1981; Naya et al., 2001; Wright and Bibb, 1992). Higher CAI values signify higher expression levels of genes in a genome (Sen et al., 2008) and generally highly expressed genes are more biased than the lowly expressed ones (Dos Reis et al., 2003; Lafay et al., 2000; Sharp and Li, 1986, 1987). It is hypothesized that, in a genome, the codon usage of highly expressed genes are governed by selection pressure for translational efficiency, whereas mutational bias influences the codon usage of the lower expressed ones (Sharp and Li, 1987). The CAI values for the mycobacterial secretomes were calculated to explore their expression tendencies. CAI values have been calculated using the CAI Calculator 2 server (http://userpages.umbc.edu/~wug1/codon/cai/cais.php) (Wu et al., 2005).

The upper plot in each page represents the GC3 versus Nc values for the whole geome of a strain under scrutiny with an insight to the ribosomal proteins and predicted secretomes. The lower plot represents the CAI vs Nc values for the secretomes with respect to the ribosomal proteins and the whole genome.

CoA graph. Correspondence analysis (CoA), a type of multivariate statistical analysis, has been very instrumental in studying the codon usage patterns in a single genome and between different genomes (Ghosh et al., 2000; Greenacre, 1984). Relative synonymous codon usage (RSCU) is a simple measure of the heterogeneity in the usage pattern of synonymous codons (Sharp and Li, 1986). RSCU values represent the number of times a particular codon is observed relative to the number of times it would have been expected in case of a uniform synonymous codon usage. Correspondence analysis on the basis of RSCU and amino acid usage of the secretomes with respect to the ribosomal protein coding genes and whole genomes were also calculated using CodonW (Ver. 1.4.2) software.

In the plot, the upper figure displays the correspondence analysis on amino acid usage of the predicted secretomes in contrast to the whole genome and the ribosomal proteins. The lower figure depicts the correspondence analysis on RSCU of the secretomes in reference to the whole genome and ribosomal proteins.

COG analysis. This page has the graphical representation of the Cluster of Orthologous Groups (COG) categories of the predicted secretomes. COGs comprise the collection of orthologous proteins from similar phylogenetic lineage (Tatusov et al., 2003). Information regarding the COG categories of the potential secretomes was obtained from the IMG database. The genes encoding the three different types of signal peptide containing proteins were sorted into different COG categories such as *Information Storage and Processing*, *Cellular Processes and Signaling*, *Metabolism,* and *Poorly characterized* in accordance with the classification scheme followed by Hsiao et al. (2005).

Sequences in FASTA. Users can retrieve and download all the gene and protein sequences predicted as secretomes for

a particular mycobacterial strain by clicking on 'Genes in FASTA' and 'Proteins in FASTA' icons respectively. An overall description of MycoSec is illustrated in Figure 2. Figure 3a, b, c depicts the snapshots of various pages of the database.

Results

MycoSec contains the predicted secretomes and various bioinformatic analysis related to secretomes of almost all 'finished' mycobacterial genomes. We have generated all relevant information regarding the codon usage indices, expressional patterns (using CAI values), and codon usage bias in the mycobacterial genomes. The results are given in both tabular as well as graphical form, which may provide the users with general information about the forces that have been instrumental in shaping the codon usage patterns in the genomes as well as the secretomes. The COG (cluster of orthlogous group) can be employed to have a brief knowhow of the functional classification of predicted secretory protein genes.

GC3 versus Nc plots

The effective number of codons (Nc) versus the GC3s graphical plot has been recommended to be an efficient way in investigating the extent of heterogeneity in a given genome. The Nc versus the GC3s graphical plots in our case depict that majority of the genes, along with the signal peptide coding genes, in all the genomes concerned, fall well below the expected curve. A few genes, however, remain on or just below the curve as evident from the plots.

CAI versus Nc plots

The CAI versus Nc plots have also been generated to provide a clear understanding of the expressional pattern of the secretomes. The CAI values for the secretomes range from 0.4 to 0.8, at the maximum, for all strains under investigation.

Correspondence analysis on the basis of RSCU and amino acid usage

Multivariate statistical analysis performed on the basis of RSCU and amino acid usage can also be employed to explore the codon bias in genes and genomes (Sen et al., 2008). Results from the CoA plots (on the basis of both RSCU and amino acid usage) portray that the ribosomal proteins cluster at one extreme end of the major principal axis and secretome-related genes were found to merge somewhat with this cluster, on plotting Axis 1 versus Axis 2, the two major principle axes of separation.

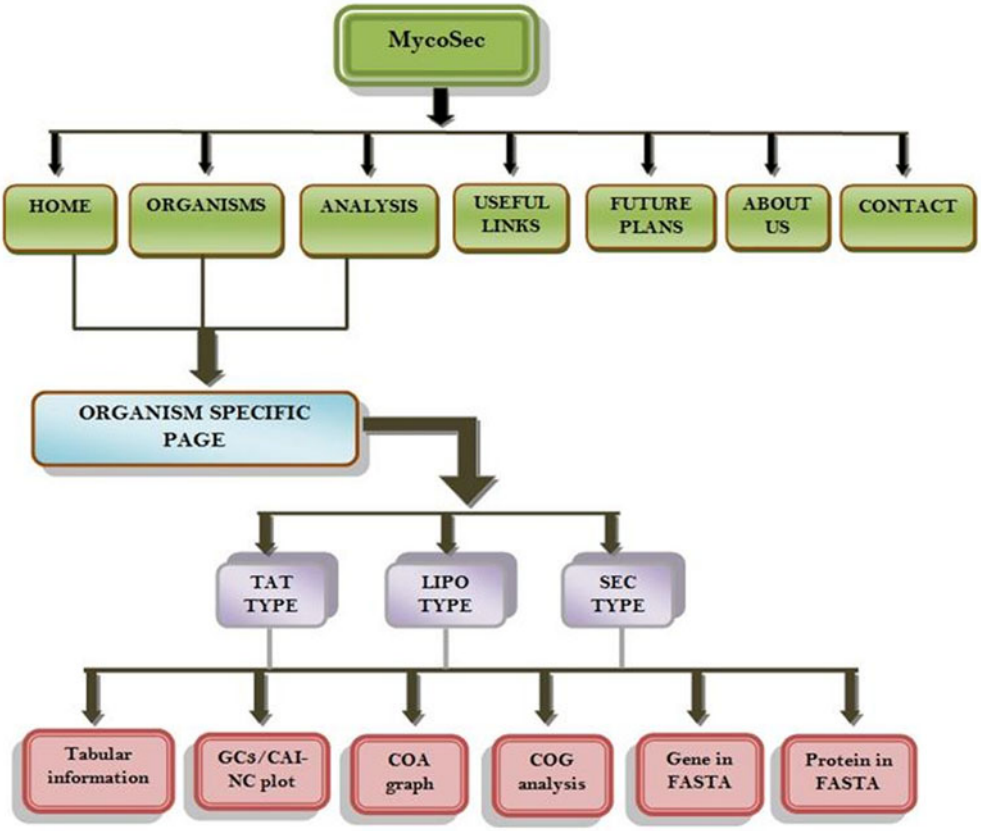

FIG. 2. Flowchart describing the database MycoSec. The database comprises of seven major interfaces as listed in *green color code* in the figure. The three interfaces—HOME, ORGANISMS, and ANALYSIS leads to the ORGANISM Specific Page. An ORGANISM Specific Page has information about a particular strain of Mycobacteria and has links to its Sec, Lipo, and TAT type signal peptides. Users can use this link to visit and retrieve specific information pertaining to each type of signal peptides.

GenBank Accession	Locus Tag	gene_oid	Product Name	AA Seq Length	Pfams
NP_214554	Rv0040c	637025209	SECRETED PROLINE RICH PROTEIN MTC28 (PROLINE RICH 28 KDA ANTIGEN)	310	pfam10738
NP_214577	Rv0063	637025232	POSSIBLE OXIDOREDUCTASE	479	pfam01565 pfam08031
YP_177692	Rv0109	637025279	PE-PGRS FAMILY PROTEIN	496	pfam00934
NP_214630	Rv0116c	637025286	POSSIBLE CONSERVED MEMBRANE PROTEIN	251	pfam03734
NP_214639	Rv0125	637025295	PROBABLE SERINE PROTEASE PEPA (SERINE PROTEINASE) (MTB32A)	355	pfam00089 pfam00595
NP_214685	Rv0171	637025341	MCE-FAMILY PROTEIN MCE1C	515	pfam02470

FIG. 3. Snapshot of various pages of MycoSec: **a)** Homepage; **b)** Genome information page; **c)** Analysis page.

COG graphs

The COG graphs reveal that the majority of the secretomes fall under the categories *'Cellular Processes and Signaling'* and *'Metabolism'*, while very few lie in the *'Information Storage and Processing'* category. Among them, COG M (cell wall/membrane/envelope biogenesis), was found to be most abundant in all types of secretomes for all the strains studied.

Discussion

Synonymous codon usage bias in prokaryotic genomes has been inferred to be shaped by the effects of translation efficiency and mutation bias. The effective number of codons (Nc) versus the GC3s graphical plots can be employed as a tool to determine the forces that govern the codon usage patterns. Genes whose codon bias are entirely governed by a mutation bias must lie on or just below the curve in a GC3 versus Nc plot, and genes lying well below the expected curve are considered to be under the influence of translational selection (Peden, 1999). It can be easily deduced from the GC3 versus Nc plots, from the present study, that a majority of the genes encoding the signal peptides are under the effect of selection for translational efficiency. However, a few genes also display the influence of mutation bias. This trend has been found in all the strains under study.

Focusing on the expressional behavior, it is quite evident from the CAI versus Nc plots that the secretomes are moderately expressed.

Correspondence analysis (CoA) is a congregated technique that highlights the major tendencies in the variation of data and places them along the continuous axes according to the variations observed (Banerjee et al., 2012). Selection force due to translational efficiency can be inferred to be acting on the genomes when the ribosomal proteins cluster at any extreme end of the major principal axis in a CoA plot based on RSCU and amino acid usage (Peden, 1999). A similar trend was noticed for all the mycobacterial genomes on plotting Axis 1 versus Axis 2, the two major principle axes of separation. Correspondence analysis reveals the crucial role of translational selection pressure in shaping the codon usage pattern of the whole genome as well as the secretomes, along with a subtle effect of mutation bias.

Conclusion

Research on mycobacterial pathogenesis has always been a topic of immense interest in biomedical sciences and has taken a giant leap with the advancement of genome sequencing programs. Numerous genomes of *Mycobacterium* have been sequenced and the number is increasing day by day. It is now a daunting task to cluster and analyze the huge amount of data that are being generated from these genome sequencing programs to a meaningful conclusion in a reasonable time. It was therefore a humble effort from our group to analyze at least the secretome-related information of all sequenced mycobacterial genomes and bring the information into one specific platform (the MycoSec) for the valued researchers. MycoSec is freely accessible at http://www.bicnbu.in/mycosec and will be updated and expanded regularly. MycoSec, a repository of potential mycobacterial signal peptides, can divulge much information underlying pathogenic infections at the molecular level and promises to provide ample avenues for developing novel therapeutics for eradication of the mycobacteria-related diseases.

Acknowledgments

The authors are grateful to the Department of Biotechnology, Government of India, for providing financial help in setting up Bioinformatics Infrastructural facility at University of North Bengal. A. Sen acknowledges the receipt of the DBT-CREST Award. Early findings were presented as an abstract in the *International Interdisciplinary Science Conference* held at Jamia Malia University, Delhi, India, in 2011.

Author Disclosure Statement

No competing financial interests exist.

References

Abdallah AM, van Pittius NCG, Champion PADG, et al. (2007). Type VII secretion—Mycobacteria show the way. Nat Rev Microbiol 5, 883–891.

Bagos PG, Nikolaou EP, Liakopoulos TD, and Tsirigos KD. (2010). Combined prediction of Tat and Sec signal peptides with hidden Markov models. Bioinformatics 26, 2811–2817.

Bagos PG, Tsirigos KD, Liakopoulos TD, and Hamodrakas SJ. (2008). Prediction of lipoprotein signal peptides in Gram-positive bacteria with a Hidden Markov Model. J Proteome Res 7, 5082–5093.

Banerjee R, Roy A, Ahmad F, Das S, and Basak S. (2012). Evolutionary patterning of hemagglutinin gene sequence of 2009 H1N1 pandemic. J Biomol Struct Dyn 29, 733–742.

Bendtsen JD, Nielsen H, Von Heijne G, and Brunak Sr. (2004). Improved prediction of signal peptides: SignalP 3.0. J Mol Biol 340, 783–795.

Bonin-Debs AL, Boche I, Gille H, and Brinkmann U. (2004). Development of secreted proteins as biotherapeutic agents. Expert Opin Biol Ther 4, 551–558.

Champion PA, and Cox JS. (2007). Protein secretion systems in Mycobacteria. Cell Microbiol 9, 1376–1384.

Cole ST, Eiglmeier K, Parkhill J, et al. (2001). Massive gene decay in the leprosy bacillus. Nature 409, 1007–1011.

De Voss JJ, Rutter K, Schroeder BG, Su H, Zhu YQ, and Barry CE. (2000). The salicylate-derived mycobactin siderophores of *Mycobacterium tuberculosis* are essential for growth in macrophages. Proc Natl Acad Sci USA 97, 1252.

Dos Reis M, Wernisch L, and Savva R. (2003). Unexpected correlations between gene expression and codon usage bias from microarray data for the whole *Escherichia coli* K-12 genome. Nucleic Acids Res 31, 6976–6985.

Ghosh TC, Gupta SK, and Majumdar S. (2000). Studies on codon usage in *Entamoeba histolytica*. Int J Parasitol 30, 715–722.

Gomez M, Johnson S, and Gennaro ML. (2000). Identification of secreted proteins of Mycobacterium tuberculosis by an ioinformatics approach. Infect Immun 68, 2323–2327.

Gore D. (2011). In silico identification of cell surface antigens in Neisseria ioinformati. Biomirror 2, 1–5.

Greenacre MJ. (1984). *Theory and Applications of Correspondence Analysis.* Academic Press, London.

Heimbeck J. (1948). BCG vaccination of nurses. Tubercle 29, 84–88.

Hsiao WW, Ung K, Aeschliman D, Bryan J, Finlay BB, and Brinkman FS. (2005). Evidence of a large novel gene pool associated with prokaryotic genomic islands. PloS Genet 1, e62.

Ikemura T. (1981). Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: A proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. J Mol Biol 151, 389–409.

Lafay B, Atherton JC, and Sharp PM. (2000). Absence of translationally selected synonymous codon usage bias in *Helicobacter pylori.* Microbiology 146, 851–860.

Lee VT, and Schneewind O. (2001). Protein secretion and the pathogenesis of bacterial infections. Genes Dev 15, 1725–1752.

Leversen NA, de Souza GA, Malen H, Prasad S, Jonassen I, and Wiker HG. (2009). Evaluation of signal peptide prediction algorithms for identification of mycobacterial signal peptides using sequence data from proteomic methods. Microbiology 155, 2375–2383.

Leversen NA, and Wiker HG. (2012). Improved signal peptide predictions in mycobacteria? Tuberculosis 92, 291–292.

Markowitz VM, Ivanova N, Palaniappan K, et al. (2006). An experimental metagenome data management and analysis system. Bioinformatics 22, e359–e367.

Mastronunzio JE, Tisa LS, Normand P, and Benson DR. (2008). Comparative secretome analysis suggests low plant cell wall degrading capacity in Frankia symbionts. BMC Genomics 9, 47.

McDonough JA, Hacker KE, Flores AR, Pavelka MS, and Braunstein M. (2005). The twin-arginine translocation pathway of *Mycobacterium smegmatis* is functional and required for the export of mycobacterial beta-lactamases. J Bacteriol 187, 7667–7679.

McDonough JA, McCann JR, Tekippe EME, Silverman JS, Rigel NW, and Braunstein M. (2008). Identification of functional Tat signal sequences in *Mycobacterium tuberculosis* proteins. J Bacteriol 190, 6428–6438.

Miller CD, Hall K, Liang YN, et al. (2004). Isolation and characterization of polycyclic aromatic hydrocarbon-degrading mycobacterium isolates from soil. Microbial Ecol 48, 230–238.

Naya H, Romero H, Carels N, Zavala A, and Musto H. (2001). Translational selection shapes codon usage in the GC-rich genome of *Chlamydomonas reinhardtii.* FEBS Lett 501, 127–130.

Niederweis M, Danilchanka O, Huff J, Hoffmann C, and Engelhardt H. (2010). Mycobacterial outer membranes: In search of proteins. Trends Microbiol 18, 109–116.

Pallen MJ, Chaudhuri RR, and Henderson IR. (2003). Genomic analysis of secretion systems. Curr Opin Microbiol 6, 519–527.

Peden J. (1999). Analysis of codon usage. PhD Thesis, The University of Nottingham, UK.

Ranganathan S, and Garg G. (2009). Secretome: Clues into pathogen infection and clinical applications. Genome Med 1, 113.

Rezwan M, Grau T, Tschumi A, and Sander P. (2007). Lipoprotein synthesis in mycobacteria. Microbiology 153, 652–658.

Rose RW, Bruser T, Kissinger JC, and Pohlschroder M. (2002). Adaptation of protein secretion to extremely high-salt conditions by extensive use of the twin-arginine translocation pathway. Mol Microbiol 45, 943–950.

Rosenkrands I, Weldingh K, Jacobsen S, et al. (2000). Mapping and identification of *Mycobacterium tuberculosis* proteins by two-dimensional gel electrophoresis, microsequencing and immunodetection. Electrophoresis 21, 935–948.

Sen A, Sur S, Bothra AK, Benson DR, Normand P, and Tisa LS. (2008). The implication of life style on codon usage patterns and predicted highly expressed genes for three Frankia genomes. Antonie Leeuwenhoek 93, 335–346.

Sharp PM, and Li WH. (1986). An evolutionary perspective on synonymous codon usage in unicellular organisms. J Mol Evol 24, 28–38.

Sharp PM, and Li WH. (1987). The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res 15, 1281–1295.

Storf S, Pfeiffer F, Dilks K, Chen ZQ, and Imam S. (2010). Mutational and ioinformatics analysis of haloarchaeal lipobox-containing proteins. Archaea 2010, 1–11.

Tatusov RL, Fedorova ND, Jackson JD, et al. (2003). The COG database: An updated version includes eukaryotes. BMC Bioinformatics 4, 41.

Tjalsma H, Antelmann H, Jongbloed JD, et al. (2004). Proteomics of protein secretion by *Bacillus subtilis*: Separating the "secrets" of the secretome. Microbiol Mol Biol Rev 68, 207–233.

Vert JP. (2002). Support vector machine prediction of signal peptide cleavage site using a new class of kernels for strings. Proc Pacific Sympos Biocomput Citeseer, pp. 649–660.

Von Heijne G. (1984). How signal sequences maintain cleavage specificity. J Mol Biol 173, 243–251.

Von Heijne G. (1989). The structure of signal peptides from bacterial lipoproteins. Protein Eng 2, 531–534.

Von Heijne G. (1990a). Protein targeting signals. Curr Opin Cell Biol 2, 604.

Von Heijne G. (1990b). The signal peptide. J Membr Biol 115, 195–201.

Wright F. (1990). The 'effective number of codons' used in a gene. Gene 87, 23–29.

Wright F, and Bibb MJ. (1992). Codon usage in the G+C-rich Streptomyces genome. Gene 113, 55–65.

Wu G, Culley DE, and Zhang W. (2005). Predicted highly expressed genes in the genomes of *Streptomyces coelicolor* and *Streptomyces avermitilis* and the implications for their metabolism. Microbiology 151, 2175–2187.

Zumla AI, and Grange J. (2002). Non-tuberculous mycobacterial pulmonary infections. Clin Chest Med 23, 369–376.

Address correspondence to:
Arnab Sen
Bioinformatics Facility
Department of Botany
University of North Bengal
Siliguri 734013
India

E-mail: senarnab_nbu@hotmail.com