Choke Point Analysis with Subtractive Proteomic Approach for Insilico Identification of Potential Drug Targets in Shigella Dysenteriae

Manmohan Pandey
AMITY Institute of Biotechnology
AMITY University Uttar Pradesh
Lucknow, UP, India

Anshul Tiwari
AMITY Institute of Biotechnology
AMITY University Uttar Pradesh
Lucknow, UP, India

Shalini Maurya
AMITY Institute of Biotechnology
AMITY University Uttar Pradesh
Lucknow, UP, India

Dipika Singh
AMITY Institute of Biotechnology
AMITY University Uttar Pradesh
Lucknow, UP, India

Prachi Srivastava, Ph.D
AMITY Institute of Biotechnology
AMITY University Uttar Pradesh
Lucknow, UP, India

ABSTRACT
Shigellosis is an endemic disease prevalent in developing and poor countries due to fecal-oral transmission resulting a significant morbidity and mortality rate. Emergence of multi-drug resistant (MDR) in Shigella sp. reveals the inefficacy towards the first line antibiotics like quinolones, co-trimoxazole and ampicillin against it. There is continuous need to monitor the characteristics and antibiotic resistance patterns of this pathogen regarding the identification of new potential therapeutic drug targets. Availability of complete protein of different Shigella species viz flexneri, body, dysentery and son has made it possible to carry out the In-silico analysis of its protein for the identification of potential vaccine and drug targets. Subtractive proteomics approach is being used to mine the list of proteins present in different Shigella species which are non-homologous to human and essential for the survival of the pathogen. The metabolic chokepoint analysis also enriches the list of essential protein and adds those proteins in the list which are uniquely found in pathogenic pathway, catalyzed by single enzyme and involved in multi pathways. Screening of essential proteins against human gut flora and approved drug targets revealed the targets which are non-homologous to human gut flora and homologous to the approved drug targets. Broad spectrum drug targets screening revealed a list of highly conserved proteins of various pathogens including different Shigella species. Probably the drug developed against these targets may be useful in treating multiple diseases or diseases which results due to co-infection of different pathogens. Subcellular localization prediction revealed a list protein, which could be potential vaccine targets in different Shigella species. Virtual screening against these identified targets might be useful in the discovery of novel Drug against MDR Shigella species.

General Terms
Shigellosis, Multi-drug resistant, WHO, in-vitro, gene knockouts, drug target, mutagenesis, non-homologous, dysentriae.

Keywords
Drug targets, DEG, metabolic choke point, subtractive proteomic, Shigella, BioCyc.

1. INTRODUCTION
Shigellosis, an enteric infectious disease, affects approximately 164.7 million patients and causes 1.1 million deaths each year globally which includes about 60% children of less than 5 years of age. The causative agent of Shigellosis is Shigella spp which includes *S. flexneri, S.boydii, S. dysentriae* and *S. soneii*. It is highly frequent in developing countries with *S. flexneri* (60%) as the most common pathogen, and rare in industrialized countries with *Shigella soneii* as the main pathogen (77%) [1][2][3]. Shigellosis was ranked 6th on the basis of morbidity caused by infectious disease in China [4]. Since last decade, Shigella species resistance have gradually increased to most of the widely used antimicrobial agents and thus resulting in reduced efficacy of antimicrobial therapies [5][6]. Ampicillin, co-trimoxazole, chloramphenicol and nalidixic acid were the most commonly used antibiotics for the treatment of Shigellosis worldwide. Since antimicrobial resistance among Shigella spp. varies from region to region, thus, WHO has recommended ciprofloxacin as the first line antibiotic for the treatment of Shigellosis. Recently various ciprofloxacin resistant isolates were also reported from different parts of India [7]. Therefore, limited options are left for the treatment of the shigellosis. Presently, cephalosporin is widely used in the case of multi resistant Shigella infection (especially in pediatric cases). But due to possibility of emergence of extended spectrum beta lactamase (ESBL) producing Shigella species, thus, there is continuous need to monitor the characteristics and antibiotic resistance patterns of this pathogen and to identify new potential therapeutic drug targets [8][9]. Drug target identification and validation in *In vitro* conditions involve whole cell or animal experiments, gene knockouts or site-directed mutagenesis experiments. Thus, traditionally identification and characterization of drug is time consuming, expensive and requires lots of manual efforts. *In silico* methods are widely accepted for systematic and large scale validation of targets as it has advantage of speed and cost effectiveness [10].

Availability of complete proteome of different Shigella species viz flexneri, boydii, dysentriae and soneii has made it possible to carry out the in silico analysis of its proteome for the identification of potential vaccine and drug targets.

The current study on different Shigella species makes use of substractive proteomics approach to mine potential and putative drug targets which are non-homologous to human and essential for the survival of pathogen. Further, this study also incorporates in finding out the targets; those are unique to the pathogen, non-homologous to human gut flora and can act as broad spectrum drug targets.

2. MATERIAL AND METHOD

The essential protein in Shigella spp. were first identified using substractive proteomics approach and further analyzed for identification of potential drug targets.

2.1 Retrieval of Host and pathogen Proteome

The complete proteome of *S. flexneri* (n= 4456) (strain: 2002017), *S. boydii* (n= 3938)   (strain: Sb227), *S. soneii* (n= 4068) (strain: strain Ss046) and *S. dysentriae* (n= 3897) (strain: Sd197) were downloaded from HAMAP project housed at Expasy and complete proteome of Homo sapiens (n=74390) was downloaded from National center for biotechnological information. The proteins with less than 50 AA residues and any protein which have non-amino-acid characters are removed [11][12].

2.2 Identification of close homologs in the human proteome

The complete proteome of *S. flexneri*, *S. boydii, S. soneii* and *S. dysentriae* were purged at 60% using CD-HIT [13] to identify the paralogs or duplicate proteins within the proteome of each of these Shigella species. The paralogs were excluded and the remaining set of protein was subjected to Blast P [14] against complete proteome of Homo sapiens with the expectation value (E-value) cut off score of 10-5. The resultant dataset obtained had no significant similarity or less than 30% similar with proteome of Homo sapiens. An in-house developed PERL script was used to parse BlastP result.

2.3 Identification of Essential protein in Shigella species

Non-homologous protein sequences of each of these Shigella species were subjected to BLASTP against the Database of Essential Gene (DEG) with E-value 10-8, 60% sequence identity and Bits score greater than 100 [15].

Further Metabolic chokepoint analysis is done using BioCyc webserver to identify Essential protein using the criteria: Exclude reactions found in human, exclude reaction catalyzed by more than one enzyme and limit to the reaction found in multiple pathways [16][17].

Chokepoint reaction is defined as a reaction that either consumes a unique substrate or produces a unique product. A chokepoint analysis provides a systematic method of identifying novel potential drug targets. Thus, the list of the targets obtained after metabolic choke point analysis is further compared with the non-homologous sets of protein and only those are retained which are found in both the sets.

Finally the targets screened with these two methods were merged into one file and thus it constitutes the final list of the Essential targets.

2.4 Further Analysis of short listed targets

The short-listed targets were subjected to further analysis, to retain only those proteins that are highly targetable.

2.4.1 Druggability Analysis

To identify the druggable protein from a list of essential protein all approved target sequences from drug bank have been downloaded [18]. The drug binding domain sequences of these proteins were used to identify the corresponding CDD (Conserved Domain database) domain of NCBI and a dataset of 1030 protein is compiled. Similarly, CDD domain of each of the essential protein has been retrieved. Comparison of identified drug binding domain with the domain of essential protein is carried out [19].

2.4.2 Broad Spectrum Drug Targets

To identify Broad spectrum targets complete list of prokaryotic proteome (organism wise) have been downloaded from DEG. CDD domain of these Essential proteins has been retrieved from NCBI CDD. Similarly, CDD domain of each of the essential protein has been retrieved. Domains which are common to at least 3 of the Shigella species including S. Flexneri are further retained. This retained dataset is compared to the compiled domain of Essential proteins retrieved from DEG and if it is present in at least 15 species than it is considered as Broad spectrum targets.

2.4.3 Subcellular Localization Prediction

Sub-cellular localization analysis of the essential proteins has been done by psortb to identify the surface membrane proteins which could be probable vaccine targets.

2.4.4 Non-homologous to gut flora

Essential proteins are subjected to BlastP against complete proteome of gutflora with the expectation value (E-value) cut off score of 10-5. The resultant dataset obtained had less than 30% of sequence similarity or no significant similarity with proteome of Human gutflora. Gutflora protein is retrieved from NCBI database using complex search criteria.

3. RESULT AND DISCUSSION

Shigella flexneri comprises of 4456 protein sequences in their proteome while *S. dysentriae*, *S soneii* and *S. boydii* have 3897, 4068 and 3938 proteins in their proteome respectively [20]. Different filter criteria used against each of these species are listed (table 1, figure 1). After initial dataset preparation of Shigella species including removal of the proteins having non aa character and the proteins which is less than 50 aa residue, the resultant dataset contains 4319 residue in case of *S. flexneri* while *S. dysentriae*, *S soneii* and *S. boydii* have 3856, 4030 and 3904 respectively. The resultant dataset is further converted to non-redundant dataset by using CD-HIT program at 60% identity [21]. Thus the non-redundant dataset have 3898 proteins in case of *S. flexneri* while 3393, 3740 and 3506 respectively in case of *S. dysentriae*, *S soneii* and *S. boydii*. Non-redundant dataset is further filtered and only those proteins which either have sequence similarity less than 30% or do not have any significant similarity reported by BlastP against human proteome is carried out. Therefore, out of 3898 proteins 3191 found to be non-homologous to human in case of *S. flexneri* while 2735 out of 3393, 3002 out of 3740 and 2817 out of 3506 respectively were found to be non-homologous with human proteome. Further these non-homologous sets to human were subjected to BLASTP against prokaryotic DEG with E-value cutoff score of 10-8, 60% identity and bit value greater than 100.

This set comprise of Essential protein which are crucial for the survival of the pathogen. Thus, 624 proteins were identified as Essential protein in case of S. flexneri while 598,

641 and 610 in case of *S. dysentriae, S soneii* and *S. boydii* respectively. Metabolic chokepoint analysis is done using BioCyc database to identify Essential protein using the criteria: Exclude reactions found in human, exclude reaction catalyzed by more than one enzyme and limit to the reaction found in multiple pathways. This results in identification of the targets which are not only essential but also unique to the pathogen pathway. In case of *S. flexneri* 58 targets were identified as metabolic chokepoints while 4, 55 and 57 in case of *S. Dysentriae, S soneii* and *S. boydii* respectively. Out of these 58 targets 34 are non-homologous to human in case of S. flexneri while 31, 34 and 4 in case of *S. boydii*, *S. soneii* and *S. dysentriae* respectively.

Merging these datasets yields a new dataset which constitute the final list of the Essential targets. Thus, the final Essential datasets contains 644 in case of *S. flexneri* while *S. dysentriae*, *S soneii* and *S. boydii* constitutes 599, 660 and 625 respectively.

Analyzing the result obtained from BioCyc for each of these species gain an insight into a list of specific proteins which are found in multiple pathways hence targeting any of these proteins may lead to non-functioning of various pathways. These enzymes are crucial for survival of pathogen as these proteins are integral part of the reaction which produces or consumes unique substrates, uniquely found in pathogen and involved in multiple pathways. A detail list of each of these pathways as well as the function associated with it is given in table 1. Some of the pathways obtained from the BioCyc result are discussed below:

Acid resistance systems help various commensal pathogens including Shigella to survive in the acidic environment of stomach and intestine of the host by neutralizing intracellular acidic fermentation products. Three targets are identified to be the part of acid resistance system in case of S. boydii while one each in case S. flexneri and *S. soneii* [22][23][24].

Glyoxalate cycle enzymes may be considered as valuable targets for the development of antimicrobial drugs because this pathway is absent in mammalian host. Since Mammals require sugar as their carbon source; therefore, Glyoxalate cycle is not needed and thus drug targeted to Glyoxalate enzymes would be more specific and less toxic than the drugs that inhibits more conserved processes. Moreover, they might be effective against a wide range of pathogens. Two targets are identified to be the part of Glyoxalate cycle in case of *S. boydii* while 1 each in case of *S. flexneri* and *S. soneii* [25][26]. Lipopolysaccharides (LPS) and peptidoglycans are the main components of the outer cell wall of gram negative bacteria and play a crucial role in pathogenesis. The enzymes involved in the pathways are unique to bacteria and does not exist in mammalian host. Four enzymes are predicted to be drug targets as they are involved in Lipopolysaccharides (LPS) and peptidoglycans biosynthesis in case of *S. soneii* while three each in case of *S. flexneri* and *S. boydii*. Reactive oxygen species is primarily used by host defense system against pathogens as it damages numerous biomolecules within cell and thus disrupts many cellular processes [27]. Pathogen cope up with this stress by activating principal oxidative stress defense mechanisms which includes enzymes that either repair damage caused by ROS or neutralize the damaging ROS [28][29][30]. One each in case of *S. flexnerri* and *S. dysentriae* are associated with SoxR gene which is activated against oxidative stress of the host. Thus, these enzymes can be potential targets as targeting these enzymes can disrupt the defense mechanism of pathogen against host.

Further analysis of shortlisted essential targets revealed sets of proteins which may be used as targets depending on different filter applied during the analysis. Druggability analysis revealed that 76 proteins of *S. Flexneri* have similar domain as the approved drug targets while 71, 74 and 77 proteins in case of *S. dysentriae, S. boydii* and *S. Soneii*. Since these identified drug targets have identical domain with the proteins which are already approved as drug targets thus, targeting these proteins may results in better selection of target which passes the filter of drug designing process. Broad spectrum drug target analysis revealed a set of proteins whose domains are found in at least 3 of the Shigella species including *S. Flexneri* and must be found in at least 15 of the other pathogen enlisted in DEG. Thus, 150 proteins are identified to be broad spectrum drug targets in case *S. flexneri* while 148, 151 and 148 in case of *S. dysentriae, S. soneii* and *S. boydii*. Due to increasing cases of co-infection it is highly recommended to use the drugs which may be effective on a broad range of pathogens. Further subcellular localization prediction through psortb [31] is done to identify those proteins which are either extracellular or periplasmic. In case of *S. flexneri* and *S. dysentriae* 7 and 9 proteins are found to be extracellular while 14 and 16 are periplasmic respectively while in case of *S. soneii* 7 and 17 while in case of *S. boydii* 8 and 13 proteins are extracellular and periplasmic respectively. These proteins can be potential vaccine targets.

Lastly, the Essential protein is subjected to BlastP with human gutflora proteome to find those targets which are non-homologous to human gut flora. This step is done to retrieve the list of drug targets non-homologous to human gut flora so that the drug administered against the targets do not bind unintentionally to the proteins of the gut flora. Inhibition of gut flora may result in adverse effect to the host by promoting pathogenic colonization of the gut [32]. Also unintentional inhibition of gut flora protein causes idiosyncratic drug toxicity and reduced bio-availability of the drug [32][33][34]. Thus 16 of the targets were identified to be non-homologous to human gut flora in case of *S. flexneri* while 7, 6 and 18 in case of *S. boydii*, *S. dysentriae* and *S. soneii*.

Thus, in order to establish the reliability of data further in vivo or in vitro assessment of the predicted proteins is needed. The complete list of the proteins identified via this in silico approach is available as supplementary data.

4. CONCLUSION

With the availability of completely sequenced genomic and proteomic data in the public domain, the Insilico analysis tools have reduced the cost and time factor and also the vast data present in different databases can be used in meaningful way. With the availability of DEG, the identification and exploration of most relevant and essential targets in the pathogen is very convenient and efficient than the conventional method. The present study not only helped to identify a list of targets essential for the survival of pathogen but also it includes the targets which are less or non-toxic to host and organisms of gut flora. Also those targets which are similar to various pathogen involved in pathogenesis of disease may provide us valuable information regarding the targets which can be chosen as a broad spectrum targets to get rid of various pathogens especially in the case of co-infection where treatment is difficult. Since the number of essential genes (Exclude reactions found in human, exclude reaction catalyzed by more than one enzyme and limit to the reaction found in multiple pathways) are small in each of these species, so it can be accessed for further analysis and their role in survival of the bacteria can be verified. Also the lists of

targets which are extracellular or periplasmic in origin can be tested in vitro as potential vaccines. Targets which are similar to already approved targets are most prominent targets for drug discovery as they have potential to sustain in the drug delivery process. Virtual screening against these novel targets might prove new insight in the discovery of novel therapeutic compounds against Shigella Species.

5. REFERENCES

[1] Kotloff, K.L., J.P. Winickoff, B. Ivanoff, J.D. Clemens, D.L. Swerdlow, P.J. Sansonetti, G.K. Adak, M.M. Levine, "Global burden of Shigella infections: implications for vaccine development and implementation of control strategies," Bulletin of the World Health Organization. vol. 77, pp. 651–666, 1999.

[2] Sur D., T. Ramamurthy, J. Deen, S.K. Bhattacharya, "Shigellosis: challenges & management issues," Indian J Med Res. vol. 120, pp. 454–462, 2004.

[3] Legros D., "Guidelines for the control of shigellosis, including epidemics due to Shigella dysenteriae type 1", Geneva: World Health Organization. 2005.

[4] Mehata S., G.C. Duan, "Molecular mechanism of multi-drug resistance in Shigella isolates from rural China," Nepal Med Coll J. vol. 13, pp. 27-29, 2011.

[5] Watanabe T., "Infective heredity of multiple drug resistance in bacteria," Bacteriol Rev. vol. 27, pp. 87–115, 1963.

[6] Sack R.B., M. Rahman, M. Yunus, E.H. Khan, "Antimicrobial resistance in organisms causing diarrheal disease," Clin Infect Dis. vol. 24, pp. 102–105, 1997.

[7] Pazhani G.P., S.K. Niyogi, A.K. Singh, B. Sen, N. Taneja, M. Kundu, S. Yamasaki, R. Thandavarayan, "Molecular characterization of multidrug-resistant Shigella species isolated from epidemic and endemic cases of shigellosis in India," J Med Microbiol. vol. 57, pp.856–863, 2008.

[8] Ahamed J., M. Kundu, "Molecular characterization of the SHV-11 Beta lactamase of Shigella dysenteriae," Antimicrob Agents Chemother. vol. 43, pp. 2081–2083, 1999.

[9] Varghese S.R., A. Aggarwal, "Extended spectrum beta-lactamase production in Shigella isolates - A matter of concern." Indian J Med Microbiol. vol. 29, pp. 76–78, 2011.

[10] Karthik R., Y. Kalidas, C. Nagasuma, "targetTB: A target identification pipeline for Mycobacterium tuberculosis through an interactome, reactome and genome-scale structural analysis." BMC Systems Biology. vol. 2, pp.109, 2008.

[11] Gasteiger E., A. Gattiker, C. Hoogland, I. Ivanyi, R.D. Appel, A. Bairoch, "ExPASy: the proteomics server for in-depth protein knowledge and analysis," Nucleic Acids Res. vol. 31, pp.3784-3788, 2003.

[12] Geer L.Y., A.M. Bauer, R.C. Geer, et al., "The NCBI BioSystems database." Nucleic Acids Res. vol. 38, pp. 492-496, 2010.

[13] Weizhong Li. & G. Adam, (2006) "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences," Bioinformatics. vol. 22, pp. 1658-1659, 2006.

[14] Altschul S.F., T.L. Madden, A.A. Schäffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Research. vol. 25, pp. 3389-3402, 1997.

[15] Ren Z., Y.O. Hong, T.Z. Chun, "DEG, a Database of Essential Genes," Nucleic Acids Research. vol. 32, pp. 271-272, 2004.

[16] Iwei Y., H. Theodor, T. Sophia, D.K. Peter, B. Russ, "Computational analysis of Plasmodium falciparum metabolism: Organizing genomic information to facilitate drug discovery," Genome Research. vol. 14, pp. 917, 2004.

[17] Caspi R.,T. Altman, J.M. Dale, K. Dreher, C.A. Fulcher, F. Gilham, P., Kaipa, A.S. Karthikeyan, A. Kothari, M. Krummenacker, "The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases," Nucleic Acids Research. vol. 38, pp. 473-479, 2010.

[18] Knox C., V. Law, T. Jewison, P. Liu, S. Ly, A. Frolkis, A. Pon, K. Banco, C. Mak, V. Neveu, Y. Djoumbou, R. Eisner, A.C. Guo, D.S. Wishart, "Drug Bank 3.0: a comprehensive resource for 'omics' research on drugs." Nucleic Acids Res. vol. 39, pp.1035-1041, 2011.

[19] Marchler B.A., C. Zheng, F. Chitsaz, M.K. Derbyshire, L.Y. Geer, R.C. Geer, et. al., "CDD: conserved domains and protein three-dimensional structure." Nucleic Acids Res. vol. 41, pp. 348-352, 2012.

[20] Telley I.A. I. Gáspár, A. Ephrussi, T. Surrey, "Aster migration determines the length scale of nuclear separation in the Drosophila syncytial embryo." Journal of Cell Biology. vol. 197, pp. 887-895, 2012.

[21] Beifang Niu., Fu. Limin, S. Shulei, Li. Weizhong, "Artificial and natural duplicates in Pyrosequencing read of metagenomic data." BMC Bioinformatics. vol. 11, pp.187, 2010.

[22] Bearson S., B. Bearson, J.W. Foster, "Acid stress responses in enterobacteria." FEMS Microbiol Lett. vol.147, pp. 173–180, 1997.

[23] Foster J.W., "Escherichia coli acid resistance: tales of an amateur acidophile." Nat Rev Microbiol. vol. 2, pp. 898–907, 2004.

[24] Rembert P., C.R. Fisher, S. Moo-Jin, S.T. Huang, P. Parmar, S.M. Payne, "Analysis of the proteome of intracellular Shigella flexneri reveals pathways important for intracellular growth". Infect Immune. vol. 81, pp. 4635-4648, 2013.

[25] Sharma, V., S. Sharma, K.H. Bentrup, J.D. McKinney, D.G.W. Russell, R. Jacobs, J.C. Sacchettini, "Structure of isocitrate lyase, a persistence factor of Mycobacterium tuberculosis." Nat. Struct. Biol. vol. 7, pp. 663-668, 2000.

[26] Michael, C. L. and R. Gerald, "Fink: Life and Death in a Macrophage: Role of the Glyoxylate Cycle in Virulence." Eukaryot Cell. vol. 1, pp. 657–662, 2002.

[27] Imlay J.A., "Pathways of oxidative damage." Annu Rev Microbiol. vol. 57, pp. 395–418, 2003.

[28] Imlay J.A.,"Cellular defenses against superoxide and hydrogen peroxide". Annu Rev Biochem. vol. 77, pp.755–776, 2008.

[29] Johnson J.R., C. Clabots, H. Rosen, "Effect of inactivation of the global oxidative stress regulator oxyR on the colonization ability of Escherichia coli O1:K1:H7 in a mouse model of ascending urinary tract infection." Infect Immun. vol. 74, pp. 461–468, 2006.

[30] Aaron D., E.S. Akamol, R.J. Laura, "The Role of OxyR and SoxRS in oxidative stress survival in Shigella flexneri." Microbiol Res. vol. 167, pp. 238–245, 2012.

[31] Yu, N.Y., J.R. Wagner, M.R. Laird, G. Melli, S. Rey, R. Lo, P. Dao, S.C. Sahinalp, M. Ester, L.J. Foster, F.S.L. Brinkman, "PSORTb 3.0: Improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes," Bioinformatics. vol. 26, pp. 1608-1615, 2010.

[32] Levy J., "The effects of antibiotic use on gastrointestinal function". Am J Gastroenterol. vol. 95, pp. 8-10, 2000.

[33] Nicholson J.K. and I.D. Wilson, "Understanding 'global' systems biology: metabonomics and the continuum of metabolism." Nat Rev Drug Discov. vol. 2, pp. 668-676, 2003.

[34] Nicholson J.K., E. Holmes, I.D. Wilson, "Gut microorganisms, mammalian metabolism and personalized health care." Nat Rev Microbiol. vol. 3, pp. 431-438, 2005.
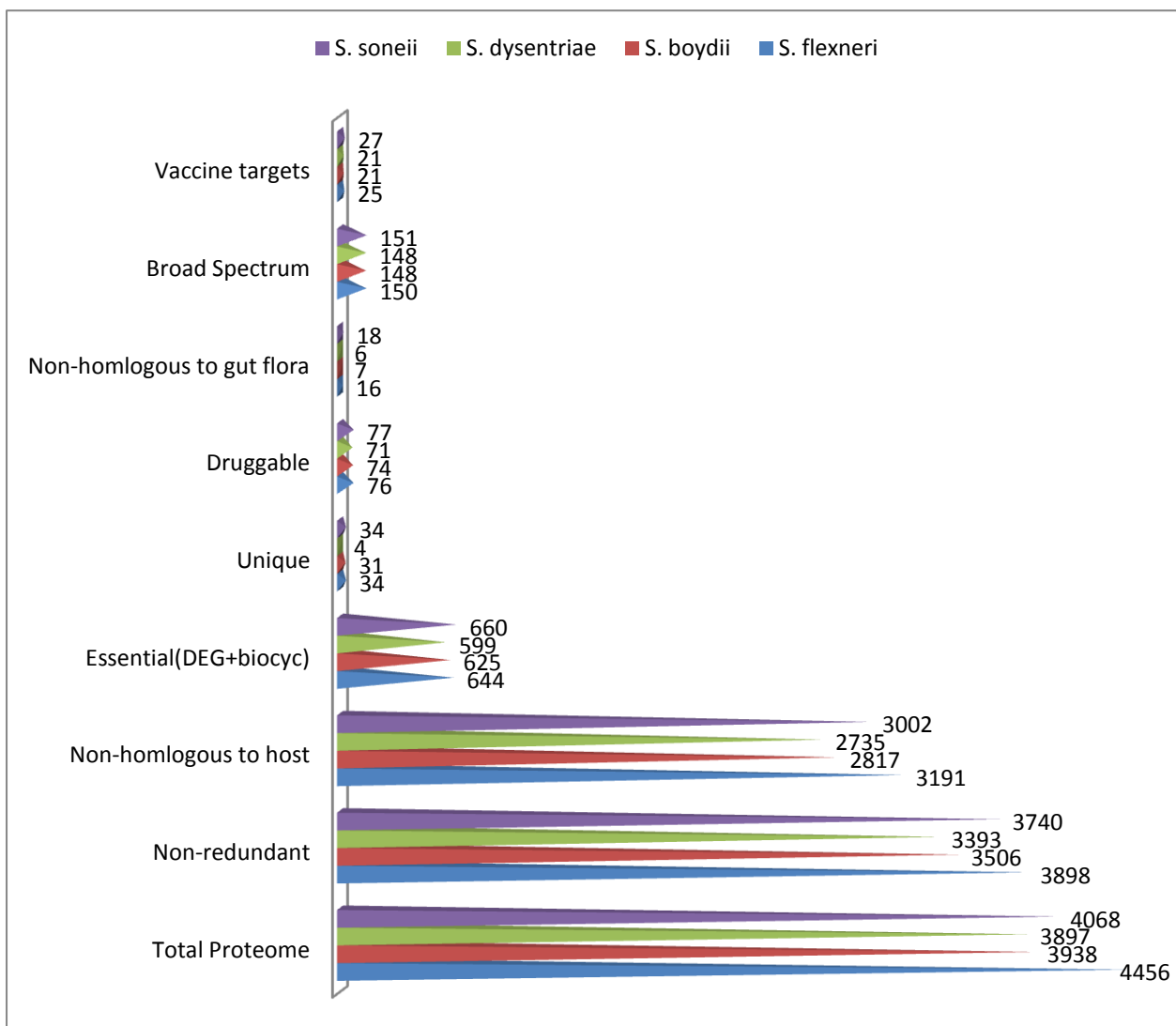
Fig 1: Proteome analysis graph of Shigella dysenteriae

Table 1. Detail list of pathways and functions

Analysis	Coverage				Pass				Fail				Computation
	SF	SB	SD	SS	SF	SB	SD	SS	SF	SB	SD	SS	
A) Initial pre-processing													
1) Removal of non-aa character	4456	3938	3897	4068	4455	3936	3895	4066	1	2	2	2	In house Perl Script
2) Removal of seq <50	4455	3936	3895	4066	4319	3904	3856	4030	136	32	39	210	In house Perl Script
B) Non- redundant sequence dataset	4319	3904	3856	4030	3898	3506	3393	3740	421	398	463	290	CD-hit
C) Non-homologous to host	3898	3506	3393	3740	3191	2817	2735	3002	707	689	658	738	BLAST against human proteome
D) Retrieval of sequences essential to survival of pathogen													
1) Homologous to DEG(A)	3191	2817	2735	3002	624	610	598	641	2567	2207	2137	2361	BLAST against DEG
2) Metabolic Choke point analysis (B)	3191	2817	2735	3002	34	31	4	34	3157	2786	2731	2968	Metabolic choke point analysis using the criteria Uniquely found in pathogen pathway, catalyzed by single enzyme and involved in multi pathways From BioCyc server
3) Total Essential Protein (A&B)	3191	2817	2735	3002	644	625	599	660	2547	2192	2136	2342	
E) Further Analysis of the target:													
1) Unique Pathway	644	625	599	660	34	31	4	34	610	594	595	626	From BioCyc server
2) Druggability Analysis	644	625	599	660	76	74	71	77	568	551	528	583	NCBI CDD
3) Broad Spectrum drug targets	644	625	599	660	150	148	148	151	494	477	451	509	NCBI CDD
4) Non-homologous to gut flora	644	625	599	660	16	7	6	18	628	618	593	642	BLAST against human gut flora
5) Vaccine targets	644	625	599	660	25	21	21	27	619	604	578	633	Subcellular localization prediction using Psortb