PREDICTION OF ESSENTIAL GENES IN THE SPECIES OF *Shigella*

Dissertation submitted in partial fulfillment of the requirement for the degree of

BACHELOR OF TECHNOLOGY

IN

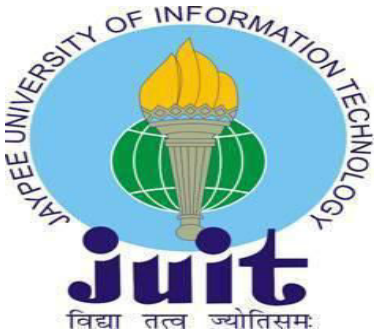
BIOINFORMATICS

By

Divya Chouhan 121504
Surbhi Sharma 121505
Srishti Jain 121511

UNDER THE GUIDANCE OF

Dr. Jayshree Ramna



DEPARTMENT OF BIOTECHNOLGY & BIOINFORMATICS

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY, WAKNAGHAT

June 2016

PREDICTION OF ESSENTIAL GENES IN THE SPECIES OF *Shigella*

Dissertation submitted in partial fulfillment of the requirement for the degree of

BACHELOR OF TECHNOLOGY

IN

BIOINFORMATICS

By

Divya Chouhan	**121504**
Surbhi Sharma	**121505**
Srishti Jain	**121511**

UNDER THE GUIDANCE OF

Dr. Jayshree Ramna

DEPARTMENT OF BIOTECHNOLGY & BIOINFORMATICS

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY, WAKNAGHAT

June 2016

TABLE OF CONTENTS

DECLARATION BY THE SCHOLAR.....**i**
CERTIFICATE.....**ii**
ACKNOWLEDGEMENT.....**iii**
LIST OF ACRONYMS.....**vi**
LIST OF FIGURES.....**v**
LIST OF TABLES.....**vi**
ABSTRACT.....**vii**
CHAPTER 1
INTRODUCTION.....**1**
CHAPTER-2
DISEASE PROCESS.....**4**
CHAPTER-3
CAUSES.....**12**
CHAPTER 4
METHOD USED.....**13**
CHAPTER 5
MACHINE LEARNING APPROACH.....**20**
CHAPTER 6
CONCLUSION & FUTURE WORK.....**28**
REFERENCES.....**29**

CERTIFICATE

This is to certify that the work tittled " **PREDICTION OF ESSENTIAL GENES IN THE SPECIES OF *Shigella*** " submitted by **Divya Chouhan(121504), Surbhi Sharma(121505) and Srishti Jain(121511)** in partial fulfilment for the award of degree of B.Tech Bioinformatics of **Jaypee University Of Information Technology, Waknaghat** has been carried out under my supervision. This work has not been submitted partially or wholly to any other university or institute for the award of this or any other dergree or diploma.

Signature of Supervisor …….

Name of Supervisor Dr. Jayshree Ramana

Designation Assistant Professor, JUIT

Date ……..

.

DECLARATION

We hereby declare that the work reported in the B-Tech thesis entitled **"PREDICTION OF ESSENTIAL GENES IN THE SPECIES OF *Shigella*"** submitted at **Jaypee University Of Information Technology** is an authentic record of my work carried out under the supervision of Dr. Jayshree Ramana**.** I have not submitted this work elsewhere for any other degree or diploma

…….

Divya Chouhan

121504

……..

Surbhi Sharma

121505

……..

Srishti Jain

121511

Department of Biotechnology & Bioinformatics

Jaypee University of Information Technology, Waknaghat, India

Date : ……..

ACKNOWLEDGEMENT

This dissertation would not have been possible without the guidance and the help of several individuals who in one way or another contributed and extended their valuable assistance in the preparation and completion of this study.

First and foremost , my utmost gratitude to **Dr. Jayshree Ramana,** Senior Lecturer of the Department of Biotechnology and Bioinformatics (JUIT) , for the unfailing support as our project adviser, and for his patience and steadfast encouragement to complete this study. He has been an inspiration as we hurdled through all the obstacles in the completion of this research work.

Apart from these countless events, countless people and several incidents have made contribution to this project that is indescribable. We again express our gratitude to them. We indebt to all those who provided reviews and suggestions for improving the results and topics covered in our project, extend our apologies to any one whom we have failed to recognize in this effort of ours.

LIST OF ACRONYMS

1. DNA- Deoxyribonucleic acid
2. RNA- Ribonucleic acid
3. PMN- Polymorphonuclear
4. HLA- Histocompatibility antigen.
5. CNS- Central nervous system
6. SDA-Stepwise discriminate analysis.
7. DEG- Database of essential genes
8. PMC- Pubmed central
9. HIV- Human immunodeficiency virus
10. STxs- Shiga toxin
11. CPS-chronic pain syndrome
12. OMP- outer membrane protein
13. SVM- support vector machine
14. FWM-Naïve Bayes model
15. WNB- Weighted naïve bayes
16. LRM-Logistic regression model
17. LIB-Laboratory of immunopathogenesis and bioinformatics.
18. GO-Gene ontology

LIST OF FIGURES

Fig.2.1: Disease Process Cycle……4

Fig.4.1: Approach Performa……..……17

Fig.4.2: Metabolic Network of interested genes ……..…….……18

Fig. 4.3: Input genes ……..……19

Fig.4.4: core genome of *S.flexneri* and *S.sonnei*……...……..…20

Fig.5.1: DAVID TOOL……...…23

Fig.5.2: Functional Annotation……...……...……24

Fig. 5.3: Annotation Summary Results……..……25

Fig. 5.4: Functional Categories……...……...……25

Fig. 5.5: Pathways……..……26

Fig. 5.6: Functional Annotation Chart……...……26

Fig. 5.7: WEKA Server……..……26

Fig.5.8: Classification of Samples……..……...……27

Fig. 5.9: Clustering…….……...……28

Fig. 5.10: Association……...……28

LIST OF TABLES

Table 5.1 Extracted Genes……...……..21

Table 5.2 Essential Genes……..……...28

ABSTRACT

Bacteria and their derivatives i.e. gram positive and gram negative have a large and expanded worldwide market. *Shigella* has the potential to become a model organism for bacteria world since its genome has been sequenced and genes coding for various trades has been predicted. In the current study we made efforts to reannotate the shigella strain i.e. *S.flexneri* and *S.Sonnei* by finding essential genes in their genome through machine learning and gene ontology based approach .In machine learning we perform classification and clustering of sample genes after processed them, on the other hand; gene ontology study was perform by finding molecular function ,biological process and cellular component of those gene . Total 24 genes are found to be involved in different metabolic and physiological process of *shigella* species which are confirmed through different statistical measures. These 24 genes found to be the area of interest regarding functional characterization.

CHAPTER 1

INTRODUCTION

WHAT ARE ESSENTIAL GENES?

Essential genes are those genes of an organism that are thought to be critical for its survival. However, being essential is highly dependent on the circumstances in which an organism lives.

For instance, a gene required to digest starch is only essential if starch is the only source of energy.

These essential genes encode proteins to maintain central metabolism, replicate DNA, translate genes into proteins, maintain a basic cellular structure, and mediate transport processes into and out of the cell. Most genes are not essential but convey selective advantages and increased fitness. [1]

The most common class of identified genes is composed of essential genes. The categories of essential genes include a broad range of lethal mutant phenotypes that block survival or reproduction. Lethal mutations range in developmental blocking stages from egg to larval, sterile, and maternal-effect. Examination of the lethal phenotypes can provide information about the function of the gene product. In one case, suppression of a lethal phenotype was used as a selection for genetic duplications.

Some categories of mutations cannot be readily defined as essential or nonessential. Examples of this are the daf mutations in which dauer-constitutive mutants cannot mature and are lethal, whereas dauer-defective mutants go through the normal life cycle and could only be considered lethal under conditions that require dauer formation.

In screens for visible mutations, lethals are generally not recovered, and in screens for lethals, visible are usually missed. In some cases, the original phenotype has been shown to be a typical. Examples include bli-4 for which only the original allele results in blistered cuticle , and all 12 other alleles result in late embryonic arrest ; unc-70 , for which the original allele results in an uncoordinated phenotype , whereas at least seven other alleles are lethal ; and rol-3, for which

two alleles result in rolling , one allele is a temperature-sensitive lethal and there are 11 non conditional lethal alleles . In contrast, unc-60 is an example of a gene for which most of the alleles give a visible phenotype. However, one allele, which proved to be a small internal deletion, is lethal. unc-60 encodes two alternatively spliced products, and it appears that both must be deleted in order for the essential function to be revealed .

Are essential genes mostly “housekeeping” genes, i.e., are they mainly required in general processes necessary for cell operation such as intermediary metabolism or are they mainly important in specific developmental processes such as determination, differentiation, or morphogenesis? Many essential genes function in both embryogenesis and later stages, and numerous genes are expressed throughout embryogenesis. These results led to the proposal that most zygotic essential genes have “housekeeping” functions. Bucher and Greenwald tested this by designing an elegant mosaic screen that circumvented the difficulties of analyzing the phenotypes of early blocking lethal mutations. These authors screened for zygotic lethal mutants in an unc-36 glp-1 ncl-1 triple-mutant genetic background using the duplication qDp3, which covers the major gene cluster on chromosome III, including the markers.

Shigella

Shigella is a group of Gram-negative, facultative intracellular pathogens. Recognized as the etiologic agents of bacillary dysentery or shigellosis in the 1890s, *Shigella* was adopted as a genus in the 1950s and subgrouped into four species: S.*dysenteriae, S.flexneri, S.boydii S.sonnei* (also designated as serogroups A to D).

Shigella grows only in the intestinal tract of humans. It's transmitted by the fecal-oral route. Fliers, fingers, and food are the usual vehicles. But because *Shigella* cells survive for a long time in contaminated water or on fomites, they transmit it too. People who live in crowded conditions where cleanliness is difficult are particularly likely to contract shigellosis. [2-4]

Children are far more likely than adults to get shigellosis. Those under 5 year old account for about half the reported cases, because they are too young to follow good hygiene habits and they are more susceptible to *Shigella* infection.

Many fecal-oral infections, including cholera and typhiod fever, have been nearly eradicated from industrialized countries, but not shigellosis. Shigellosis is difficult to eradicate partly because it is so infectious. A person must ingest thousands to millions of bacterial cells to contract typhoid fever or cholera, but only 200 cells are sufficient to cause shigellosis.

It is well-established that the virulence plasmid (VP) carries the primary virulence genes that enable the invasiveness of the bacteria in the colon and the rectum and the induction of apoptosis to resident macrophages and dentritic cells, leading to inflammatory infection.

CHAPTER 2

DISEASE PROCESS

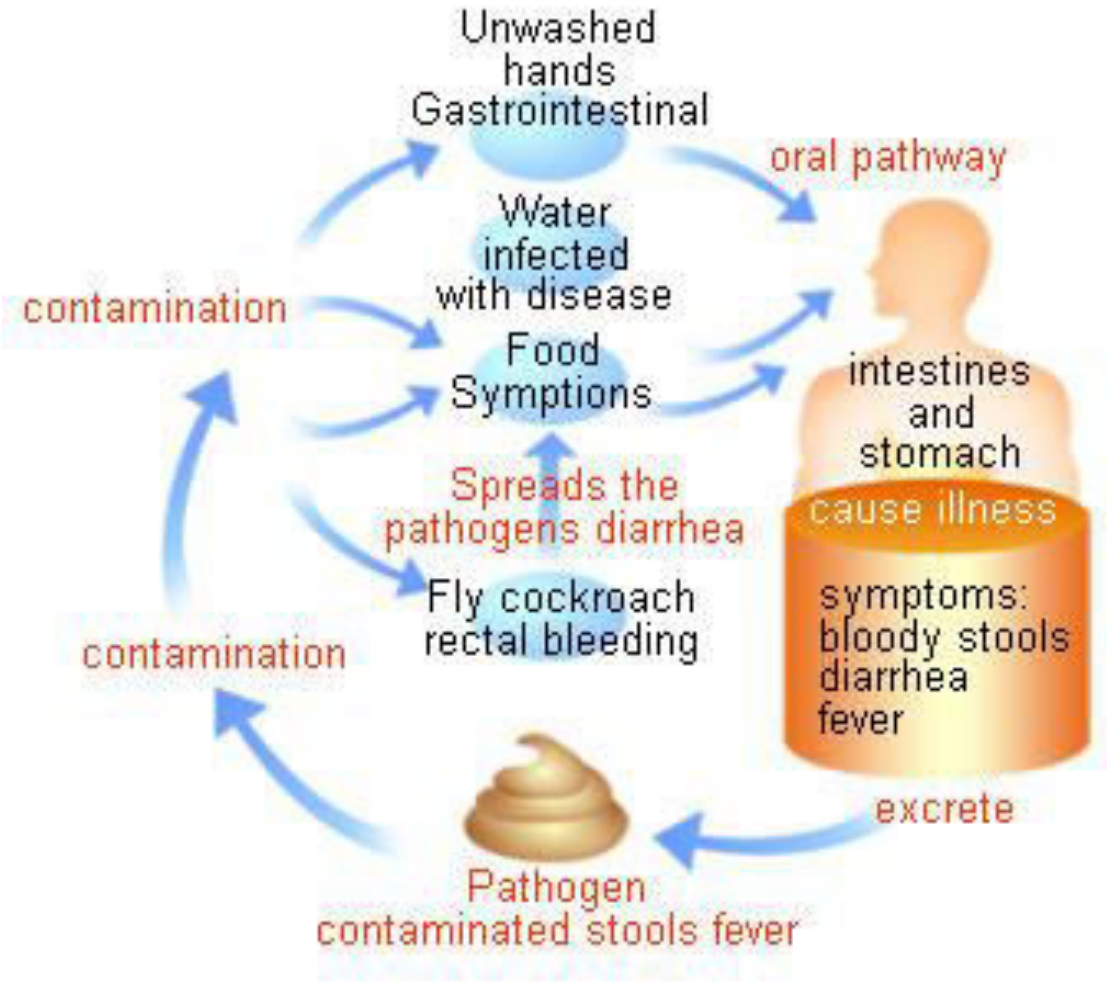


Fig.2.1: Disease Process Cycle

Shigellosis is spread by means of fecal-oral transmission. Other modes of transmission include ingestion of contaminated food or water (untreated wading pools, interactive water fountain), contact with a contaminated inanimate object, and certain mode of sexual contact. Vectors like the housefly can spread the disease by physically transporting infected feces.

The infectivity dose (ID) is extremely low. As few as 10 *S.dysenteriae* bacilli can cause clinical disease, whereas 100-200 bacilli are needed for *S.sonnei* or *S. flexneri* infection. The reasons for this low-dose response are not completely clear. One possible explanation is that virulent

Shigella can withstand the low pH of gastric juice. Most isolates of *Shigella* survive acidic treatment at pH 2.5 for at least 2 h.

The incubation period varies from 12 hours to 7 days but is typically 2-4 days; the incubation period is inversely proportional to the load of ingested bacteria. The disease is communicable as long as an infected person excretes the organism in the stool, which can extend as long as 4 weeks from the onset of illness. Bacterial shedding usually ceases within 4 weeks of the onset of illness; rarely, it can persist for months. Appropriate antimicrobial treatment can reduce the duration of carriage to a few days. [5-7]

VIRULENCE

Virulence in *Shigella* species involves both chromosomal-coded and plasmid-coded genes. Virulent *Shigella* strains produce disease after invading the intestinal mucosa; the organism only rarely penetrates beyond the mucosa.

The characteristic virulence trait is encoded on a large (220 kb) plasmid responsible for synthesis of polypeptides that cause cytotoxicity. *Shigella* that lose the virulence plasmid are no longer pathogenic. Escherichia coli that harbor this plasmid clinically behave as *Shigella* bacteria.

Siderophores, a group of plasmid-coded genes, control the acquisition of iron from host cells from its protein-bound state. In the extra intestinal phase of infection by gram-negative bacteria, iron becomes one of the major factors that limit further growth. This limitation occurs because most of the iron in human body is sequestered in hemoproteins (i.e., hemoglobin, myoglobin) or iron-chelating proteins involved in iron transport (transferrin and lactoferrin). These siderophores are under the control of plasmids and are tightly regulated by genes such that, under low iron conditions, expression of the siderophore system is high.

Regulatory genes control expression of virulence genes. Shiga toxin (Stx) is not essential for virulence of S dysenteriae type 1 but contributes to the severity of dysentery. Both plasmid-encoded virulence traits and chromosome-encoded factors are essential for full virulence of shigellae.[7-9]

Regarding chromosomally encoded enterotoxin, many pathogenic features of *Shigella* infection are due to the production of potent cytotoxins known as Stx, a potent protein synthesis–inhibiting exotoxin. *Shigella* strains produce distinct enterotoxins. These are a family of cytotoxins that contain 2 major immunologically non–cross-reactive groups called Stx1 and Stx2.

These toxins are lethal to animals; enterotoxic to ligated rabbit intestinal segments; and cytotoxic for vero, HeLa, and some selected endothelial cells (human renal vascular endothelial cells) manifesting as diarrhea, dysentery, and hemolytic-uremic syndrome (HUS).Stx1 is synthesized in significant amount by *S.dysenteriae* serotype 1 and *S.flexneri* 2a and E.coli (*Shigella* toxin–producing E.coli [ShET]).

Stx1 and Stx2 are both encoded by a bacteriophage inserted into the chromosome. Stx1 increases inflammatory cytokine production by human macrophages, which, in turn, leads to a burst of interleukin (IL)-8. This could be relevant in recruiting neutrophils to the lamina propria of the intestine in hemorrhagic colitis and accounts for elevated levels of IL-8 in serum of patients with diarrhea-associated HUS. [10-12]

Stxs have 2 subunits. Stx is transported into nucleoli. Stx nucleolar movement is carrier-dependent and energy-dependent. Subunit A is a 32-kD polypeptide that, when digested by trypsin, generates A1 with a 28-kD fragment and another small fragment, A2, which is 4 kD. A1 fraction acts like N -glycosidase; it removes single adenine residue from 28S rRNA of ribosome and inhibits protein synthesis. [13]

In summary, events that occur on exposure to *Shigella* toxin are as follows:

The B subunit of holotoxin binds to the Gb3 receptor on the cell surface of brush-border cells of the intestines.

The receptor-holotoxin complex is endocytosed.

The complex moves to Golgi apparatus and then to the endoplasmic reticulum.

The A1 subunit is released and it targets 28S RNA of the ribosome, inhibiting protein synthesis. Stxs may play a role in the progression of mucosal lesions after colonic cells are invaded, or they may induce vascular damage in the colonic mucosa. Stx adheres to small-intestine receptors and

blocks the absorption of electrolytes, glucose, and amino acids from intestinal lumen. The B subunit of Stx binds the host's cell glycolipid in the large intestine and in other cells, such as renal glomerular and tubular epithelia. The A1 domain internalized by means of receptor-mediated endocytosis and causes irreversible inactivation of the 60S ribosomal subunit, inhibiting protein synthesis and causing cell death, microvascular damage to the intestine, apoptosis in renal tubular epithelial cells, and hemorrhage (as blood and mucus in the stool).

Chromosomal genes control LPS antigens in cell walls. LPS plays an important role in resistance to nonspecific host defense encountered during tissue invasion. These genes help in invasion, multiplication, and resistance to phagocytosis by tissue macrophages.

A 3-kb plasmid that harbors information for the production of bacteriocin by *S.flexneri* strains has been described. [14-17]

INTESTINAL ADHERENCE FACTOR

Intestinal adherence factor favors colonization in vivo and in animal models. This is 97-kD outer-membrane protein (OMP) encoded by each gene on chromosomes. This codes for intimin protein, and an anti-intimin response is observed in children.

PATHOLOGY

The host response to primary infection is characterized by the induction of an acute inflammation, which is accompanied by PMN infiltration, resulting in massive destruction of the colonic mucosa. Apoptotic destruction of macrophages in subepithelial tissue allows survival of the invading *shigellae*, and inflammation facilitates further bacterial entry.

Gross pathology consists of mucosal edema, erythema, friability, superficial ulceration, and focal mucosal hemorrhage involving the rectosigmoid junction primarily.

Microscopic pathology consists of epithelial cell necrosis, goblet cell depletion, PMN infiltrates and mononuclear infiltrates in lamina propria, and crypt abscess formation.

Shigella bacteria invade the intestinal epithelium through M cells and proceed to spread from cell to cell, causing death and sloughing of contiguously invaded epithelial cells and inducing a potent inflammatory response resulting in the characteristic dysentery syndrome. In addition to

this series of pathogenic events, only *S dysenteriae* type 1 has the ability to elaborate the potent Shiga toxin that inhibits protein synthesis in eukaryotic cells and that may lead to extraintestinal complications, including hemolytic-uremic syndrome and death. [18]

EPIDEMIOLOGY

UNITED STATES

The reported incidence of *Shigella* infections in 2010 was 1,780, which is 3.8 cases per 100,000 population. Most cases are reported during summer months. S sonnei accounts for approximately 78% of all *Shigella* isolates in recent surveys from the Center for Disease Control and Prevention (CDC); *S.flexneri* and *S.boydii* account for most of the remainder. *S.flexneri* causes 18% of *Shigella* infections in the United States. *S.dysenteriae* is rare in the United States. The highest incidence per 100,000 population for shigellosis (27.77 cases) was among children younger than 5 years.

State public health laboratories reported 7.746 laboratory confirmed *Shigella* infections to the CDC in 2012. Of the 7,746 laboratory confirmed isolates, 687 were identified to species level. Distribution by species was similar to previous years, with *S sonnei* accounting for the largest percentage of infections (75%), followed by *S.flexneri* (12%), *S.boydii* (0.8%), and *S.dysenteriae* (0.3%).The reporting jurisdictions with the highest incidence rates were Nebraska (13.2 %), New Jersey (7.6%), and Minnesota (7.1%).

The overall incidence of *Shigella* infection in 2012 was 2.5 cases per 100,000 population, and the rate of HUS in pediatric patients younger than 15 years is 0.49 cases per 100,000 population. Compared with the previous 10 years (2002–2011), a larger portion of *Shigella* infections in 2012 were reported from January.

INTERNATIONAL

Worldwide, the incidence of shigellosis is estimated to be 164.7 million cases per year, of which 163.2 million were in developing countries, where 1.1 million deaths occurred. About 60% of all episodes and 61% of all deaths attributable to shigellosis involved children younger than 5 years. The incidence in developing countries may be 20 times greater than that in developed countries.

Although the relative importance of various serotypes is not known, an estimated 30% of these infections are caused by *S.dysenteriae.*

Case-fatality rates for *S.dysenteriae* infections may approach 30%. Patients with malnutrition are at increased risk of having complicated course. *Shigella* infection in malnourished children often causes a vicious cycle of further impaired nutrition, recurrent infection, and further growth retardation. [19-20]

MORTALITY/MORBIDITY

Although shigellosis-related mortality is rare in developed countries, *S.dysenteriae* infection is associated with substantial morbidity and mortality rates in the developing world. Case fatality is as high as 15% among patients with *S.dysenteriae* type 1 who require hospitalization; this rate is increased by delayed arrival and treatment with ineffective antibiotics. Infants, non-breastfed children, children recovering from measles, malnourished children, and adults older than 50 years have a more severe illness and a greater risk of death

The overall mortality rate in developed countries is less than 1%.In the Far East and Middle East, the mortality rates for *S.dysenteriae* infections may be as high as 20-25%.

Race: No racial predilection is known.

Sex: No sexual predilection is known.

Age: According to recent CDC reports, *Shigella* infection accounted for 28% of all the enteric bacterial infections. Children younger than 5 years had 7% of total reported cases, a rate indicating a disproportionate disease burden in this population.

Populations that are at high-risk for shigellosis include the following:

-Children in daycare centers (< 5 y) and their caregivers

-Persons in custodial institutions

-International travelers

-Homosexual men

-People living in crowded conditions with poor sanitary facilities and inadequate clean water supply (eg, refugee camps, shelters for displaced people)

-People with HIV infection.

Symptoms include the following:

-Sudden onset of severe abdominal cramping, high-grade fever, emesis, anorexia, and large-volume watery diarrhea. Seizures may be an early manifestation.

-Abdominal pain, tenesmus, urgency, fecal incontinence, and small-volume mucoid diarrhea with frank blood (fractional stools) may subsequently occur.

-Elevated temperatures (as high as 106 º F) are documented in approximately one third of cases, and a generally toxic appearance is noticed.

-Tachycardia and tachypnea may occur secondary to fever and dehydration. Depending on the degree of dehydration, dry mucous membranes, hypotension, prolonged capillary refill time, and poor skin turgor may be present.

-Abdominal tenderness is usually central and lower, although it may be generalized.

-Extra intestinal manifestations are as follows:

-CNS symptoms include severe headache, lethargy, meningismus, delirium, and convulsions lasting less than 15 minutes, especially with *S.dysenteriae*. Severe toxic encephalopathy is rare, but lethal complications occur when initial symptoms are followed by sensory obtundation, seizures, coma, and death in 6-48 hours. The pathogenesis of neurologic manifestations during shigellosis is unclear. However, data now clearly demonstrate that *Shigella* toxin is not responsible. [21-22]

-Septicemia is rare, except in malnourished children with *S.dysenteriae* infection. Septicemia is sometimes caused by other gram-negative organisms and is related to loss of mucosal integrity by *Shigella* infection. Profound dehydration and hypoglycemia is more common with *S.dysenteriae* infection.

-*Shigellasepsis* may be complicated with Disseminated Intravascular Oagulation (DIC), bronchopneumonia, and multiple organ failure in lethal cases.

-Arthritis, urethritis, conjunctivitis syndrome is commonly observed in adults carrying Human Leukocyte Antigen (HLA)-B27 histocompatibility antigen.

-Cholestatic hepatitis, if present, is usually mild.

-Myocarditis is identified with cardiogenic shock, arrhythmias, and heart block.

-Rectal prolapse, toxic megacolon, and intestinal obstruction may be present.

-Shigellosis in the first 6 months of life is rare probably due to presence of antibodies to both virulence plasmid-coded antigens and lipopolysaccharides in the breast milk. Shigellosis in the neonatal period results from mother-to-infant fecal-oral transmission during labor and delivery, usually from asymptomatic mothers.

-Symptoms usually begin on the third day of life.

-Septicemia and chronic diarrhea are common.

-Fever may be absent.

-Diarrhea is not usually bloody.

-Intestinal perforation and mortality are more common in this group than in older children.

-Shigellosis in patients with HIV infection is often a protracted, chronic, relapsing disease (even when treated with antibiotics). Bacteremia is rare, although it can occur in immune compromised or malnourished patients. [23]

PHYSICAL EXAMINATION

Physical examination during acute illness reveals a febrile ill-appearing child. Fever with a temperature as high as 39-40 ° C may be noted. The patient's hydration status should be carefully assessed. Especially note dryness of the oral mucosa, lack of tears, decreased urine output, and loss of skin turgor. Abdominal examination may reveal generalized mild-to-moderate tenderness with no guarding or rigidity. In a child who presents with febrile seizures, careful neurologic examination is mandatory to exclude meningitis. [24]

CHAPTER 3

CAUSES

-The primary mode of transmission of *Shigella* infection is fecal-oral contamination by the gram-negative aerobic bacilli.

-Contaminated food usually looks and smells normal. Food may become contaminated by infected food handlers who forget to wash their hands with soap after using the bathroom. Vegetables can become contaminated if they are harvested from a field with sewage in it.

-Outbreaks of shigellosis have also occurred among men who have sex with men.

-Travellers from developed to developing regions and soldiers serving under field conditions are also at an increased risk to develop shigellosis.

-Shigellosis can be caused by exposure to contaminated treated water and, more likely, from untreated recreational water. [24]

CHAPTER 4

METHOD USED

1) Comparative genomics

2) Machine learning

Two types of approaches are mainly used to predict and identify essential genes: experimental laboratory techniques and computational techniques. The former is randomly or systematically used to inactivate potential essential genes, and gene essentiality could be determined based on the living situation of the organism. General gene disruption strategies include single gene knockouts, conditional knockouts, RNA interference, and transposon mutagenesis. Unfortunately, experimental techniques have significant drawbacks, such as long durations and high costs. In addition, the spectrum of gene essentiality varies under different growth conditions.

Computational techniques have become popular over the past years for several reasons. First, known essential genes from dozens of microorganisms provide instructional and training materials. Second, the available genome sequences obtained by high-throughput sequencing provide unprecedented opportunities for investigating the minimal subset of genes in various organisms. Finally and most importantly, the development of bioinformatics tools improves our capability for exploring essential genes.

Several prediction models have been developed in silico to identify essential genes. Among these models, the simplest one is prediction of essential genes based on the known essentiality of homologous genes. Although these prediction models show high confidence levels, they still have two limitations: first, the conserved orthologs between species only account for a small portion of the genome and, second, the orthologs, especially in distantly related species, often show variations in gene regulations and functions, which lead to potential diversity in gene essentiality. To circumvent these limitations, feature-based models have been constructed to

distinguish essential genes from non-essential ones based on common or similar features among essential genes.

In previous models, feature selection was often based on significant correlations between gene essentiality and gene features or the significant distribution difference between essential and non-essential genes. A common disadvantage of such selection method, however, is that feature–feature interactions and strong correlations among features are ignored. Moreover, because of evolutionary divergence among species, the linkages between features and gene essentiality might have changed. For example, arguments on whether or not younger genes are less likely to be essential than older genes or whether or not duplicate genes are less likely to be essential than singletons , demonstrate that gene essentiality associations with origin time and number of duplications are diverse among different species.

Aside from feature selection, machine learning algorithms have also been introduced into feature-based classification models to identify essential genes in many studies, such as Naïve Bayes, decision tree, SVM.

In the present study, we first collected 16 features that were widely used in previous models, and demonstrated that the predictions exhibit at least two problems:

(1) Strong correlations among gene features.

(2) Different and even contrasting associations of gene features with gene essentiality among different species. We then presented a novel approach, the feature-based weighted FWM, which can address multicollinearity impacts among gene features and feature divergence between species. In the proposed model, prior information was collected to determine the weight of each feature by logistic regression and genetic algorithm. Afterward, essential genes in target organisms were predicted using a WNB classifier. We applied FWM to reciprocally predict essential genes between and within 21 species and compared its performance with those of other models including SVM, NBM, and LRM. Results showed that FWM can significantly improve the accuracy and robustness of essential gene prediction. Finally, using SDA, we demonstrated why FWM outperforms these other classifiers.

DEGG DATABASE

Essential genes are those indispensable for the survival of an organism, and therefore are considered a foundation of life. DEG hosts records of currently available essential genomic elements, such as protein-coding genes and non-coding RNAs, among bacteria, archaea and eukaryotes. Essential genes in a bacterium constitute a minimal genome, forming a set of functional modules, which play key roles in the emerging field, synthetic biology.

The definition of the minimal gene set needed to sustain a living cell is of considerable interest not only because it represents a fundamental question in biology, but also because it has much significance in practical use. For example, since most antibiotics target essential cellular processes, essential gene products of microbial cells are promising new targets for antibacterial drugs. One of the applications is the prediction of essential genes based on homologous sequence search against DEG. The functions encoded by essential genes are considered to be generally essential for all cells.

IMPORTANCE

- Bacterial dysentery due to *Shigella* species is a major cause of morbidity and mortality; 165 million cases occur annually worldwide, with 1 million associated deaths.

- In developed countries, most cases are transmitted by fecal-oral spread from people with symptomatic infection. Outbreaks in the United States occur predominantly in institutions such as day care centers or custodial institutions and less commonly by common source contamination of food or drinking water.

- In developing countries, both fecal-oral spread and contamination of common food and water supplies are important mechanisms of transmission.

- The annual number of *Shigella* episodes throughout the world was estimated to be 164.7 million, of which 163.2 million were in developing countries (with 1.1 million deaths) and 1.5 million in industrialized countries.

 The median percentages of isolates of *S. flexneri, S. sonnei, S. boydii*, and *S. dysenteriae* were, respectively, 60%, 15%, 6%, and 6% (30% of *S.dysenteriae* cases were type 1) in developing countries; and 16%, 77%, 2%, and 1% in industrialized countries.

 In developing countries, the predominant serotype of*S. flexneri* is 2a, followed by 1b, 3a, 4a, and 6. In industrialized countries, most isolates are *S. flexneri* 2a or other unspecified type 2 strains. Shigellosis, which continues to have an important global impact.

 Innovative strategies, including development of vaccines against the most common serotypes, could provide substantial benefits.

APPROACH USED

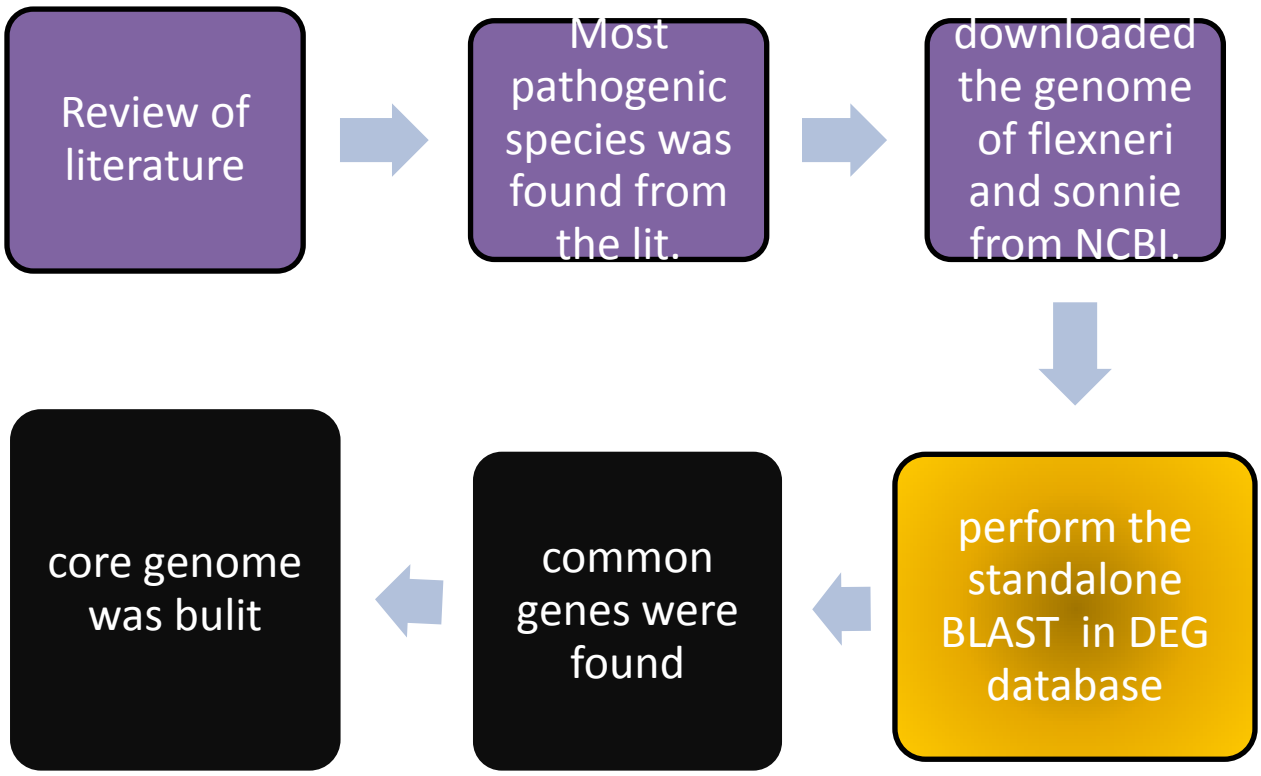


Fig.4.1: Approach Performa

RESULTS

-the common genes found are then searched for the common pathways in which they are involved.

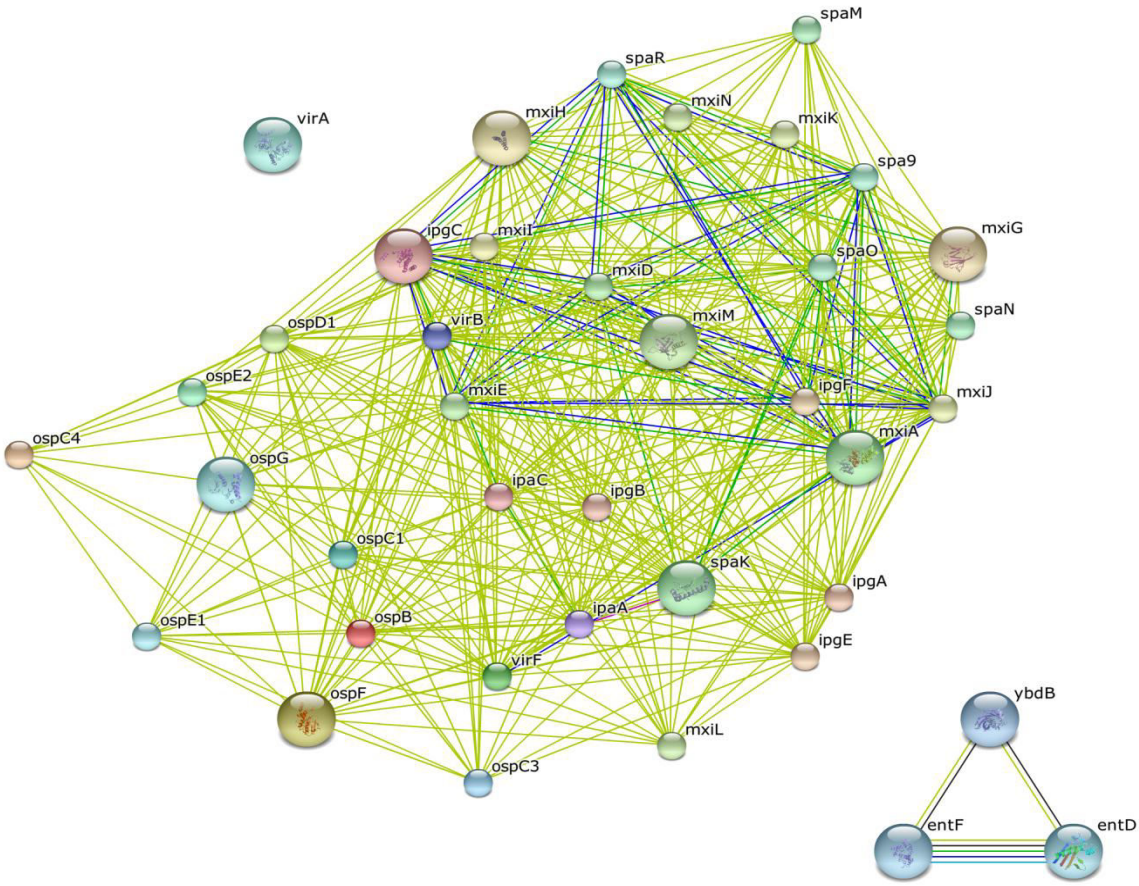


Fig.4.2: Metabolic Network of interested genes

Your Input:

ospB	OspB protein (288 aa)
ospC4	OspC4 (424 aa)
ospF	OspF; Catalyzes the removal of the phosphate group from the phosphothreonine in the mitogen-activated protein kinases such as MAPK2/ERK2, MAPK3/ERK1, MAPK8 and MAPK14 in an irreversible reaction, thus preventing the downstream phosphorylation of histone H3. This epigenetic modification results in inhibition of the transcription of a specific subset of proinflammatory genes, and ultimately to a reduced immune response against the invading pathogen. The diminished immune response enhances the bacterium's ability to disseminate and multiply within the host (239 aa)
ospD1	protein OspD1 (225 aa)
virF	transcriptional activator VirF; Primary regulator of plasmid-encoded virulence genes. Activates the transcription of icsA (virG) and of virB, which is an activator of the ipaABCD virulence regulon (262 aa)
ospE2	OspE2 (88 aa)
ospC1	protein OspC1 (470 aa)
ospC3	OspC3 (484 aa)
virB	transcriptional activator VirB; Transcription activator for the invasion antigens IpaB, IpaC and IpaD. VirB is itself regulated by VirF (309 aa)
ipaA	hypothetical protein; Rapidly associates with the first 265 amino acids of vinculin after bacteria-cell contact. This interaction is critical for efficient Shigella uptake. IpaA acts as a potent activator of vinculin and increase its ability to interact with F-actin. The complex IpaA-vinculin induces F-actin depolymerization along with the occasional formation of actin filament bundles (633 aa)
ipaC	IpaC, secreted by the Mxi-Spa secretion machinery, required for entry into epithelial cells; Forms a pore with IpaB, which is inserted into host cell membrane through the Mxi/Spa apparatus, during cell contact. This pore probably allows the translocation of IpaA. The C-terminus of IpaC could be involved in actin polymerization that leads to the cell extension formation and its N-terminus could be involved in the conversion of filopodial extensions into lamellipodial extension. Most extensions results from Cdc42 activation by IpaC whereas the lamellipodial extensions result from the Rac [...] (363 aa)
ipgC	IpgC, cytoplasmic chaperone for IpaB and IpaC; Assists the correct folding of nascent IpaB. Once it is bound to IpaB, it binds to IpaC and impedes their premature association that would lead to their degradation in the absence of IpcG (155 aa)
ipgB	protein IpgB1 (208 aa)
ipgA	chaperone IpgA; Molecular chaperone required for IcsB stabilization and secretion (129 aa)
ipgE	IpgE, cytoplasmic chaperone for IpgD; Molecular chaperone required for IpgD stabilization and secretion (120 aa)
ipgF	periplasmic protein IpgF (152 aa)
mxiG	MxiG protein; Involved in the secretion of the Ipa antigens. Involved in the intracellular dissemination of Shigella (371 aa)
mxiH	Mxi-Spa secretion machinery protein MxiH; Necessary for the secretion of IPA invasins (83 aa)
mxiI	Mxi-Spa secretion machinery protein MxiI; Necessary for the secretion of IPA invasins (97 aa)
mxiJ	Mxi-Spa secretion machinery protein MxiJ; Involved in the secretion of the Ipa antigens (241 aa)
mxiK	Mxi-Spa secretion machinery protein MxiK; Necessary for the secretion of IPA invasins (175 aa)
mxiN	Mxi-Spa secretion machinery protein MxiN (231 aa)

Fig. 4.3: Input genes

1	query	subject	%id	query cov	score	e-value	query star	query end	sub start	sub end	gap
2	NC_00433	NC_00738	98.75	52364	602	19	1122431	1174772	1176011	1228341	0
3	NC_00433	NC_00738	98.68	42802	545	7	3310672	3353454	3497562	3540363	0
4	NC_00433	NC_00738	99.05	40600	373	6	939456	980046	971373	1011968	0
5	NC_00433	NC_00738	99.28	39324	274	6	4178299	4217615	4377673	4416993	0
6	NC_00433	NC_00738	98.45	38143	541	4	76570	114684	87288	125409	0
7	NC_00433	NC_00738	98.77	36789	437	6	337773	374547	393356	430143	0
8	NC_00433	NC_00738	98.06	36580	669	7	3562947	3599493	3869695	3906268	0
9	NC_00433	NC_00738	99.13	33710	281	8	2397883	2431591	2460772	2494470	0
10	NC_00433	NC_00738	98.88	32504	357	7	2431582	2464081	2495238	2527738	0
11	NC_00433	NC_00738	98.78	31421	358	7	2658225	2689636	2777331	2808736	0
12	NC_00433	NC_00738	98.64	28953	390	5	2964860	2993811	3200752	3229700	0
13	NC_00433	NC_00738	98.64	28802	382	3	662528	691320	649283	620483	0
14	NC_00433	NC_00738	98.81	28235	327	5	4028512	4056742	4232817	4261047	0
15	NC_00433	NC_00738	99.25	27497	204	2	412497	439993	468163	495656	0
16	NC_00433	NC_00738	98.68	26646	322	7	2575540	2602182	2689550	2716167	0
17	NC_00433	NC_00738	98.29	26186	445	2	838011	864195	888056	914240	0
18	NC_00433	NC_00738	99.03	25361	226	7	1329340	1354682	1964480	1939123	0
19	NC_00433	NC_00738	98.65	25613	306	11	4433907	4459495	4681097	4655500	0
20	NC_00433	NC_00738	99.08	24525	226	0	176035	200559	200012	224536	0
21	NC_00433	NC_00738	99.17	23510	189	3	1265966	1289473	2064601	2041095	0
22	NC_00433	NC_00738	98.8	23678	280	2	3969240	3992916	4171366	4195041	0
23	NC_00433	NC_00738	98.98	23028	215	2	2479132	2502138	2589218	2612245	0
24	NC_00433	NC_00738	99.29	22551	161	0	3377819	3400369	3563035	3585585	0
25	NC_00433	NC_00738	98.95	22687	220	2	4315026	4337693	4499197	4476511	0
26	NC_00433	NC_00738	98.68	22879	297	4	41611	64488	51631	74504	0

Fig.4.4: core genome of *S.flexneri* and *S.sonnei*

CHAPTER-5

MACHINE LEARNING APPROACH

1) Review the literature through NCBI,PUBMED,PMC etc
2) Found out the most infectious species ie. *S.flexneri* and *S.soneii*
3) Complete annotated genes of *S.flexneri* was available so we took it as a model organism
4) Download the gene sets for both the strains.
5) Perform the standalone BLAST of gene sets.
6) Set the threshold 95%-100% for similarity criteria.
7) Extract the genes above this threshold.

Total genes after BLAST	GENES between threshold 95%-100%	100%similar	Total genes required
39839	14776	2102	12674

Table 5.1: Extracted Genes

8) Take out the 500 genes on the basis of length of the sequence that 60 because below this range the gene will not be having any biological significance for the further analysis.
9) Perform gene ontology studies for the enrichment of genes which includes:
-cellular location
-molecular function
-biological processes
Gene Ontology Tool-
-DAVID Tool

Description about DAVID Tool

DAVID (the Database for Annotation, Visualization and Integrated Discovery) is a free online bioinformatics resource developed by the LIB. All tools in the DAVID Bioinformatics Resources aim to provide functional interpretation of large lists of genes derived from genomic studies, e.g. microarray and proteomics studies.

The DAVID Bioinformatics Resources consists of the DAVID Knowledgebase and five integrated, web-based functional annotation tool suites: the DAVID Gene Functional Classification Tool, the DAVID Functional Annotation Tool, the DAVID Gene ID Conversion Tool, the DAVID Gene Name Viewer and the DAVID NIAID Pathogen Genome Browser. The expanded DAVID Knowledgebase now integrates almost all major and well-known public bioinformatics resources centralized by the DAVID Gene Concept, a single-linkage method to agglomerate tens of millions of diverse gene/protein identifiers and annotation terms from a variety of public bioinformatics databases. For any uploaded gene list, the DAVID Resources now provides not only the typical gene-term enrichment analysis, but also new tools and functions that allow users to condense large gene lists into gene functional groups, convert between gene/protein identifiers, visualize many-genes-to-many-terms relationships, cluster redundant and heterogeneous terms into groups, search for interesting and related genes or terms, dynamically view genes from their lists on bio-pathways and more.

FUNCTONALITY

DAVID provides a comprehensive set of functional annotation tools for investigators to understand biological meaning behind large list of genes. For any given gene list, DAVID tools are able to:

1) Identify enriched biological themes, particularly GO terms

2) Discover enriched functional-related gene groups

3) Cluster redundant annotation terms.

4) Visualize genes on BioCarta & KEGG pathway maps.

5) Display related many-genes-to-many-terms on 2-D view.

6) Search for other functionally related genes not in the list.

7) List interacting proteins.

8) Explore gene names in batch.

9) Link gene-disease associations.

10) Highlight protein functional domains and motifs.

11) Redirect to related literatures.

12) Convert gene identifiers from one type to another.

How do we use DAVID tool:

1) Take out the accession number of *S.flexneri* from BLAST result .

2) Put this accession number in NCBI and take out the UI list of whole genes set.

3) Take out the first 500 genes from the UI list for gene ontology analysis.

4) Search David tool on google, open the first link.

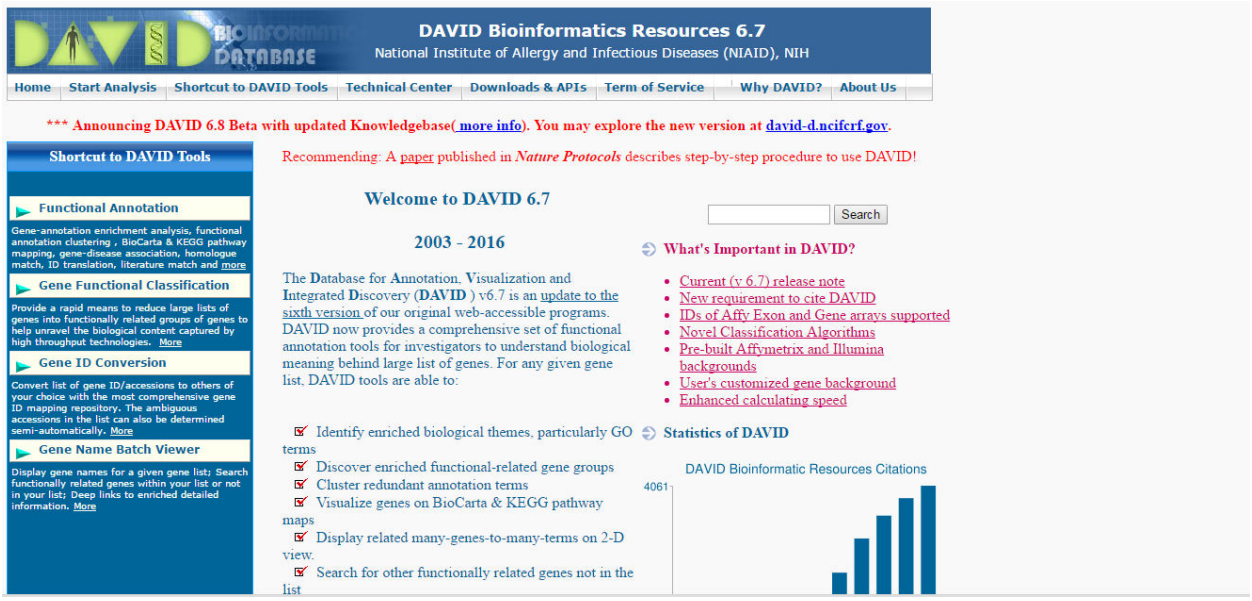


Fig.5.1: DAVID TOOL

5) Then, do gene functional annotation.

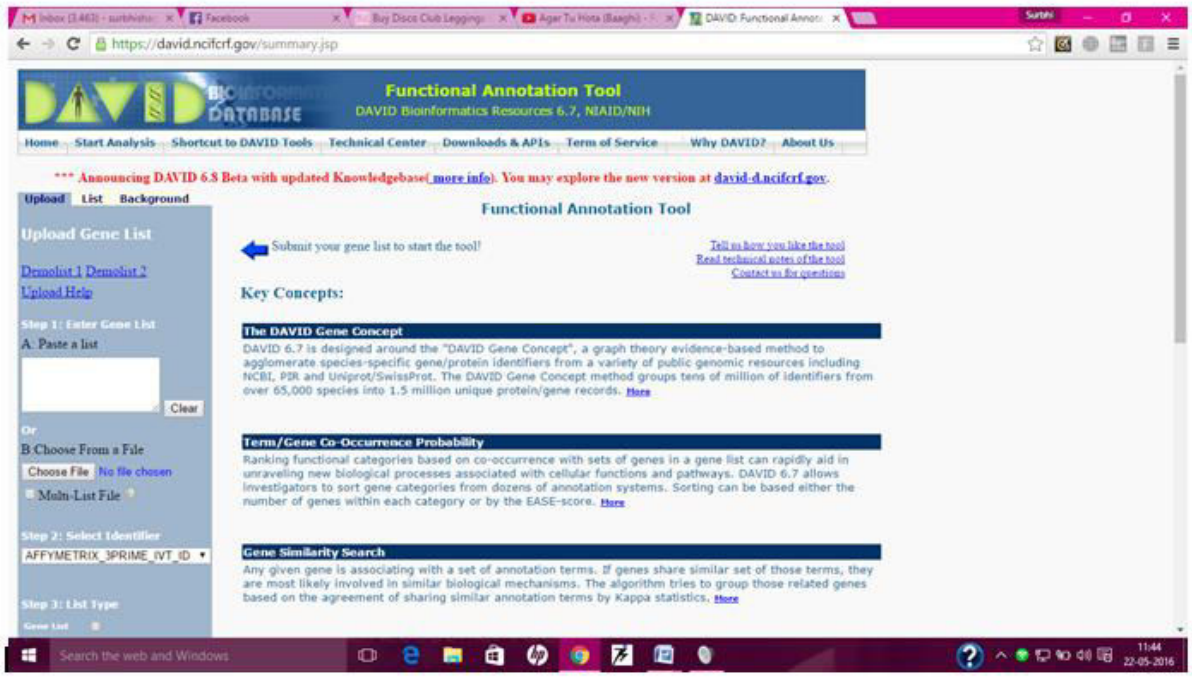


Fig.5.2: Functional Annotation

Paste first 500 genes from the UI list and select 'Entrez gene ID' in select identifier.

Now submit the result.

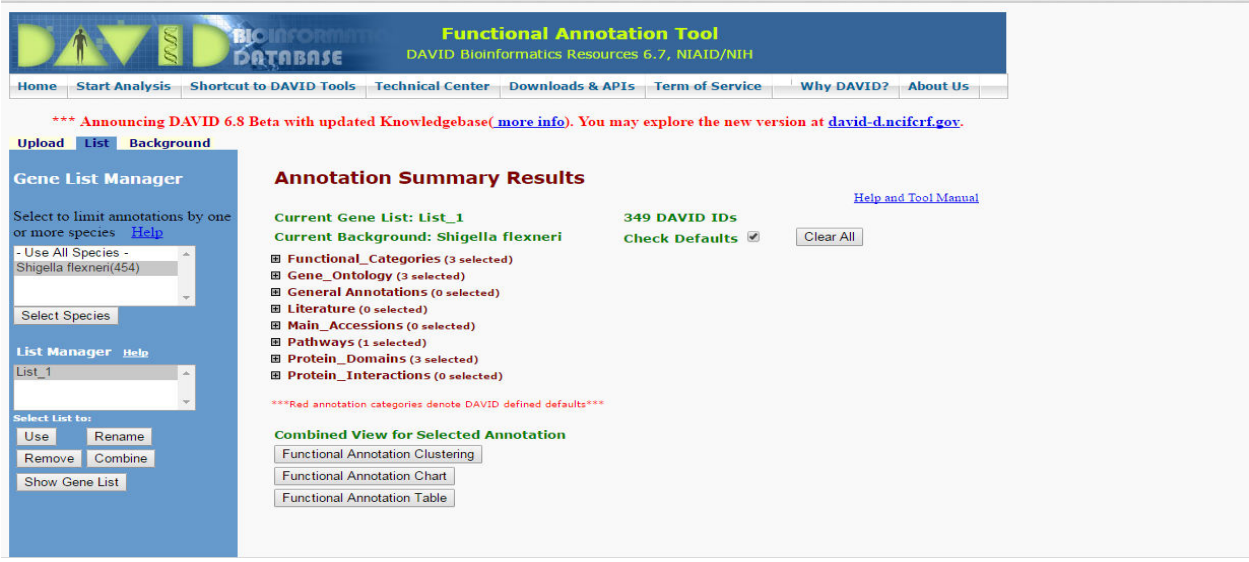


Fig. 5.3: Annotation Summary Results

Now click on functional_cateogories,

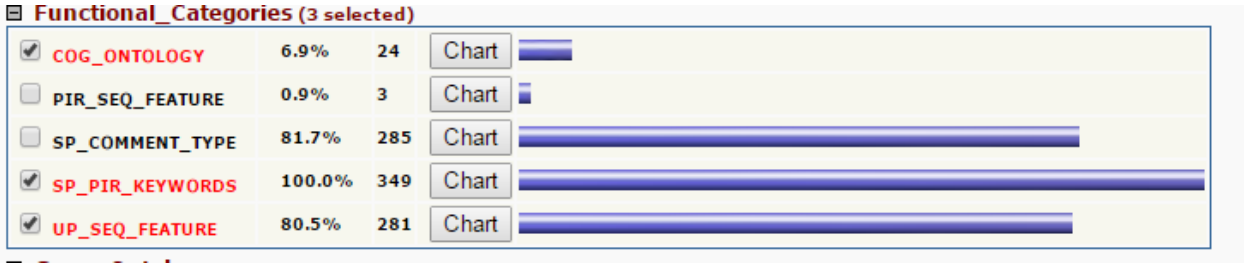


Fig. 5.4: Functional Categories

Now click on pathways :

Fig. 5.5: Pathways

Now do functional clustering :

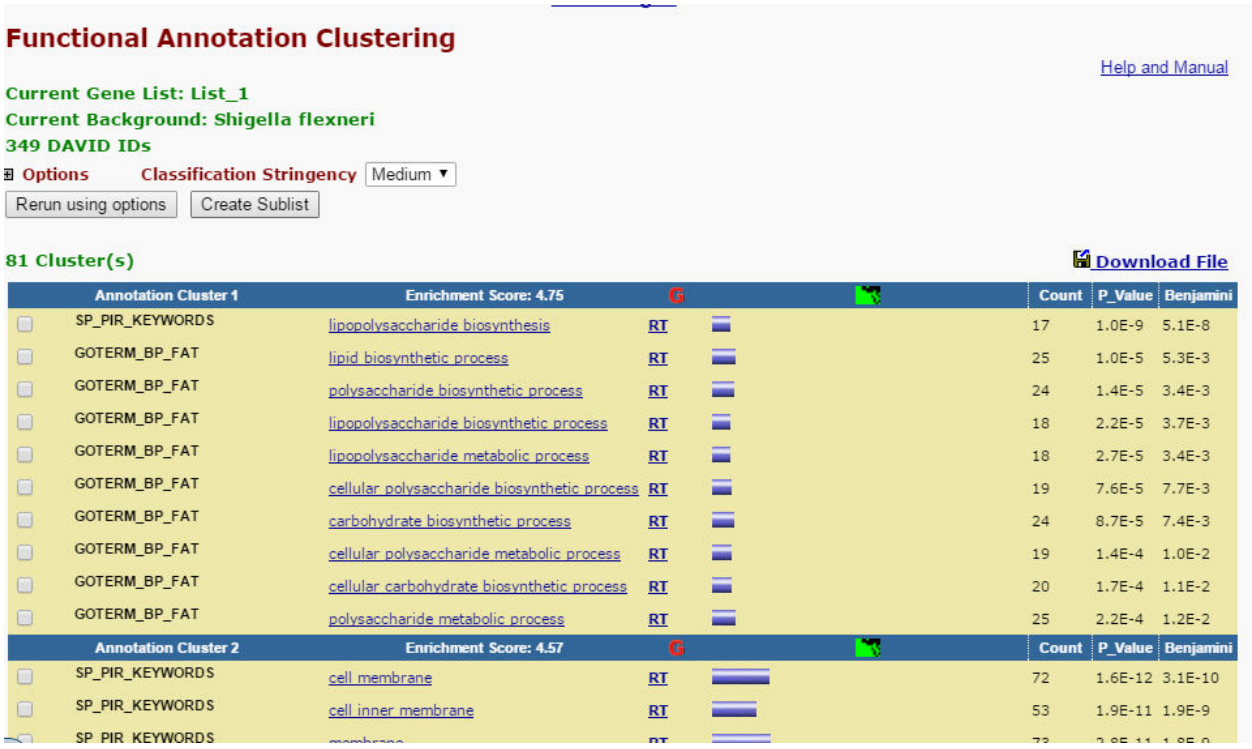


Fig. 5.6: Functional Annotation Chart

Making of Decision tree

With the help of weka, we are applying machine learning techniques to classify the essential genes of *Shigella* in classes.

Weka: Waikato Environment for Knowledge Analysis (Weka) is a popular suite of machine learning software written in Java, developed at the University of Waikato, New Zealand. It is free software licensed under the GNU General Public License.

In weka, we did the classification of genes using J 4.8 algorithm and naïve bayes algorithm. The classifer were directly applied without any feature (gene) selection. The number of top-ranked genes selected using feature selection techniques.

Methodology to perform task in WEKA

1) In the weka server, we click on filter and the attributes are supervised on the basis of 'ADD Classification'.

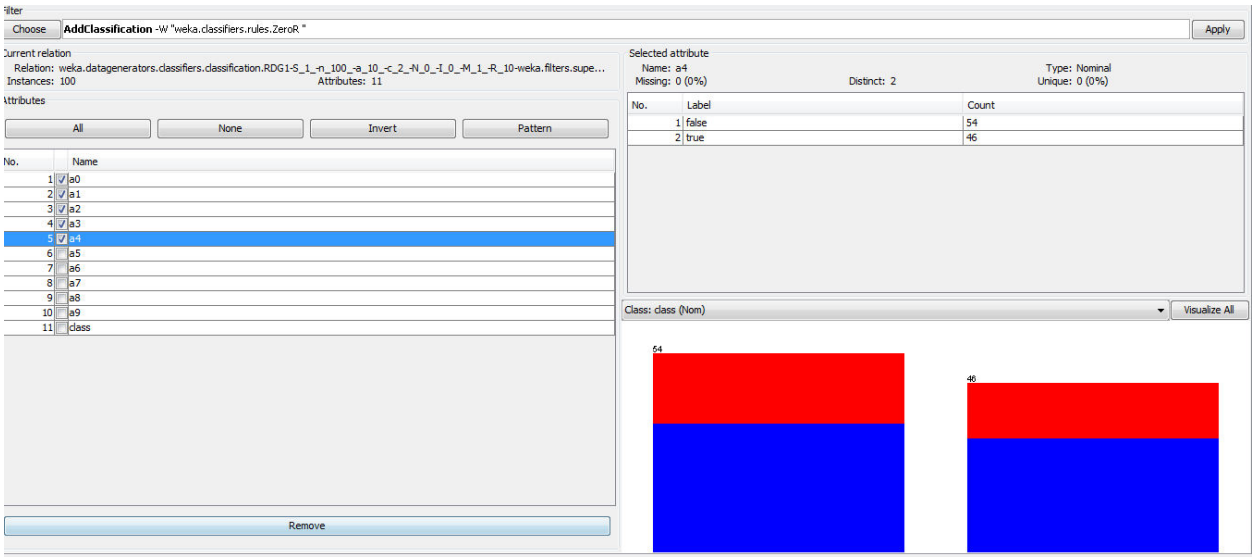


Fig. 5.7: WEKA Server

Then we select the first 5 classes of genes. Here, we get 46% which are true label and 54% which are false label.

2) Then we classify the samples genes given in complimentary files on the basis of 'Random Forest Algorithm'. The classifier was validated through '5 – fold cross validation'. Results are shown below:-

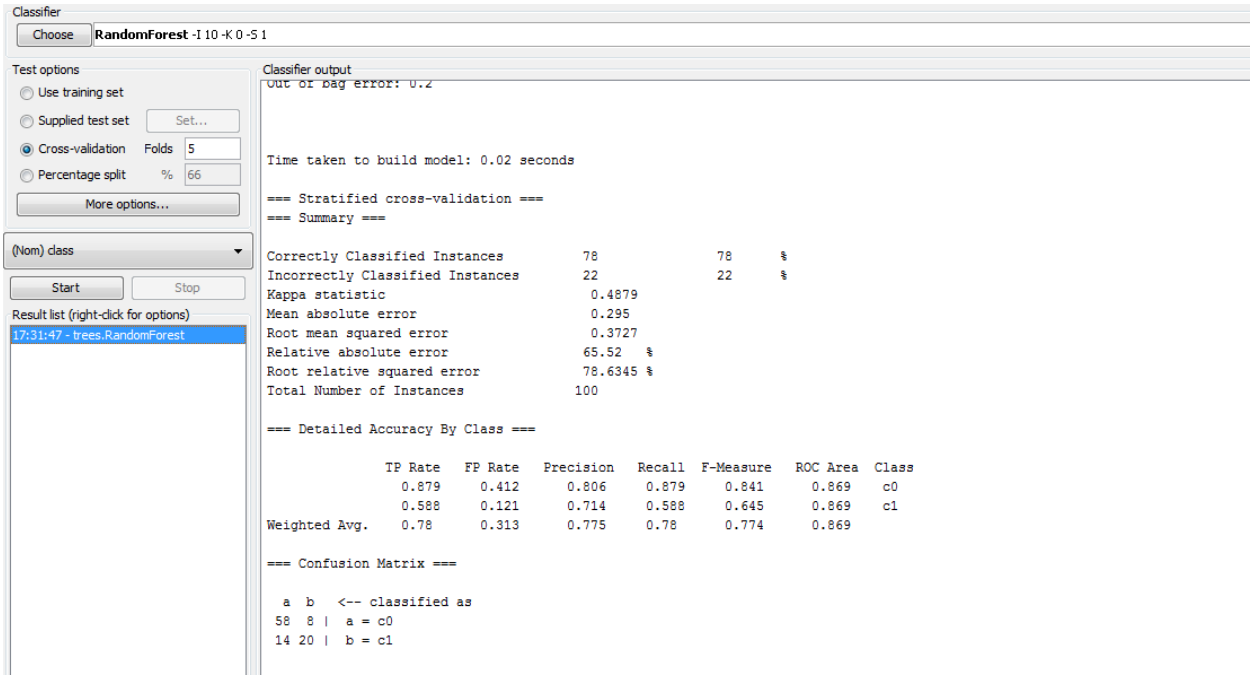


Fig.5.8: Classification of Samples

Here, we get the confusion matrix which is used to validate our build model on the basis of sensitivity and specificity.

3) Then we did the cluster on the basis of 'Expectation Maximization Algorithm'.

```
Time taken to build model (full training dat

=== Model and evaluation on training set ===

Clustered Instances

0        68 ( 68%)
1        32 ( 32%)


Log likelihood: -7.28232
```

Fig. 5.9: Clustering

4) Then we did association on the basis of 'Predictive apriori'. Here ,we get 99% accuracy of genes .

```
PredictiveApriori
==================


Best rules found:

  1. a1=false a5=false 24 ==> class=c0 24    acc:(0.99481)
```

Fig. 5.10: Association

LIST OF ESSENTIAL GENES

Gene	Gene Description
thrB	homoserine kinase
thrC	threonine synthase
yadQ	chloride chain protein
yadR	iron-sulfur cluster insertion protein ErpA
yadS	hypothetical protein
yadT	vitamin B12-transporter protein BtuF
pfs	5'-methylthioadenosine/S-adenosylhomocysteine nucleosidase
htrA	serine endoprotease
yaeG	Pseudo
yaeH	hypothetical protein
dapD	2,3,4,5-tetrahydropyridine-2,6-dicarboxylate N-succinyltransferase
glnD	PII uridylyl-transferase
map	methionine aminopeptidase
tsf	elongation factor Ts
pyrH	uridylate kinase
frr	ribosome recycling factor
yaeM	1-deoxy-D-xylulose 5-phosphate reductoisomerase
yaeS	undecaprenyl pyrophosphate synthase
cdsA	CDP-diglyceride synthetase
yaeL	zinc metallopeptidase RseP
yaeT	outer membrane protein assembly factor YaeT
hlpA	periplasmic chaperone
nhaA	pH-dependent sodium/proton antiporter

Table 5.2: Essential genes

CHAPTER 6

CONCLUSION

For human understanding of complex gemone it is important to use find essentiality of the whole genome .For this, it is essential to find vital genes involved in organism. Computational approach made this task possible. By the ontology study we are somewhat success in finding essential genes by means of finding their biological process, cellular components and molecular function of sample gene set of *Shigella* species. Classification of genes on the basis of decision tree we are succeed in finding 24 essential genes and cluster those 24 gene from set of 500 genes belong to particular class. Our proposed classification model and enrichment study on the basis of ontology of those 24 target genes will serve as standard for computing biological phenomenon related to metabolism of bacteria.

REFERENCES

1. Lindberg, Alf A., Anders Karnell, and Andrej Weintraub. "The lipopolysaccharide of Shigella bacteria as a virulence factor." *Review of Infectious Diseases* 13.Supplement 4 (1991): S279-S284.
2. Yang, Fan, et al. "Genome dynamics and diversity of Shigella species, the etiologic agents of bacillary dysentery." *Nucleic acids research* 33.19 (2005): 6445-6458.
3. Ali, Amjad, et al. "Pan-genome analysis of human gastric pathogen H. pylori: comparative genomics and pathogenomics approaches to identify regions associated with pathogenicity and prediction of potential core therapeutic targets." *BioMed research international* 2015 (2015).
4. Liu, Wei, et al. "Comparative genomics of mycoplasma: Analysis of conserved essential genes and diversity of the pan-genome." *PLoS One* 7.4 (2012): e35698.
5. Sasakawa, C. H. I. H. I. R. O., et al. "Eight genes in region 5 that form an operon are essential for invasion of epithelial cells by Shigella flexneri 2a."*Journal of bacteriology* 175.8 (1993): 2334-2346.
6. Nakata, N., et al. "Identification and characterization of virK, a virulence-associated large plasmid gene essential for intercellular spreading of Shigella flexneri." *Molecular microbiology* 6.16 (1992): 2387-2395.
7. Plaimas, Kitiporn, Roland Eils, and Rainer König. "Identifying essential genes in bacterial metabolic networks with machine learning methods." *BMC systems biology* 4.1 (2010)
8. Lett, M. C., et al. "virG, a plasmid-coded virulence gene of Shigella flexneri: identification of the virG protein and determination of the complete coding sequence." *Journal of bacteriology* 171.1 (1989): 353-359.
9. Lett, M. C., C. Sasakawa, N. Okada, T. Sakai, S. Makino, M. Yamada, K. Komatsu, and M. Yoshikawa. "virG, a plasmid-coded virulence gene of Shigella flexneri: identification of the virG protein and determination of the complete coding sequence." *Journal of bacteriology* 171, no. 1 (1989): 353-359.
10. Kaniga, Kone, David Trollinger, and Jorge E. Galan. "Identification of two targets of the type III protein secretion system encoded by the inv and spa loci of Salmonella

typhimurium that have homology to the Shigella IpaD and IpaA proteins." *Journal of Bacteriology* 17Kaniga, K., Trollinger, D., & Galan, J. E. (1995). Identification of two targets of the type III protein secretion system encoded by the inv and spa loci of Salmonella typhimurium that have homology to the Shigella IpaD and IpaA proteins. *Journal of Bacteriology*, *177*(24), 7078-7085.7.24 (1995): 7078-7085.

11. Knuth, Karin, et al. "Large-scale identification of essential Salmonella genes by trapping lethal insertions." *Molecular microbiology* 51.6 (2004): 1729-1744.
12. Knuth K, Niesalla H, Hueck CJ, Fuchs TM. Large-scale identification of essential Salmonella genes by trapping lethal insertions. Molecular microbiology. 2004 Mar 1;51(6):1729-44.
13. Yang, Fan, et al. "Genome dynamics and diversity of Shigella species, the etiologic agents of bacillary dysentery." Nucleic acids research 33.19 (2005): 6445-6458.
14. BRENNER, DON J., et al. "Polynucleotide sequence relatedness among Shigella species." International Journal of Systematic and Evolutionary Microbiology 23.1 (1973): 1-7.
15. Jin, Qi, et al. "Genome sequence of Shigella flexneri 2a: insights into pathogenicity through comparison with genomes of Escherichia coli K12 and O157." Nucleic Acids Research 30.20 (2002): 4432-4441.
16. Wei, J., et al. "Complete genome sequence and comparative genomics of Shigella flexneri serotype 2a strain 2457T." Infection and immunity 71.5 (2003): 2775-2786.
17. Venkatesan, Malabi M., et al. "Complete DNA sequence and analysis of the large virulence plasmid of Shigella flexneri." Infection and immunity 69.5 (2001): 3271-3285.
18. Lan, Ruiting, et al. "Molecular evolutionary relationships of enteroinvasive Escherichia coli and Shigella spp." Infection and immunity 72.9 (2004): 5080-5088.
19. Juhas, Mario, et al. "High confidence prediction of essential genes in Burkholderia cenocepacia." PloS one 7.6 (2012): e40064.
20. 登录, et al. "Predicting essential genes in fungal genomes."
21. Seringhaus, Michael, et al. "Predicting essential genes in fungal genomes.
22. " Genome research 16.9 (2006): 1126-1135.Nishino, Tatsuya, and Kosuke Morikawa. "Structure and function of nucleases in DNA repair: shape, grip and blade of the DNA scissors." Oncogene 21.58 (2002): 9022-9032.

23. Sidaway-Lee, Kate, et al. "Direct measurement of transcription rates reveals multiple mechanisms for configuration of the Arabidopsis ambient temperature response." Genome biology 15.3 (2014): R45.
24. Shlyakhovenko, V. O. "Ribonucleases. Possible new approach in cancer therapy." Experimental oncology 38,№ 1 (2016): 2-8.