International Journal of Biometrics and Bioinformatics (IJBB)

ISSN : 1985-2347

VOLUME 4, ISSUE 5

PUBLICATION FREQUENCY: 6 ISSUES PER YEAR

International Journal of Biometrics and Bioinformatics (IJBB)

Volume 4, Issue 5, 2010

Edited By
Computer Science Journals
www.cscjournals.org

Editor in Chief Professor João Manuel R. S. Tavares

International Journal of Biometrics and Bioinformatics (IJBB)

Book: 2010 Volume 4, Issue 5

Publishing Date: 20-12-2010

Proceedings

ISSN (Online): 1985-2347

IJBB Journal is a part of CSC Publishers

http://www.cscjournals.org

Published in Malaysia

Typesetting: Camera-ready by author, data conversation by CSC Publishing Services – CSC Journals, Malaysia

CSC Publishers

Editorial Preface

This is the fifth issue of volume four of International Journal of Biometric and Bioinformatics (IJBB). The Journal is published bi-monthly, with papers being peer reviewed to high international standards. The International Journal of Biometric and Bioinformatics are not limited to a specific aspect of Biology but it is devoted to the publication of high quality papers on all division of Bio in general. IJBB intends to disseminate knowledge in the various disciplines of the Biometric field from theoretical, practical and analytical research to physical implications and theoretical or quantitative discussion intended for academic and industrial progress. In order to position IJBB as one of the good journal on Bio-sciences, a group of highly valuable scholars are serving on the editorial board. The International Editorial Board ensures that significant developments in Biometrics from around the world are reflected in the Journal. Some important topics covers by journal are Bio-grid, biomedical image processing (fusion), Computational structural biology, Molecular sequence analysis, Genetic algorithms etc.

The coverage of the journal includes all new theoretical and experimental findings in the fields of Biometrics which enhance the knowledge of scientist, industrials, researchers and all those persons who are coupled with Bioscience field. IJBB objective is to publish articles that are not only technically proficient but also contains information and ideas of fresh interest for International readership. IJBB aims to handle submissions courteously and promptly. IJBB objectives are to promote and extend the use of all methods in the principal disciplines of Bioscience.

IJBB editors understand that how much it is important for authors and researchers to have their work published with a minimum delay after submission of their papers. They also strongly believe that the direct communication between the editors and authors are important for the welfare, quality and wellbeing of the Journal and its readers. Therefore, all activities from paper submission to paper publication are controlled through electronic systems that include electronic submission, editorial panel and review system that ensures rapid decision with least delays in the publication processes.

To build its international reputation, we are disseminating the publication information through Google Books, Google Scholar, Directory of Open Access Journals (DOAJ), Open J Gate, ScientificCommons, Docstoc and many more. Our International Editors are working on establishing ISI listing and a good impact factor for IJBB. We would like to remind you that the success of our journal depends directly on the number of quality articles submitted for review. Accordingly, we would like to request your participation by submitting quality manuscripts for review and encouraging your colleagues to submit quality manuscripts for review. One of the great benefits we can

provide to our prospective authors is the mentoring nature of our review process. IJBB provides authors with high quality, helpful reviews that are shaped to assist authors in improving their manuscripts.

Editorial Board Members
International Journal of Biometrics and Bioinformatics (IJBB)

Editorial Board

Editor-in-Chief (EiC)

Associate Editors (AEiCs)

Editorial Board Members (EBMs)

Table of Content

Volume 4, Issue 5, December 2010

Pages

Structure and Function Predictions of Hypothetical Proteins in Vibrio Phages

Swapnil G. Sanmukh swamukh1985in@rediffmail.com
Applied Aquatic Ecosystem Division,
National Environmental Engineering Research Institute,
Nehru Marg, Nagpur-440020, India

Waman N. Paunikar wn_paunikar@neeri.res.in
Applied Aquatic Ecosystem Division,
National Environmental Engineering Research Institute,
Nehru Marg, Nagpur-440020, India

Tarun K. Ghosh tk_ghosh@neeri.res.in
Applied Aquatic Ecosystem Division,
National Environmental Engineering Research Institute,
Nehru Marg, Nagpur-440020, India

Tapan Chakrabarti t_chakrabarti@neeri.res.in
National Environmental Engineering Research Institute,
Nehru Marg, Nagpur-440020, India

Abstract

The Vibriophages are the potential agents for the transfer of the virulence factor to their host through lateral gene transfer. The complete genome sequencing of various known vibriophages has been done which deciphered the presence of various gene sequences for hypothetical proteins whose function is not yet understood. We analyzed complete genome of 21such Vibriophages for hypothetical proteins from which 13 phages were sorted for our studies. Our attempt is to predict the structure and function of these hypothetical proteins by the application of computational methods and Bioinformatics. The probable function prediction of the hypothetical protein was done by using Bioinformatics web tools like CDD-BLAST, INTERPROSCAN, PFAM and COGs by searching sequence databases for the presence of orthologous enzymatic conserved domains in the hypothetical sequences. While tertiary structures were constructed using PS^2 Server (Protein Structure Prediction server). These study revealed presences of enzymatic functional domain in 92 uncharacterized proteins; their roles are yet to be discovered in Vibriophages. These deciphered enzymatic data for hypothetical proteins can be used for the understanding of functional, structural, evolutionary and metabolic development of Vibriophages and its life cycle along with their role in host evolution and pathogenicity.

Keywords: Bioinformatics Web Tools, Conserved Domains, Protein Structure Prediction, Uncharacterized Proteins, Life Cycle and Pathogenicity.

1. INTRODUCTION

The etiologic agent of cholera, *Vibrio cholerae* is a gram negative bacterium which has been reported to be infected by various specific filamentous phages (Campos, et al., 2003, Faruque, et al., 2005, Waldor, et al., 1997, Ikema, et al., 1998, Jouravleva, et al., 1998, Kar, et al., 1996, Honma, et al., 1996). CTXΦ phage has been the most studied due to its role in pathogenicity and horizontal gene transfer (Davis, et al., 2003). The phage is potentially responsible for transducing the cholera toxin genes into nonpathogenic environmental strains along with replicating directly from the bacterial chromosome for producing infective phage particles (Davis, et al., 2003, Waldor, et al., 2003). The VGJΦ is able to recombine with the CTXΦ genome to originate a hybrid phage with the full potential for virulence conversion. The hybrid phage shows an increased infectivity due to its specificity for the receptor mannose-sensitive hemagglutinin (receptor mannose- sensitive hemagglutinin pilus), which is ubiquitous among environmental strains (Campos, et al., 2003a, Campos, et al., 2003b). The vibriophages KVP40 differs from many described vibriophages in having a broad host range and is reported to infect eight *Vibrio* species, including *Vibrio cholerae* and *Vibrio parahaemolyticus*, the nonpathogenic species *Vibrio natriegens*, and *Photobacterium leiognathi* (Matsuzaki, et al., 1992).

Vibriophages (family Vibrionaceae) contains the greatest number of reported phage-host systems for the marine environment (Moebus 1987), with the genus *Vibrio* comprising most of the hosts (Moebus & Nattkemper 1981). The phage VpVs phage infect only *V. parahaemolyticus* strains (Koga et al., 1982; Kellogg et al., 1995), phage P4 (Baross et al., 1974) and KVP20 (Matsuzaki et al., 1998) infect other *Vibrio* spp. (as the VpVs in this study), whereas phage V14 (Nakanishi et al., 1966) and KVP40 (Matsuzaki et al., 1992) have been reported to infect other genera. Vibriophage has also proved to be useful in studying the host chromosomes (Guidolin and Manning, 1987).

Vibrio cholera-specific filamentous bacteriophages CTXf was first identified in 1996 (Waldor and Mekalanos, 1996). Its genome includes the genes encoding cholera toxin, an AB 5- subunit type toxin secreted by *V. cholera* during its growth in the small intestine which causes secretory diarrhoea (Lencer and Tsai, 2003). The acquisition of CTXf is an important factor for *V. cholera* virulence. Virulence factors are frequently encoded within mobile genetic elements such as phages and plasmids (Davis and Waldor, 2002). The first reported filamentous phage horizontally transmitting a virulence factor that results in lysogenic conversion of a host to become virulent was CTXf (Waldor and Mekalanos, 1996; Ochman et al., 2000). Most of the characterized phages that integrate into their respective host chromosomes also undergo a reverse reaction wherein the phage genome excises from the chromosome (Azaro and Landy, 2002). However, excision of the CTXf prophage from the *V. cholera* chromosome has never been observed (Davis and Waldor, 2000). Instead, the chromosomally integrated CTXf prophage acts as a template for synthesis of viral DNA (Davis and Waldor, 2000; Moyer et al., 2001).

The study of Vibriophages is limited to the expressed genetic characteristics which are observed through experimental studies, but to get some insight of the Vibriophages and how its acquisition imparts host to gain various new characteristics leading to virulence and evolution of both phage-host systems, the study of phage genome is essential. The *in-silico* studies of hypothetical proteins (Uncharacterized proteins) for identifying their structure and function is an attempt to understand Vibriophages and their genomes with some possible implications.

Computational biology assists us to predict the functionality in the uncharacterized sequences using the different strategies of comparative proteomics. The program's ability of homology searching using defined databases and by choosing standard parameters, the presence of the enzymatic conserved domain/s in the sequences could be searched out and it may assist in the categorizing protein into specific enzymatic family.

Bioinformatics web tools like CDD-BLAST,INTERPROSCAN, PFAM and COGs can search the orthologous sequence in biological sequence databases for the target sequence, while assist in classification of target sequence in particular family (Edward et al., 2000; Dilip and Alankar,

2009). This study will helps us to understand the probable functions of hypothetical proteins in Vibriophages.

Several online automated servers are available which can predict the three dimensional structures for protein sequences by using the strategy of aligning target sequences with orthologous sequences by virtue of sequence homology and based on that, constructs the 3Dstructure for target protein using best scored template of orthologous family member. Here, we have predicted 3-D structure using Protein Structure Prediction Server (PS^2 server) (Dilip and Alankar, 2009; Zafer et al., 2006; Chih-Chieh et al., 2006).

2. MATERIALS AND METHODS

2.1 Sequence Retrieval

The Complete protein sequences for 21 different Vibrio phages *were* downloaded from the Database of KEGG (http://www.genome.jp/kegg/). The phages under study includes Vibrio phage kappa (Ehara, et. al., unpublished), Vibrio phage VP93,Vibrio phage VEJphi (Campos, 2010), Vibrio phage N4 (Das, et. al., Unpublished),Vibrio phage fs1 (Honma, et. al., 1997),Vibrio phage K139 (Kapfhammer, et. al., 2002),Vibrio phage KVP40 (Miller, et. al., 2003),Vibrio phage fs2 (Ikema, et. al., 1992),Vibrio phage VfO3K6,Vibrio phage VfO4K68,Vibrio phage Vf33,Vibrio phage Vf12,Vibrio phage VSK (Basu, Unpublished), Vibrio phage VpV262 (Hardies, et. al., 2003),Vibrio phage VHML, Vibrio phage VGJphi (Campos, et. al., Unpublished), Vibrio phage VP2 (Wang, Unpublished),Vibrio phage VP5,Vibrio phage VP882,Vibrio phage KSF-1phi (Faruque, et. al., 2005) and Vibrio phage VP4.

2.2 Functional Annotations

Hypothetical proteins were screened for the presence of enzymatic conserved domains using sequence similarity search with close orthologous family members available in various protein databases using the web-tools. Four bioinformatics web tools like CDD-BLAST (http://www.ncbi.nlm.nih.gov/BLAST/) (Altschul et al., 1997; Schaffer et al., 2001; Aron et al., 2006), INTERPROSCAN (http://www.abi.ac.uk/interpro) (Zdobnov and Rolf, 2001), Pfam (http://www.pfam.sanger.ac.uk/) (Alex et al., 2004) and COGs (http://www.ncbi.nih, gov/cog) (Roman et al., 2000) were used, which shows the ability to search the defined conserved domains in the sequences and assist in the classification of proteins in appropriate family.

2.3 Functional Categorization

Hypothetical proteins analyzed by the function prediction web tools such as CDD-BLAST, INTERPROSCAN, PFAM and COGs have shown the variable results when searched for the conserved domains in hypothetical sequences.

2.4 Protein Structure Prediction

Several online protein structure prediction servers are available. Out of that, online PS^2 (PS Squared) Protein Structure Prediction Server was used (http://www.ps2.life.nctu.edu.tw/) (Chih-Chieh et al., 2006; Altschul et al.,1997; Schaffer et al., 2001; Cédric et al., 2000; Wendy et al.,2000), which accepts the protein (query) sequences in FASTA format and uses the strategies of Pair-wise and multiple alignment by combining powers of the programs PSI-BLAST, IMPALA and T-COFFEE in both target – template selection and target–template alignment and resultant target proteins 3D structures were constructed using structural positioning information of atomic coordinates for known template in PDB format using best scored alignment data. Where the selection of template was based on the same conserved domain detected in the functional annotations and which must be available in the structure alignment for modeling purpose.

3. RESULTS AND DISCUSSION

The *in silico* structure and function of the Vibriophages was worked out for 21 phages. Out of 21 Vibriophages, conserved domain prediction in hypothetical proteins was possible in 13 phages. The hypothetical proteins were screened for the presence of enzymatic conserved domains using

sequence similarity search with close orthologous family members available in various protein databases using the web tools. The 3-D structure prediction of protein (query) sequences in FASTA format and uses the strategies of Pair-wise and multiple alignment by combining powers of the programs PSI-BLAST, IMPALA and T-COFFEE in both target – template selection and target– template alignment and resultant target proteins 3D structures were constructed using structural positioning information of atomic coordinates for known template in PDB format using best scored alignment data. Where the selection of template was based on the same conserved domain detected in the functional annotations and which must be available in the structure alignment for modeling purpose.

3.1 Functional Annotations and Protein Structure Prediction

The analysis of hypothetical proteins of Vibriophages was accomplished by using web tools for their classification into particular enzymatic family based on enzymatic conserved domain available in the sequence which are represented in respective Table 1 through 13. In 13 different Vibriophages, 215 hypothetical proteins resulted in 205 functional annotations out of which 92 are showing enzymatic conserved domains.

The $(PS)^2$ Server built the three dimensional structures for hypothetical proteins. Where in 17 different Vibriophage genome analyzed, $(PS)^2$ satisfactorily predicted structures of 54 hypothetical proteins using best scored orthologous template. The resulted 10 structures out of 54 showed no functional conserved domains may be due to lack of due to the lack of defined 3D structures for the aligned templates. The 3-D structures built are represented sequentially in respective Vibriophage specific gene. The templates with best scoring with hypothetical protein sequences are represented in the order as Template ID, Identity, Score and E-value which represented in structure column of each Vibriophage gene analyzed. The structure and functional data for Vibrio phage VfO3K6 (Table 1), Vibrio phage Vf33 (Table 2), Vibrio phage KSF-1phi Table 3), Vibrio phage VP4 (Table 4), Vibrio phage kappa (Table 5), Vibrio phage fs1 (Table 6), Vibrio phage K139 (Table 7), Vibrio phage KVP40 (Table 8), Vibrio phage VP93 (Table 9) , Vibrio phage N4 (Table 10) , Vibrio phage VP2 (Table 11) , Vibrio phage VP5 (Table 12) and Vibrio phage VP882 (Table 13) are given in their respective tables.

4. CONCLUSION

This study sorted some functional hypothetical proteins of Vibriophages applying the parameters of pair-wise and multiple sequence alignment tools along with structure prediction tools, which suggests that many probable functional uncharacterized proteins are available in the Vibriophages. Development in sequence analysis programming and ever growing genome sequence databases enhanced this methodology to draw conclusive functional relationships in the hypothetical proteins under study. Bioinformatics Web Tools like CDD-BLAST, INTERPROSCAN, PFAM and COGs have shown the ability to predict structure and functions in 215 hypothetical proteins of Vibriophages, in that sense assisted in predicting functional activity in 205 hypothetical proteins, out of which 10 showed only structural results and no functional activity was found in them. In all 54 3-D structures for hypothetical proteins was constructed using $(PS)^2$ serves as fast automated homology modeling web server. This predicted three dimensional structures may assist in establishing their role in life cycle of Vibriophages whose exact role in phage-host lifecycle is still unclear and can be used in future for the study of virulence and evolution of both phage-host systems.

5. DISCUSSION

The in-silico analysis of the hypothetical proteins is proved only on expression of the selective gene through cloning. The results obtained are concluded on the bases of available information in different databases and are valid till date.

Table 1 :Vibrio phage VfO3K6

NCBI gene ID	CDD BLAST	INTERPROSCAN	PFAM	COGS	Structures
1262767	No	No	Phage related protein & MraW methylase family	No	3cecA -34-35 -0.005

Table 2 Vibrio phage Vf33

NCBI gene ID	CDD BLAST	INTERPROSCAN	PFAM	COGS	Structures
2853318	No	No	Chromate transporter	No	No

Table 3 Vibrio phage KSF-1phi

NCBI gene ID	CDD BLAST	INTERPROSCAN	PFAM	COGS	Structures
3031573	No	No	Retrograde transport protein Dsl1 N & CRISPR-associated protein	No	No
3031575	No	No	Baculovirus 11 kDa family	No	No
3031578	No	No	Archaeal ATPase	ABC-type multidrug/protein/lipid transport system, ATPase component	No

Table 4 Vibrio phage VP4

NCBI gene ID	CDD BLAST	INTERPROSCAN	PFAM	COGS	Structures
3800005	Nucleoside/nucleotide kinase (NK)	No	No	No	No
3800011	No	No	No	No	1dekA-17- 132-6e-32

Table 5 Vibrio phage kappa

NCBI gene ID	CDD BLAST	INTERPROSCAN	PFAM	COGS	Structures
5850542	No	No	Thiamin pyrophosphokinase, catalytic domain & Sporulation related domain	No	No
5850551	S-adenosylmethionine-dependent methyltransferases	DNA methylase, N-6 adenine-specific, conserved site & N6 adenine-specific DNA methyltransferase	Methyltransferase small domain & Ribosomal L32p protein	No	2okcB -13- 64- 1e-11
5850544	Helix-turn-helix	Helix-turn-helix & Lambda repressor-like, DNA-binding	Helix-turn-helix	Predicted transcriptional regulators	1b0nA-20- 51- 1e-07
5850553	phage zinc-binding transcriptional activators	Phage transcriptional activator, Ogr/Delta	Ogr/Delta-like zinc finger,Insertion element protein & Dam-replacing	No	No
5850575	P2_Phage_GpR super family[cl06104	P2 phage tail completion R	P2 phage tail completion protein R (GpR)	No	No
5850569	No	No	MerC mercury resistance protein ,Diacylglycerol acyltransferase	No	No
5850560	No	No	Anti-sigma-K factor rskA , Bacteriophage lysis protein , Hepatic lectin, N-terminal domain	No	1i84S- 22-33- 0.010
5850548	Baseplate_J super family[cl01294	Baseplate assembly protein J-like, predicted	Baseplate J-like protein	No	No
5850543	No	No	Phage tail protein (Tail_P2_I)	No	No
5850540	No	No	Baculovirus polyhedron envelope protein, PEP, C terminus , FlgN protein	Methyl-accepting chemotaxis protein	No
5850550	No	No	BRO family, N-terminal domain , NTF2-like N-terminal transpeptidase domain	No	No
5850584	No	No	No	No	3cddD-14- 35 -3e-04

Table 6 Vibrio phage fs1

NCBI gene ID	CDD BLAST	INTERPROSCAN	PFAM	COGS	Structures
955575	No	No	Procyclic acidic repetitive protein (PARP) & Potato leaf roll virus readthrough protein	No	No
955576	No	No	Exonuclease VII,Bacillus transposase protein ,Reovirus sigma C capsid protein,Allexivirus 40kDa protein ,Baculovirus polyhedron envelope protein,Filoviridae VP35,Biogenesis of lysosome & Nucleopolyhedrovirus P10 protein	No	1m1jC-17- 38-0.002

955584	DNA replication initiation protein	No	No	Putative phage replication protein RstA	2gtqA- 24- 37- 0.002
955585	Rep_trans super family, Plasmid replication is initiated by the replication initiation factor (REP).	Replication initiation factor	Replication initiation factor	Putative phage replication protein RstA	No

Table 7 Vibrio phage K139

NCBI gene ID	CDD BLAST	INTERPROSCAN	PFAM	COGS	Structures
929070	No	No	Thiamin pyrophosphokinase, catalytic domain & Sporulation related domain	No	No
929074	No	No	Indoleamine 2,3-dioxygenase	No	No
929077	Helix-turn-helix	Helix-turn-helix	Helix-turn-helix	Predicted transcriptional regulators	1b0nA- 20- 51- 1e-07
929087	P2_Phage_GpR super family	P2 phage tail completion R	P2_Phage_GpR super family	No	No
929093	No	No	MerC mercury resistance protein & Diacylglycerol acyltransferase	No	No
929096	No	No	No	No	1i84S- 22- 33- 0.010
929101	Baseplate_J super family[cl01294], The P2 bacteriophage J protein lies at the edge of the baseplate.	Baseplate assembly protein J-like, predicted	Baseplate J-like protein	No	No
929102	No	No	Phage tail protein (Tail_P2_I)	No	No
929106	No	No	FlgN protein & Baculovirus polyhedron envelope protein	Methyl-accepting chemotaxis protein	No
929107	No	No	NTF2-like N-terminal transpeptidase domain & BRO family, N-terminal domain BRO-A and BRO-C are DNA binding proteins that influence host DNA replication and/or transcription	No	No
929108	No	No	No	No	3cddD- 14- 35- 3e-04
929110	Breast Cancer Suppressor Protein (BRCA1) & NAD-dependent DNA ligase	BRCT	BRCA1 C Terminus (BRCT) domain	NAD-dependent DNA ligase	No

Table 8 Vibrio phage KVP40

NCBI gene ID	CDD blast	Interproscan	Pfam	Cogs	Structure
2545647	F420_ligase super family[cl00644]	No	DUF218 domain	No	No
2545650	No	No	Caf1 Capsule antigen	No	No
2545653	No	No	Transglycosylase associated protein	No	No
2545654	Phage_head_chap super family[cl12668]	Bacteriophage T4, Gp40, head assembly	Head assembly gene product	Predicted ATP-dependent protease	No
2545674	No	No	Gap junction channel protein cysteine-rich domain	No	No
2545675	No	No	Plasmid conjugative transfer entry exclusion protein TraS	No	No
2545680	No	No	Fibronectin type III domain	No	No
2545681	DUF3307 super family[cl13235]	No	Protein of unknown function (DUF3307)	No	No
2545684	No	No	Glycosyltransferase family 52,Monogalactosyldiacylglycerol (MGDG) synthase	No	No
2545686	PRTase_typeII super family[cl12019], Phosphoribosyltransferase (PRTase) type II	Nicotinate phosphoribosyltransferase-like	Nicotinate phosphoribosyltransferase (NAPRTase) family	Nicotinic acid phosphoribosyltransferase	2g95B- 18- 320- 2e-88
2545687	30.2 super family[cl14359], hypothetical protein	No	No	No	2jarA- 14- 63- 5e-11
2545688	Radical_SAM super family[cl14056],NrdG[COG0602], Organic radical activating enzymes	No	Radical SAM superfamily	Organic radical activating enzymes	1tv8A- 16- 41- 3e-04
2545689	No	No	Poly(ADP-ribose) polymerase catalytic domain, RNA 2'-phosphotransferase, Tpt1 / KptA family	No	3c4hA- 15- 37- 0.008
2545691	No	No	Frag1/DRAM/Sfk1 family,(DUF2976), (DUF1625),(DUF2569), (DUF373)	No	1ii7A- 12- 40- 6e-04
2545694	MPP_superfamily super family[cl13995], Metallophosphatases	Metallo-dependent phosphatase	Calcineurin-like phosphoesterase	Predicted phosphohydrolases	No

	(MPPs),Metallophos[pfam00149]				
2545704	MPP_superfamily super family[cl13995], Metallophosphatases (MPPs)	No	No	Predicted phosphohydrolases	1xm7A- 24- 103- 2e-23
2545705	No	Prepilin-type cleavage/methylation, N-terminal	Prokaryotic N-terminal methylation motif	No	2hi2A- 63- 39- 3e-04
2545706	No	No	(DUF1469), (DUF973),(HAP),(DUF2614),(DUF2062),Vpu protein,UNC-50 family,ABC-2 type transporter,Secretion system effector C (SseC) like family	No	No
2545718	No	No	Post-segregation antitoxin CcdA,Ribosomal L29 protein, Leucine permease transcriptional regulator helical domain, Region found in RelA / SpoT proteins	No	No
2545725	No	No	Biofilm regulator BssS	No	No
2545728	No	No	PKC-activated protein phosphatase-1 inhibitor	No	No
2545733	DUF458 super family[cl00861]	Protein of unknown function DUF458, RNase H-like	Protein of unknown function (DUF458)	No	No
2545735	DUF2828[pfam11443]	No	Domain of unknown function (DUF2828)	No	1yvrA- 12- 94- 1e-19
2545736	No	No	Special lobe-specific silk protein SSP160	No	No
2545743	No	No	Acyl-ACP thioesterase	No	No
2545744	No	No	GatB domain,Septum formation initiator ,TATA element modulatory factor 1 DNA binding ,PspA/IM30 family,,She9 / Mdm33 family, Flagellar protein FliT	No	No
2545748	No	No	(DUF1014),ZF-HD protein dimerisation region	No	No
2545749	No	No	M protein trans-acting positive regulator (MGA) HTH domain	No	2hbtA- 13- 38- 0.002
2545750	No	No	CYTH domain	No	2fblB- 22- 79- 1e-15
2545751	No	No	Bacterial virulence protein (VirJ)	No	No
2545752	No	No	ORF6C domain, Gal4-like dimerisation domain,Bacillus transposase protein ,Acetyl co-enzyme A carboxylase carboxyltransferase alpha subunit, Tetrahydromethanopterin S-methyltransferase subunit B ,Toxic anion resistance protein (TelA),Baculovirus polyhedron envelope protein	Chromosome segregation ATPases	No
2545753	Adenine nucleotide alpha hydrolases superfamily including N type ATP PPases	No	No	Predicted ATPase (PP-loop superfamily), confers aluminum resistance	2pg3A- 14- 202- 3e-53
2545759	No	No	Bacterial alpha-L-rhamnosidase , ARP2/3 complex 16 kDa subunit (p16-Arc)	No	2i6gB- 15- 36- 0.006
2545760	No	No	Histidine kinase ,DUF576	No	No
2545761	30.2 super family[cl14359], hypothetical protein, COG1011[COG1011], Predicted hydrolase (HAD superfamily)	No	haloacid dehalogenase-like hydrolase	No	1q92A- 18- 52- 1e-07
2545763	No	No	Ubiquitin-fold modifier-conjugating enzyme 1	No	No
2545764	No	No	KRAB box ,Cytochrome C biogenesis protein,Septum formation topological specificity factor MinE	No	No
2545767	23 super family[cl14344], major capsid protein	No	Sucrose-6F-phosphate phosphohydrolase, Major capsid protein Gp23	No	No
2545768	No	No	XisH protein	No	No
2545771	No	Thioredoxin-like fold	Glutaredoxin	No	No
2545773	No	No	No	1,4-alpha-glucan branching enzyme	No
2545779	No	No	No	No	1potA -27- 36 -0.004
2545780	No	No	Iron dependent repressor, N-terminal DNA binding domain	No	No
2545782	GatB_Yqey super family[cl11497]	Aspartyl/glutamyl-tRNA amidotransferase subunit B-related,Protein of unknown function YOR215C, mitochondrial	Yqey-like protein	Uncharacterized ACR	1ng6A- 25- 125- 2e-30
2545788	No	No	RIO1 family , Beta-trefoil ,(DUF2972)	No	No
2545795	No	No	Herpes virus protein UL24	No	No
2545796	57B super family[cl14352]	RNA ligase/cyclic nucleotide phosphodiesterase	2',5' RNA ligase family	No	1jh6A- 18- 44- 2e-05

2545798	No	No	No	ATP-dependent protease Clp, ATPase subunit	No
2545799	No	No	Flagellar motor switch protein FliM	No	No
2545801	No	No	Nickel-containing superoxide dismutase	No	1qgrA- 10- 32- 0.004
2545805	No	No	Serpentine type 7TM GPCR chemoreceptor Srbc	No	No
2545809	No	No	PAAR motif, phosphotransferase system, EIIB	No	
2545814	Band_7 super family[cl02525]	Band 7 protein	SPFH domain / Band 7 family	membrane protease subunits, somatin/prohition homologs	3bk6A- 13- 128- 9e-31
2545825	NT5C super family[cl01869]	5'(3')-deoxyribonucleotidase,	5' nucleotidase, deoxy (Pyrimidine), cytosolic type C protein (NT5C)	No	3bwvB- 25- 159- 2e-40
2545826	DUF1768 super family[cl01271],	Bacteriophage GP30.3	Bacteriophage protein GP30.3	No	2b3wA- 20- 89 -5e-19
2545830	No	No	Thymidine kinase	No	No
2545834	alt[PHA02566]	No	No	No	1r45A- 16- 41- 4e-04
2545837	No	No	Integral membrane protein	No	No
2545840	No	No	Enterobacterial protein of unknown function	No	No
2545841	No	No	Zinc-binding domain of primase-helicase , PHD-finger	No	No
2545843	No	No	Ion transport protein ,(DUF2530) , (DUF1119)	No	No
2545844	No		Saposin-like type B, region 1	No	1oygA- 23- 34- 0.008
2545848	No	No	Flagellar protein FliT , MerR, DNA binding, Potyviridae polyprotein, Proteasome complex subunit Rpn13 ubiquitin receptor	No	No
2545849	No	Zinc finger, C2H2-like	Ribosomal protein L33	No	No
2545851	No	No	Enhancer of rudimentary, Lysophospholipase catalytic domain,Pyrrolo-quinoline quinone coenzyme N-terminus ,(DUF3228)	No	No
2545854	No	No	Predicted permease, (DUF2899)	No	
2545861	No	No	Ca2+ regulator and membrane fusion protein	No	No
2545863	No	No	Predicted membrane protein (DUF2324)	No	No
2545865	No	No	Epstein-Barr virus nuclear antigen 3 (EBNA-3)	No	No
2545866	No	No	M61 glycyl aminopeptidase	No	No
2545871	No	No	Protein of unknown function (DUF2682)	No	No
2545876	No	No	Periplasmic binding protein	No	No
2545877	No	DNA methylase, C-5 cytosine-specific, active site	No	No	No
2545878	No	No	Baculoviral E56 protein, specific to ODV envelope, Orbivirus NS3 ,Calcium-activated chloride channel	No	No
2545879	Nuc-transf super family[cl01417	Nucleotidyltransferase, predicted	Predicted nucleotidyltransferase	No	2v3cC- 14 -36- 0.007
2545884	No	No	WW domain	No	No
2545885	Lysine_decarbox super family[cl00695]	No	No	No	No
2545886	No	Neuraxin/MAP1B repeat	No	No	No
2545887	No	No	Agenet domain	No	No
2545889	No	No	CobW/HypB/UreG, nucleotide-binding domain , Borrelia burgdorferi BBR25 lipoprotein	ATPases with chaperone activity, ATP-binding subunit	1qvrA- 27- 35- 0.004
2545897	TFold super family[cl00263],	6-pyruvoyl tetrahydropterin synthase-related	6-pyruvoyl tetrahydropterin synthase	6-pyruvoyl-tetrahydropterin synthase	1y13A- 17- 49- 1e-06
2545899	No	Glutamine amidotransferase, type II	No	No	1ao0A- 17- 113- 2e-26
2545900	No	No	Prokaryotic membrane lipoprotein lipid attachment site	No	No
2545901	No	No	Queuine tRNA-ribosyltransferase	No	No
2545905	No	No	Glycosyl hydrolase family 98 ,(DUF1795)	No	No
2546061	NO	NO	DUF1219,Orthopoxvirus A49R protein,	NO	No
			DUF1967, Terpene synthase family, metal binding domain		
2546059	NO	No	Prokaryotic membrane lipoprotein lipid attachment site ,MLTD_N	NO	No
2546055	UvsW super family[cl13141]	DNA helicase, ATP-	ATP-dependant DNA helicase UvsW	NO	No

		dependent, UvsW			
	This family of proteins represents the DNA helicase UvsW from bacteriophage T4				
2546054	NO	NO	Phage portal protein, lambda family	NO	2jpnA- 28- 41- 2e-04
2546050	NO	No	Type IV leader peptidase family	NO	No
2546049	NO	No	Type IV leader peptidase family	NO	No
2546033	NO	NO	Adenoviral DNA terminal protein,FFD and TFG box motifs	NO	No
2546032	NO	NO	Quinohemoprotein amine dehydrogenase, gamma subunit	NO	No
2546029	NO	No	DNA binding domain of tn916 integrase	NO	No
2546021	NO	NO	Fibritin C-terminal region	NO	1nayA- 70- 39- 0.003
2546017	NO	NO	DNA gyrase C-terminal domain, beta-propeller	NO	No
2546012	NO	NO	Restriction endonuclease EcoRII, N-terminal Opioid growth factor receptor (OGFr) conserved region	NO	No
2546007	NO	No	Predicted membrane protein	NO	No
2546006	NO	No	General secretion pathway, M protein ,Bacterial protein of unknown function (DUF948) Cytomegalovirus TRL10 protein ,Sodium ion transport-associated	NO	No
2546005	NO	No	Colicin V production protein ,Srg family chemoreceptor SNARE associated Golgi protein ,Protein of unknown function (DUF3590)	NO	No
2546004	Macro_Poa1p_like[cd02901], Macro domain, Poa1p_like family	Appr-1-p processing	Macro domain	NO	1vhuA- 18- 40- 5e-04
2546002	SprT super family[cl01182], Predicted to have roles in transcription elongation	No	SprT-like family	Uncharacterized BCR	No
2546001	NO	NO	Uncharacterized protein conserved in archaea, NAF domain	NO	No
2545995	NO	NO	Cancer susceptibility candidate 1	NO	No
2545993	NO	NO	Iron/manganese superoxide dismutases, C-terminal domain	NO	No
2545991	NO	NO	Aldehyde dehydrogenase family ,Phage GP30.8 protein	NO	No
2545990	NO	No	Glycosyl transferases group 1	NO	No
2545987	NO	ATPase, AAA+ type, core,	NO	NO	No
2545983	DUF2829 super family[cl12744],This proteins found in bacteria and bacteriphages.	No	No	NO	No
2545981	NO	Hedgehog/DD-peptidase, zinc-binding motif Peptidase M15A, C-terminal	Peptidase M15	NO	1lbuA- 25- 42- 4e-05
2545979	NO	NO	NO	ATPase involved in DNA repair	No
2545978	NO	No	Heat shock factor binding protein 1 , Bacterial flagellin N-terminal helical region	No	No
2545977	NO	NO	Ribonuclease R winged-helix domain ,DUF1514	No	No
2545976	NO	NO	G10 protein	Predicted GTPase	No
2545970	NO	NO	AP2 domain	No	No
2545957	AdoMet_MTases[cd02440],S-adenosylmethionine-dependent methyltransferases	No	Methyltransferase small domain,DNA N-6-adenine-methyltransferase (Dam)	No	No
2545954	NO	NO	Sporulation lipoprotein YhcN/YlaJ (Spore_YhcN_YlaJ)	No	No
2545950	NO	NO	NO	Molecular chaperone	No
2545949	NO	No	Invasion associated locus B (IalB) protein	No	No
2545946	NO	No	Sugar (and other) transporter	No	No
2545940	NO	NO	Peptidyl-tRNA hydrolase PTH2	No	1rzwA - 23- 40- 4e-04
2545939	NO	NO	DUF2536,Uncharacterized protein conserved in bacteria (DUF2312)	No	No
2545936	NO	NO	NO	No	2hbtA- 14- 39- 0.001

2545929	NO	NO	MSV199 domain	No	No
2545926	NO	Armadillo-type fold	HEAT repeat	No	1lrvA- 21 - 49- 2e-06
2545924	NO	NO	Mor transcription activator family	No	No
2545923	NO	NO	Merozoite surface protein (SPAM) ,DUF2392, CTP synthase N-terminus	No	No
2545922	NO	NO	MULE transposase domain	No	No
2545917	NO	No	SUR7/PalI family ,DUF2499	No	No
2545914	NO	No	Colicin V production protein,Transmembrane amino acid transporter protein Wnt-binding factor required for Wnt secretion,DUF3021	No	No
2545913	NO	NO	Rad4 beta-hairpin domain 3 ,DUF2585	No	No
2545907	NO	NO	Arabidopsis thaliana protein of unknown function (DUF821)	No	No

Table 9 Vibrio phage VP93

NCBI Gene ID	CDD BLAST	INTERPROSCAN	PFAM	COGS	Structures
7853570	NO	NO	POPLD (NUC188) domain	NO	No
7853571	NO	NO	Ribosomal protein S9/S16 CENP-B N-terminal DNA-binding domain	NO	No
7853573	NO	NO	tRNA synthetases class II (A)	NO	No
7853580	NO	NO	SOCS box	NO	No
7853581	No	No	No	No	1u3eM - 16- 41- 4e-05
7853583	NO	NO	Prolyl 4-Hydroxylase alpha-subunit, N-terminal region	NO	No
7853584	NT_Pol-beta-like super family[cl11966]	NO	Poly A polymerase head domain	tRNA nucleotidyltransferase/poly(A) polymerase	1ou5A- 41- 42- 1e-04
7853585	PHA02030[PHA02030], hypothetical protein	NO	NO	NO	No
7853587	NO	Phosphoribosyl-ATP pyrophosphohydrolase-like	Phosphoribosyl-ATP pyrophosphohydrolase	Predicted pyrophosphatase	1vmgA- 27- 47- 3e-06
7853592	cyt_kin_arch[TIGR02173]	NO	AAA domain (dynein-related subfamily)	ATPase involved in DNA replication	1dekA- 15 -42- 5e-05
7853594	NO	NO	6-phosphofructo-2-kinase	NO	No
7853600	NO	NO	Rickettsia 17 kDa surface antigen Bacteriocin class II with double-glycine leader peptide LMBR1-like membrane protein	NO	2i1kA- 14- 37- 0.004
7853605	PHA02046 super family[cl10354]	NO	EspA-like secreted protein	DNA-directed RNA polymerase sigma subunits (sigma70/sigma32)	No
7853607	NO	NO	IRSp53/MIM homology domain ,Histidine kinase Phi29 scaffolding protein,Centromere protein H (CENP-H)	NO	No
7853608	NO	NO	Ribosomal protein S30	NO	No
7853609	Peptidase_M15_3 super family[cl01194], Peptidase M15	Hedgehog/DD-peptidase, zinc-binding motif Peptidase M15A, C-terminal	Peptidase M15	TPR-repeat-containing proteins	1lbuA- 21- 50- 2e-07
7853613	NO	NO	Cyclin, N-terminal domain	NO	No

Table 10 Vibrio phage N4

NCBI Gene ID	CDD BLAST	INTERPROSCAN	PFAM	COGS	Structures
8676422	NO	NO	Villin headpiece domain	NO	No
8676425	Wzz super family[cl01623]	NO	DUF848, STAT protein, all-alpha domain Autophagy protein 16 (ATG16),DUF641 Tumour-suppressor protein CtIP N-terminal domain TATA element modulatory factor 1 DNA binding Afadin- and alpha -actinin-Binding Spc7 kinetochore protein,Kinesin-related Cobalamin adenosyltransferase,TMPIT-like protein CorA-like Mg2+ transporter protein Erp protein C-terminus	ATPase involved in DNA repair	No
8676426	No	No	No	No	1dekA-

					17- 125- 8e-30
8676431	NO	Beta tubulin, autoregulation binding site	Endoplasmic reticulum-based factor for assembly of V-ATPase	NO	No
8676439	NO	NO	DUF2675,Brucella outer membrane protein 2	NO	No
8676447	NO	Bacteriophage T7-like, gene 6.7	NO	NO	No
8676464	NO	NO	DUF2133,GHMP kinases N terminal domain	NO	No
8676465	NO	NO	Lysis protein,Prokaryotic membrane lipoprotein lipid attachment site	NO	No

Table 11 Vibrio phage VP2

NCBI Gene ID	CDD BLAST	INTERPROSCAN	PFAM	COGS	Structures
2948097	NO	NO	Outer membrane lipoprotein LolB	NO	No
2948099	NO	NO	VRR-NUC domain	NO	No
2948114	NO	NO	D-Ala-teichoic acid biosynthesis protein,Rap-phr extracellular signalling	NO	No
2948115	NO	Peptidase S26A, signal peptidase I, serine active site	NO	NO	No
2948116	NO	NO	Minor capsid	NO	No
2948120	NO	NO	Tim17/Tim22/Tim23 family,Mitochondrial ribosomal protein L28	NO	No
2948127	NO	NO	(DUF2459), short chain dehydrogenase	NO	No
2948137	NO	NO	Mediator complex subunit 3 fungal	NO	No
2948139	NO	NO	Sulfolobus plasmid regulatory protein	NO	No

Table 12 Vibrio phage VP5

NCBI Gene ID	CDD BLAST	INTERPROSCAN	PFAM	COGS	Structures
5741329	NO	No	Glycosyl hydrolases family 6	NO	No
5741330	NO	NO	SYF2 splicing factor	NO	No
5741333	HDc super family[cl00076] Metal dependent phosphohydrolases with conserved 'HD' motif	Metal-dependent phosphohydrolase, HD domain	HDOD domain	Predicted hydrolases of HD superfamily	2gz4A- 22- 52- 7e-08
5741338	NO	NO	Sulfolobus plasmid regulatory protein	NO	No
5741340	NO	NO	PaaX-like protein,Glycosyl transferase family, helical bundle domain	NO	No
5741347	NO	Hedgehog/DD-peptidase, zinc-binding motif Peptidase M15A, C-terminal	Peptidase M15	NO	1lbuA- 27-43- 2e-05
5741353	NO	NO	DUF2459, short chain dehydrogenase	NO	No
5741361	NO	NO	VRR-NUC domain	NO	No
5741365	NO	NO	Mediator complex subunit 3 fungal	NO	No
5741366	NO	NO	Minor capsid	NO	No

Table 13 Vibrio phage VP882

NCBI Gene ID	CDD BLAST	INTERPROSCAN	PFAM	COGS	Structures
5076227	NO	NO	Probable metal-binding protein (DUF2387) Zn-ribbon-containing, possibly nucleic-acid-binding protein (DUF2310)	NO	No
5076229	NO	No	Organic Anion Transporter Polypeptide (OATP) family	NO	No
5076232	NO	NO	Caenorhabditis protein of unknown function, DUF268	NO	No
5076244	NO	Phage DNA packaging Nu1 Winged helix-turn-helix transcription repressor DNA-binding	Phage DNA packaging protein Nu1	NO	1j9iA- 40-63- 5e-11
5076267	NO	NO	Vps23 core domain	NO	No
5076268	NO	NO	Asp/Glu/Hydantoin racemase , PsbP	NO	No

6. ACKNOWLEDGEMENT

We are thankful to Miss. Kimi Patel and Miss. Lekha Patel for their help and assistance in this work.

7. REFERENCES

1. A. Guidolin and P.A.Manning: Genetics of *Vibrio cholerae* and its bacteriophages. *Microbiol Rev* (1987), 51:285-298.

2. Alex, B., Lachlan, C., Richard, D., Robert, D. F., Volker, H., Sam, G.J., Ajay, K., Mhairi, M., Simon, M., Erik, L. L. S., David, J. S., Corin Y., Sean, R. E., (2004). The Pfam families' database. Nucleic Acids Research, Vol. 32, D138-D141.

3. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., Lipman, D. J., (1997). Gapped BLAST and PSI-BLAST: "*a new generation of protein database search programs"*. Nucleic Acids Res. 25 (17), 3389-402.

4. Aron, M. Bauer., John, B. A., Myra, K. D., Carol, D. S., Noreen, R. G., Marc, G., Luning, H., Siqian, H., David, I. H., John, D. J., Zhaoxi, K., Dmitri, K., Christopher, J. L.,Cynthia A. L., Chunlei, L., Fu, L., Shennan, L., Gabriele, H. M., Mikhail, M., James, S. S., Narmada, T., Roxanne, A. Y., Jodie, J. Y., Dachuan, Z., Stephen, H. B., (2006). CDD: "*a conserved domain database for interactive domain family analysis.* "Nucleic Acids Research, Vol. 35, D237–D240.

5. Azaro, M.A., and Landy, A. (2002) l integrase and the l Int family. In: *Mobile DNA II*. Craig, N.L. (ed.). Washington, DC: American Society for Microbiology Press, pp. 118–148

6. B. M. Davis and M. K. Waldor. "*Filamentous phages linked to virulence of Vibrio cholera"*. Curr. Opin. Microbiol. (2003), 6:35–42.

7. Basu,N., Kar,S. and Ghosh,R.K. "*Molecular analysis of filamentous phage VSK of Vibrio cholerae 0139*: "A possible clue to genetic transmission Unpublished

8. C. A. Kellogg, J.B. Rose, S.C. Jiang, J.M. Thurmond and J.H. Paul," *Genetic diversity of vibriophages isolated from marine environments around Florida and Hawaii,"* USA. *Mar Ecol Prog Ser* (1995), 120: 89–98.

9. Campos, J., Martinez, E., Izquierdo, Y. and Fando, R. VEJ {phi*}, "A novel filamentous phage of Vibrio cholerae able to transduce the cholera toxin genes*." Microbiology (Reading, Engl.) 156 (PT 1), 108-115 (2010)

10. Campos,J., Martinez,E., Suzarte,E., Rodriguez,B.L., Marrero,K., Silva,Y.K., Ledon,T.Y., Del Sol,R.E. and Fando,R.A. VGJphi: "*A Novel Lysogenic Filamentous Phage of Vibrio cholerae which Shares the Same Integration Site with CTXphi* " Unpublished

11. Cédric, N., Desmond, G. H., Jaap, H., (2000). T-coffee: "*a novel method for fast and accurate multiple sequence alignment*." J. Mol. Biol. 302, 205-217.
12. Chih-Chieh, C., Jenn-Kang, H., Jinn-Moon, Y., (2006). $(PS)^2$: "*protein structure prediction server Nucl*." Acids Res. 34, W152-W157.

13. Das,M., Bhowmick,T.S., Sarkar,B.L., Nair,G.B., Yamasaki,S. and Nandy,R.K. "*Complete genome sequence of lytic vibriophage N4 indicates close relativeness of T7 viral supergroup Unpublished*

14. Davis, B.M., and Waldor, M.K. "*(2000) CTXf contains a hybrid genome derived from tandemly integrated elements." Proc Natl Acad Sci USA* 97: 8572–8577.

15. Davis, B.M., and Waldor, M.K. (2002) '"*Mobile genetic elements and bacterial pathogenesis".* In: *Mobile DNA II*. Craig, N.L. (ed.). Washington, DC: American Society for Microbiology Press, pp. 1040–1059.

16. Dilip, G., Alankar, R., (2009). "*Computational Function and Structural Annotations for Hypothetical Proteins Bacillus anthracis.* Biofrontiers", 1, 27-36.

17. E. A. Jouravleva, G. A. McDonald, C. F. Garon, M. B. Finkelstein, and R. A. Finkelstein. "*Characterization and possible function of a new filamentous bacteriophage from Vibrio cholera".* Microbiology (1998), 144:315–324.

18. Edward, E., Gary, L. G., Osnat, H., John, M., John, O., Roberto, J. P., Linda, B., Delwood, R., Andrew, J. H., (2000). "*Biological function made crystal clear- annotation of hypothetical proteins via structural genomics."* Current Opinion in Biotechnology 11, 25-30.

19. Ehara, M., Nguyen, M.B., Nguyen,T.D., Ngo,C.T., Le,H.T., Nguyen,T.H. and Iwami,M. Integrated kappa phage genome Unpublished

20. Faruque,S.M., Bin Naser., Fujihara., Diraphat., Chowdhury., Kamruzzaman., Qadri., Yamasaki,S., Ghosh,R.K. and Mekalanos,J.J. Genomic sequence and receptor for the Vibrio cholerae phage KSF-1phi: "*evolutionary divergence among filamentous vibriophages mediating lateral gene transfer."* J. Bacteriol. 187 (12), 4095-4103 (2005)

21. H. Nakanishi, Y. Iida, K. Maeshima, T. Teramoto, Y. Hosaka and M. Ozaki. "*Isolation and properties of bacteriophages of Vibrio parahaemolyticus*. "*Biken J* (1966), 9: 149– 157.

22. Hardies, S.C., Comeau, A.M., Serwer, P. and Suttle, C.A. "*The complete sequence of marine bacteriophage VpV262 infecting vibrio parahaemolyticus indicates that an ancestral component of a T7 viral supergroup is widespread in the marine environment";* Virology 310 (2), 359-371 (2003)

23. Honma, Y., Ikema, M., Toma, C., Ehara, M. and Iwanaga, M." *Molecular analysis of a filamentous phage (fsl) of Vibrio cholerae O139*;" Biochim. Biophys. Acta 1362 (2-3), 109-115 (1997)

24. J. Campos, E. Martinez, E. Suzarte, B. L. Rodriguez, K. Marrero, Y. Silva, T. Ledo´n, R. del Sol, and R. Fando. VGJ_, "*a novel filamentous phage of Vibrio cholerae, integrates into the same chromosomal site as CTX"*_. J. Bacteriol. (2003), 185:5685–5696.

25. J. Campos, E. Martinez, K. Marrero, Y. Silva, B. L. Rodriguez, E. Suzarte, T. Ledon, and R. Fando*.' Novel type of specialized transduction for CTX_ or its satellite phage RS1 mediated by filamentous phage VGJ_ in Vibrio cholera"*. J. Bacteriol. (2003), 185:7231–7240.

26. J.A. Baross, J. Liston, and R.Y. Morita. "*Some implications of genetic exchange among marine vibrios, including Vibrio parahaemolyticus, naturally occurring in the Pacific oyster".* In *International Symposium on* Vibrio parahaemolyticus. Fujino, T., Sakaguchi, G., Sakazaki, R., and Takeda, Y. (eds). Tokyo, Japan: Saikon Publishing, (1974), pp. 129–137.

27. K. Moebus and H. Nattkemper. '*Bacteriophage sensitivity patterns among bacteria isolated from marine waters. Helgolander Meeresunters.*"(1981), 34: 375-385

28. K. Moebus. Ecology of manne bacteriophages. In: Goyal, S. M., Gerba. C. P, Bitton, G. (eds.) Phage ecology. John Wiley & Sons. New York, (1987), p. 137-156

29. Kapfhammer, D., Blass, J., Evers, S. and Reidl, J. Vibrio cholerae phage K139: "*complete genome sequence and comparative genomics of related phages";* J. Bacteriol. 184 (23), 6592-6601 (2002)

30. M. Ikema and Y. Honma*. "A novel filamentous phage, fs2, of Vibriocholerae O139*." Microbiology (1998), 144:1901–1906.

31. M. K. Waldor and J. J. Mekalanos"*. Lysogenic conversion by a filamentous phage encoding cholera toxin".* Science (1996), 272:1910–1914.

32. Miller,E.S., Heidelberg,J.F., Eisen,J.A., Nelson,W.C., Durkin,A.S., Ciecko,A., Feldblyum,T.V., White,O., Paulsen,I.T., Nierman,W.C., Lee,J., Szczypinski,B. and Fraser,C.M. "*Complete genome sequence of the broad-host-range vibriophage KVP40: comparative genomics of a T4-related bacteriophage"* J. Bacteriol. 185 (17), 5220-5233 (2003)

33. Moyer, K.E., Kimsey, H.H., and Waldor, M.K. (2001) "*Evidence for a rolling-circle mechanism of phage DNA synthesis from both replicative and integrated forms of CTXf. Mol Microbiol 41*:" 311–323.

34. Ochman, H., Lawrence, J.G., and Groisman, E.A. (2000)" *Lateral gene transfer and the nature of bacterial innovation*." *Nature* 405: 299–304.

35. Roman, L. T., Michael, Y., Galperin, Darren A. Natale, Eugene V. Koonin (2000). "*The COG database: a tool for genome –scale analysis of protein functions and evolution. Nucleic Acid Research*." 28, 33-36.

36. S. Kar, R. K. Ghosh, A. N. Ghosh, and A. Ghosh." *Integration of the DNA of a novel filamentous bacteriophage VSK from Vibrio cholerae O139 into the host chromosomal DNA. FEMS Microbiol*." Lett. (1996), 145:17–22.

37. S. M. Faruque, I. Bin Naser, K. Fujihara, P. Diraphat, N. Chowdhury, M. Kamruzzaman, F. Qadri, S. Yamasaki, A. N. Ghosh, and J. J. Mekalanos. "*Genomic sequence and receptor for the Vibrio cholerae phage KSF-1_: evolutionary divergence among filamentous vibriophages mediating lateral gene transfer"*. J. Bacteriol. (2005), 187:4095–4103.

38. S. Matsuzaki, S. Tanaka, T. Koga, and T. Kawata. "*A broad-host-range vibriophage, KVP40, isolated from sea wate*r". Microbiol. Immunol. (1992), 36:93–97.

39. S. Matsuzaki, T. Inoue, M. Kuroda, S. Kimura and S. Tanaka. "*Cloning and sequencing of major capsid protein (mcp) gene of a vibriophage, KVP20, possibly related to Teven coliphages" Gene* (1998), 222: 25–30.

40. Schaffer, A. A., Aravind, L., Madden, T. L., Shavirin, S. Spouge, J. L., Wolf, Y. I., Koonin, E. V., Altschul, S. F., (2001*)." Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements".* Nucleic Acids Res. 29(14), 2994-3005.

41. T. Koga, S. Toyoshima and T. Kawata. "*Morphological varieties and host ranges of Vibrio parahaemolyticus bacteriophages isolated from seawater". Appl Environ Microbiol* (1982), 44: 466–470.

42. W.I. Lencer and B. Tsai. "*The intracellular voyage of cholera toxin: going retro.* "Trends Biochem Sci (2003), 28: 639– 645.

43. Wang,D., Kan,B., Li,Y., Liu,Z., Gao,S., Liu,Y., Liang,W., Zhang,L., Yan,M., Li,W., Liu,G., Liu,Y., Li,J., Diao,B., Zhu,Z. and Qiu,H. Vibrio cholerae phage VP2 complete genome Unpublished

44. Wendy, B. et al., (2000). "*The EMBL Nucleotide Sequence Database".* Nucleic Acid Research. 28, 19-23

45. Zafer, A., Yucel, A., Mark, B., (2006). "*Protein secondary structure prediction for a single-sequence using hidden semi-Markov models, BMC Bioinformatics*, "7, 178.

46. Zdobnov, E. M., Rolf, A., (2001*)." Interproscan- an integration platform for the signatures recognition methods in InterPro. Bioinformatics* "17, 847-848

Multiple Features Based Two-stage Hybrid Classifier Ensembles for Subcellular Phenotype Images Classification

Bailing Zhang bailing.zhang@xjtlu.edu.cn
Department of Computer Science and Software Engineering,
Xi'an Jiaotong-Liverpool University,
Suzhou,215123, P.R.Chin

Tuan D. Pham t.pham@adfa.edu.au
School of Engineering and Information Technology,
The University of New South Wales,
Canberra, ACT 2600, Australia.

Abstract

Subcellular localization is a key functional characteristic of proteins. As an interesting ``bio-image informatics'' application, an automatic, reliable and efficient prediction system for protein subcellular localization can be used for establishing knowledge of the spatial distribution of proteins within living cells and permits to screen systems for drug discovery or for early diagnosis of a disease. In this paper, we propose a two-stage multiple classifier system to improve classification reliability by introducing rejection option. The system is built as a cascade of two classifier ensembles. The first ensemble consists of set of binary SVMs which generalizes to learn a general classification rule and the second ensemble, which also include three distinct classifiers, focus on the exceptions rejected by the rule. A new way to induce diversity for the classifier ensembles is proposed by designing classifiers that are based on descriptions of different feature patterns. In addition to the Subcellular Location Features (SLF) generally adopted in earlier researches, three well-known texture feature descriptions have been applied to cell phenotype images, which are the local binary patterns (LBP), Gabor filtering and Gray Level Co-occurrence Matrix (GLCM). The different texture feature sets can provide sufficient diversity among base classifiers, which is known as a necessary condition for improvement in ensemble performance. Using the public benchmark 2D HeLa cell images, a high classification accuracy 96% is obtained with rejection rate 21% from the proposed system by taking advantages of the complementary strengths of feature construction and majority-voting based classifiers' decision fusions.

Keywords: subcellular phenotype images classification; hybrid classifier; local binary patterns; Gabor filtering; Gray level co-occurrence matrix; support vector machine; multiple layer perceptron; random forest

1. INTRODUCTION

Eukaryotic cells have a number of subcompartments termed organelles, each of which contains a unique localization of proteins and hence different biochemical properties. Determining a protein's location within a cell is critical to understanding its function and to build models that capture and simulate cell behaviors. It has been shown that mislocalization of proteins correlates with several diseases that range from metabolic disorders to cancer [1], thus knowledge of the location of all proteins will be essential for early diagnosis of disease and/or monitoring of therapeutic effectiveness of drugs. Given that mammalian cells are believed to express tens of thousands of proteins, a comprehensive analysis of protein locations requires the development of an automated massive analysis method. If such analyses can be converted into high throughput ``location proteomics'' assays, the resulting information would help us to understand the functions, properties and distribution of proteins in cells, and how a protein changes its characteristics in response to drugs, diseases and various stages of the cell cycle.

The most widely used method for determining protein subcellular location is fluorescence microscopy, which combines fluorescence detection with high-powered digital microscopy. Advances in fluorescent probe chemistry, protein chemistry, and imaging techniques have made fluorescence microscopy a valuable method for determining protein subcellular locations [2,3]. Over the past decade, there has been much progress in the classification of subcellular protein location patterns from fluorescence microscope images. The pioneering contributions to this problem should be attributed to Murphy and his colleagues [4-8]. Machine learning methods such as artificial neural networks and Support Vector Machine (SVM) have been utilized for the predictive task of protein localization in conjunction with various feature extraction methods from fluorescence microscopy images. Most of the proposed approaches employed feature set which consist of different combinations of morphological, edge, texture, geometric, moment and wavelet features. For example, [5] used images of ten different subcellular patterns to train a neural network classifier, which has been shown to correctly recognize an average of 83% of the patterns.

In previous studies of subcellular phenotype images classification, classification accuracy was the only pursuit, aiming to produce a classifier with the smallest error rate possible. In many applications, however, reject option for classifiers by allowing for an extra decision expressing doubt is important. For instance, in early diagnosis of disease or monitoring of therapeutic effectiveness of drugs, it is more important to be able to reject an example of subcellular phenotype image when there is no sufficiently high degree of accuracy, since the consequences of misclassification are severe and scientific expertise is required to exert control over the accuracy of the classifier thus making reliable determination. Therefore, we are motivated to investigate the option of classification scheme with rejection paradigm to meet the desirable functionality of automated subcellular phenotype images classification whereby the system generates decisions with confidence larger than some prescribed threshold and transfers the decision on cases with lower confidence to a human expert. For the 2D HeLa images [5,6], evidence from many published works and our own extensive experiments confirmed that no single method of classification could achieve high classification accuracy for all localizations. It has become a consensus in machine learning community that an integrative approach by combining multiple learning systems often offer higher and more robust classification accuracy than a single learning system [19]. The so-called ensemble system that combines the outputs of several diverse classifiers or experts has been broadly applied and proven an efficient approach to improve the performance of recognition systems. The intuition is that the diversity in the classifiers allows different decision boundaries to be generated, which can be implemented by using different learning algorithms corresponding to different errors or by using different representations of the same input to make different features apparent and provide supplementary information.

As a typical multi-class classification issue, subcellular phenotype images classification involves two interweaved parts: feature representation and classification. Many of the off-the-shelf standard classifiers such as multiple layer perceptron can be directly applied together with

different possible feature sets which are potentially useful for separating different classes of subcellular phenotype [32]. However, a multi-class subcellular phenotype images dataset is often featured with large intra-class variations and inter-class similarities, which poses serious problems for simultaneous multi-class separation using the standard classifiers. On the other hand, it is almost impossible to find a feature set that is universally informative for separating all classes simultaneously. A better alternative solution to the problem, therefore, is to train different classifiers on distinct feature sets to fit the different characteristics. In our study, three kind of texture feature representations were considered, together with the Subcellular Location Features (SLF) [5,7]. The three texture feature expressions are the local binary patterns (LBP) [12], Gabor filtering [17] and Gray Level Co-occurrence Matrix (GLCM) [18] . The LBP operator has been proved a powerful means of texture description, which is relatively invariant with respect to changes in illumination and image rotation, and computationally simple [13, 14]. Gabor filter is another widely adopted operator for texture properties description and has been shown to be very efficient in many applications [17]. The Gray Level Co-occurrence Matrix (GLCM) method is characterized by its capability of extracting second order statistical texture features when considering the spatial relationship of pixels and has been proved to be a promising method in many image analysis tasks. These kinds of texture features alone might, however, have limited power in describing the complex features from microscopy images related to the subcellular protein location patterns. This again strengthens our avocations to propose a two-stage classifiers system which cater for a design-based method to fuse the features from LBP, Gabor filter, GLCM and SLF in order to obtain an improved classification performance.

Our work follows the hybrid classification paradigm, which combines classifiers to yield more accurate recognition rates when different classifiers contributes partially with different features. Unlike relative works that combine different base classifiers (trained with same samples) for image recognition systems, we use an effective approach to utilize complementary texture information and provide sufficient diversity among base classifiers of ensemble. With the 2D HeLa images, a sample can be either classified or rejected. The objective of reject option is to improve classification reliability and leave the control of classification accuracy to human expert. Comparing with some earlier cascading classifier paradigms, our proposed system is composed of different classifiers each specializes with different set of features. In our implementation, one-vs-all SVMs are employed in the first stage to obtain high accuracy for easier inputs and reject a subset of class assignments which is harder or ambiguous. A second stage classifier ensemble consists of three different kind of multi-class classifiers working in parallel (random forest, neural networks and support vector machines) and the final decision is based on the majority voting for the final combination.

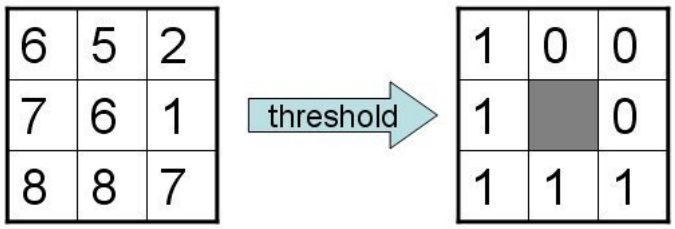
The paper is organized as follows. In Section 2, we introduce feature descriptions, including three texture descriptors LBP, Gabor filter and GLCM, together with the Subcellular Location Features (SLF). In Section 3, we elaborate the details of the proposed two-stage hybrid classification system. Experiments using the 2D HeLa images are provided in Section 4 and conclusion is outlined in Section 5.

2. FEATURE DESCRIPTIONS FOR CELL PHENOTYPE IMAGES

In order to automated analyse and classify microscopic cellular images, some kind of features have to be extracted to express the statistical characteristics in the image. And given two sets of sub-cellular localization images under differing experimental conditions, an efficient image feature can be used to evaluate if there is a statistically significant difference, even to the extent that visually indistinguishable images of distinct localizations may be differentiated [4] .The feature sets proposed in the literature include, for instance, morphological data of binary image structures, Zernike moments and edge information [5,6]. Use of a single technique for the extraction of diverse features in an image usually exhibits limited information description. Features extracted using different techniques can be combined in an attempt to enhance their description capability.

2.1 Local Binary Pattern

Local Binary Pattern (LBP) operator was introduced as a texture descriptor for summarizing local gray-level structure [12]. LBP labels pixels of an image by taking a local neighborhood around each pixel into account, thresholding the pixels of the neighborhood at the value of the central pixel and then using the resulting binary-valued image patch as a local image descriptor. In another word, the operator assigns a binary code of 0 and 1 to each neighbor of the neighborhoods. The binary code of each pixel in the case of 3x3 neighborhoods would be a binary code of 8 bits and by a single scan through the image for each pixel the LBP codes of the entire image can be calculated. Figure 1 shows an example of an LPB operator utilizing 3x3 neighborhoods.



Binary code = **11110001**
LBP = 1 + 16 +32 + 64 + 128 = **241**

Figure 1. Illustration of the basic LBP operator.

Formally, the LBP operator takes the form

$$\mathrm{LBP}(x_c, y_c) = \sum_{n=0}^{7} s(i_n - i_c)2^n$$

where in this case n runs over the 8 neighbors of the central pixel c, i_c and i_n are the gray-level values at c and n, and s(u) is 1 if u ≥0 and 0 otherwise.

An useful extension to the original LBP operator is the so-called uniform patterns [12]. An LBP is ``uniform'' if it contains at most two bitwise transitions from 0 to 1 or vice versa when the binary string is considered circular. For example, 11100001 is a uniform pattern, whereas 11110101 is a non-uniform pattern. The uniform LBP describes those structures which contain at most two bitwise (0 to 1 or 1 to 0) transitions. Uniformity is an important concept in the LBP methodology, representing important structural features such as edges, spots and corners. Ojala et al. [12] observed that although only 58 of the 256 8-bit patterns are uniform, nearly 90 percent of all observed image neighbourhoods are uniform. We use the notation $LBP^{u}_{P,R}$ for the uniform LBP operator. $LBP^{u}_{P,R}$ means using the LBP operator in a neighborhood of P sampling points on a circle of radius R. The superscript u stands for using uniform patterns and labeling all remaining patterns with a single label. The number of labels for a neighbourhood of 8 pixels is 256 for standard LBP and 59 for $LBP^{u}_{8,1}$.

A common practice to apply the LBP coding over an image is by using the histogram of the labels, where a 256-bin histogram represents the texture description of the image and each bin can be regarded as a micro-pattern. Local primitives which are coded by these bins include different types of curved edges, spots, flat areas, etc. The distribution of these patterns represents the whole structure of the texture. The number of patterns in an LBP histogram can be reduced by only using uniform patterns without losing much information. There are totally 58 different uniform patterns at 8-bit LBP representation and the remaining patterns can be assigned in one non-uniform binary number, thus representing the texture structure with a 59-bin histogram.

LBP scheme has been extensively applied in face recognition, face detection and facial expression recognition with excellent success, outperforming the state-of-the-art methods [13].

The methodology can be directly extended to microscopy image representations as outlined in the following. First, a microscopy image is divided into M small no-overlapping rectangular blocks R_0, R_1, …, R_M. On each block, the histogram of local binary patterns is calculated. The procedure can be illustrated by Figure 2. The LBP histograms extracted from each block are then concatenated into a single, spatially enhanced feature histogram defined as:

$$H_{ij} = \sum_{x,y} I(f_l(x,y) = i) \qquad i = 0, \ldots, L-1,\ j = 0, \ldots, M-1$$

where L is the number of different labels produced by the LBP operator and I(A) is 1 if A is true and 0 otherwise. The extracted feature histogram describes the local texture and global shape of microscopy images.

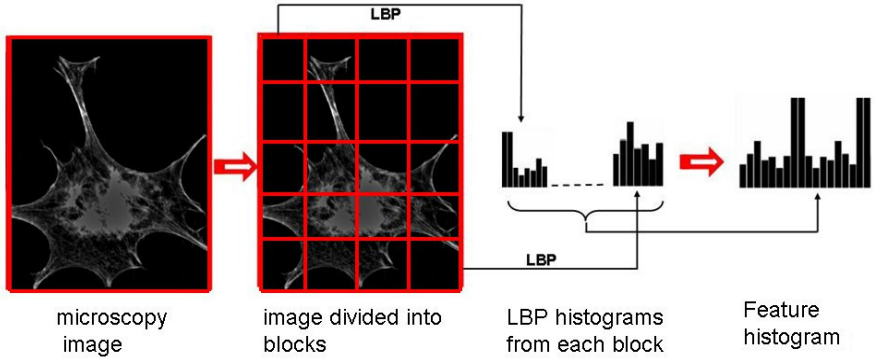


FIGURE 2: Feature extraction diagram for image recognition with local binary patterns.

LBP has been proved being a good texture descriptor with high extra-class variance and low intra-class variance. Recently, a number of variants of LBP have been proposed [15]. In [16], a completed modeling of the local binary pattern operator is proposed and an associated completed LBP (CLBP) scheme is developed for texture classification. In this scheme, a local region is represented by its center pixel and a local difference sign-magnitude transform. And the center pixels represent the image gray level and they are converted into a binary code by global thresholding. For many applications like face recognition, CLBP can offer better performance.

2.2 Gabor Based Texture Features

Gabor filters [17] have been used extensively to extract texture features for different image processing tasks. Image representation using Gabor filter responses minimises the joint space-frequency uncertainty. The filters are orientation- and scale-tunable edge and line detectors. Statistics of these local features in a region relate to the underlying texture information. The convolution kernel of Gabor filter is a product of a Gaussian and a cosine function, which can be characterized by a preferred orientation and a preferred spatial frequency:

$$g_{\lambda,\theta,\varphi}(x,y) = \exp\left(-\frac{(x'^2 + \gamma y'^2}{2\sigma^2}\right)\cos\left(2\pi\frac{x'}{\lambda} + \varphi\right)$$

where

$$x' = x\cos\theta + y\sin\theta$$
$$y' = -x\sin\theta + y\cos\theta$$

The standard deviation σ determines the effective size of the Gaussian signal. The eccentricity of the convolution kernel g is determined by the parameter λ, called the spatial aspect ratio. λ determines the frequency (wavelength) of the cosine. θ determines the direction of the cosine function and finally, φ is the phase offset.

There exists several useful properties with Gabor functions which are important for texture analysis. Gabor function optimally concentrate both in space and space-frequency domain by the smallest time-bandwidth product of the Gaussian function. Due to the ability to tune a Gabor filter to specific spatial frequency and orientation, and achieve both localization in the spatial and the spatial-frequency domains, textures can be encoded into multiple channels each having narrow spatial frequency and orientation. The local information regarding the texture elements is described by the orientations and frequencies of the sinusoidal grating and the global properties are captured by the Gaussian envelope of the Gabor function. Hence the local and global properties of the texture regions can be simultaneously represented by making use of the Gabor filters.

Typically, an image is filtered with a set of Gabor filters of different preferred orientations and spatial frequencies that cover appropriately the spatial frequency domain, and the features obtained form a feature vector that is further used for classification. Given an image I(x,y), its Gabor wavelet transform is defined as

$$W_{mn}(x,y) = \int I(x_1,y_1) g^*_{mn}(x-x_1, y-y_1) dx_1 dy_1$$

where * indicates the complex conjugate. With assumption of spatially homogeneous local texture regions, the mean μ_{mn} and standard deviation σ_{mn} of the magnitude of transform coefficients can be used to represent the regions [17]. A feature vector f (texture representation) is thus created using μ_{mn} and σ_{mn} as the feature components.

2.3 Gray Level Co-occurrence Matrices

Gray level co-occurrence matrix (GLCM) proposed by Haralick [18] is another common texture analysis method which estimates image properties related to second-order statistics. GLCM matrix is defined over an image to be the distribution of co-occurring values at a given offset. Mathematically, a co-occurrence matrix C is defined over an nxm image I, parameterized by an offset

$$C_{\Delta x,\Delta y}(i,j) = \sum_{p=1}^{n}\sum_{q=1}^{m} \begin{cases} 1, & \text{if } I(p,q)=i \text{ and } I(p+\Delta\, x, q+\Delta\, y)=j \\ 0, & \text{otherwise} \end{cases}$$

Note that the ($\triangle$x, $\triangle$y) parameterization makes the co-occurrence matrix sensitive to rotation. An offset vector can be chosen such that a rotation of the image not equal to 180 degrees will result in a different co-occurrence distribution for the same (rotated) image.

In order to estimate the similarity between different GLCM matrices, Haralick proposed 14 statistical features extracted from them [18]. To reduce the computational complexity, only some of these features will be selected. The 4 most relevant features that are widely used in literature include: (1)Energy, which is a measure of textural uniformity of an image and reaches its highest value when gray level distribution has either a constant or a periodic form; (2) Entropy, which measures the disorder of an image and achieves its largest value when all elements in C matrix are equal; (3) Contrast, which is a difference moment of the C and measures the amount of local variations in an image; (4) Inverse difference moment (IDM) that measures image homogeneity.

2.4 Subcellular Location Features (SLF)

Murphy group has developed and published several sets of informative features, termed Subcellular Location Features (SLFs), that describe protein subcellular location patterns in 2D fluorescence microscope images [5-7]. There are three major subsets of features. The first set is 49 Zernike moment features through order 12, which are calculated from the moments of each image relative to the Zernike polynomials, an orthogonal basis set defined on the unit circle. The second set is 13 Haralick texture features [18], which is related to intuitive descriptions of image texture, such as coarseness, complexity and isotropy. The third set of 22 features was derived from morphological and geometric analysis that correspond better to the terms used by biologists,

including the number of objects, the ratio of the size of the largest object to the smallest object, the average distance of an object from the center of fluorescence, and the fraction of above-threshold pixels along an edge et al. Each cell in the dataset is thus represented by a SLF feature vector x of length d = 84. Though SLF includes a much simplified Haralick texture features, we still applied GLCM analysis in a general scenario by specifying the different distance between the pixel of interest and its neighbor and including more statistical measurements as introduced in last subsection.

3. TWO-STAGE HYBRID CLASSIFICATION ENSEMBLES

After feature extraction, a statistical model needs to be learned from data that accurately associates image features with predefined phenotype classes. Some supervised learning algorithms such neural networks, k-nearest neighbor algorithm and SVM [5-8] have been applied to solve this problem. In pattern recognition systems, it has been proven that ensemble of classifiers have the potential to improve classification performance. How to combine multiple classifiers has been studied for decades, with a number of successful methods proposed in the literature [19]. The most popular method for creating an ensemble classifier is to build multiple parallel classifiers, and then to combine their outputs according certain decision fusion strategy. Alternatively, serial architecture can be adopted with different classifiers arranged in cascade and the output of each classifier is the input to the classifier of the next stage of the cascade.

Our approach is based on a hybrid topology that combine parallel and serial schemes. The idea is motivated by a human category learning theory rule-plus-exception model (RULEX) proposed in [20]. According to RULEX, people learn to classify objects by forming simple logical rules and remembering occasional exceptions to those rules. In machine learning, many off-the-shelf methods like support vector machine (SVM) and multi-layer perceptron (MLP) are able to approximate the Bayes optimal discriminant function, which is equivalent to discover the knowledge or patterns hidden in the dataset. Such a knowledge can be represented in terms of a set of rules underlying most of the training examples [22]. A rule consists of an antecedent (a set of attribute values) and a consequent (class):

$$\text{IF} < attrib = value > \text{AND} \cdots \text{AND} < attrib = value >$$
$$\text{THEN} < class > .$$

It is not realistic to expect such a rule to explain all of the data. The examples which are failed to be explained should be considered as exceptions and processed with a rejection option separately. For many real-world applications, such a rejection option is important to satisfy the classification constraints and many multi-stage classifier architectures have been proposed to automatically treating the rejects [23, 25 , 26] .

Extending from the previous works, we proposed a two-stage hybrid classifier ensemble in which a second classifier ensemble is concatenated to the first ensemble. At all stages, a pattern can be either classified or rejected. Rejected patterns are fed into the next stage. The overall system can be illustrated in Figure 3, which shows that second stage need only operate on the surviving inputs from the previous stage.

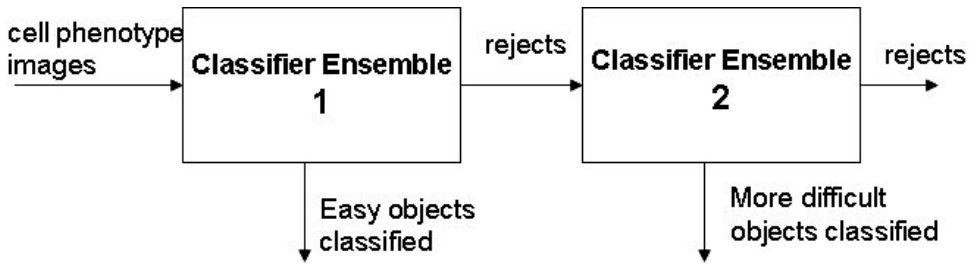


FIGURE 3: Illustration of the overall system which is a cascade of classifier ensembles. Samples rejected at first stage are passed on to second stage during classification.

The major issue for designing the above hybrid classification system is to decide when a pattern is covered by the rule and should be learned by the first classifier ensemble and when it is an exception and should be learned by the second classifier ensemble. The reject option has been formalized in the context of statistical pattern recognition, under the minimum risk theory [31, 23]. It consists in withholding the automatic classification of a pattern, if the decision is considered not sufficiently reliable. Intuitively, objects should be rejected when the confidence in their classification is too low. The standard approach to rejection in classification is to estimate the class posteriors, and to reject the most unreliable objects, that is, the objects that have the lowest class posterior probabilities [24, 23] . As the posteriors sum to 1, there will be complete ambiguity if all posteriors are equal to 1/d with d classes and complete certainty when one posterior is equal to 1 and all others equal to 0.

To simplify the design of the first stage ensemble with appropriate posteriors estimation, we can decompose the multi-label classification problems with k classes into k independent two-class problems, each one consisting in deciding whether an object should be assigned or not to the corresponding class. This is the idea of the *one-versus-all* approach to divide the classes into two groups each time, with one group consisting of a single class and the other group consisting of samples in all the other classes. In other words, a set of k independent binary classifiers are constructed for k classes where the i^{th} classifier is trained to separate samples belonging to class i from all others. Then the multiclass classification is carried out according to the maximal output of the binary classifiers. Though there are many candidates to implement such a scheme, we choose to apply SVMs due to their ability to map features into arbitrarily complex spatial dimensions to find the optimal margin of separation. To estimate class posteriors from SVM's outputs, a mapping can be implemented using the following sigmoid function [28]:

$$P(y = +1|\mathbf{x}) = \frac{1}{1 + \exp(\mathrm{a}\rho(\mathbf{x}) + \mathrm{b})}$$

where the class labels are denoted as y = +1, -1, while a and b are constant terms to be defined on the basis of sample data. Such a method provides estimates of the posterior probabilities that are monotonic functions of the output ρ(x) of a SVM. This implies that Chow's rule applied to such estimates is equivalent to the rejection rule obtained by directly applying a reject threshold on the absolute value of the outputρ(x) [27].

In our scheme, M binary SVM classifiers are constructed for M different image features. The ith SVM output function P_i is trained taking the examples from i^{th} class as positive and the examples from all other classes as negative. In another word, each binary SVM classifier in the ensemble was trained to act as a class label detector, outputting a positive response if its label is present and a negative response otherwise [21]. So, for example, a binary SVM trained as a ``Nuclei detector'' would classify between cell phenotypes which are Nuclei and not Nuclei. For a new example x, the corresponding SVM assigns it to the class with the largest value of P_i following

$$Class = \arg\max \ P_i, \quad i = 1, \ldots, n$$

where Pi is the signed confidence measure of the ith SVM classifier. The maximum confidence rule with $P(Y_i = 1)$ is used as the confidence measure.

We assume that k classifier ensemble or experts are deployed in the first stage, and that for each input sample, each expert produces a unique decision regarding the identity of the sample. This identity could be one of the allowable classes, or a rejection when no such identity is considered possible. In the event that the decision can contain multiple choices, the top choice would be selected [29]. In combining the decisions of the k experts, the sample is assigned the class for which there is a consensus or when at least t of the experts are agreed on the identity, where

$$t = \begin{cases} \frac{k}{2} + 1 & \text{if } k \text{ is even} \\ \frac{k+1}{2} & \text{if } k \text{ is odd} \end{cases}$$

Otherwise, the sample is rejected. Since there can be more than two classes, the combined decision is correct when a majority of the experts are correct, but wrong when a majority of the decisions are wrong and they agree. A rejection is considered neither correct nor wrong, so it is equivalent to a neutral position or an abstention. Figure 2 further explains the process chart of the stage 1 classifier ensemble.

It is worthy to emphasize that different representations of same set of images were considered for different ``expert", which allow a single expert to take decision about class memberships and thus have different probable decisions. This presents a way to use fusion to have more authenticated decisions by considering many representations of set of patterns.

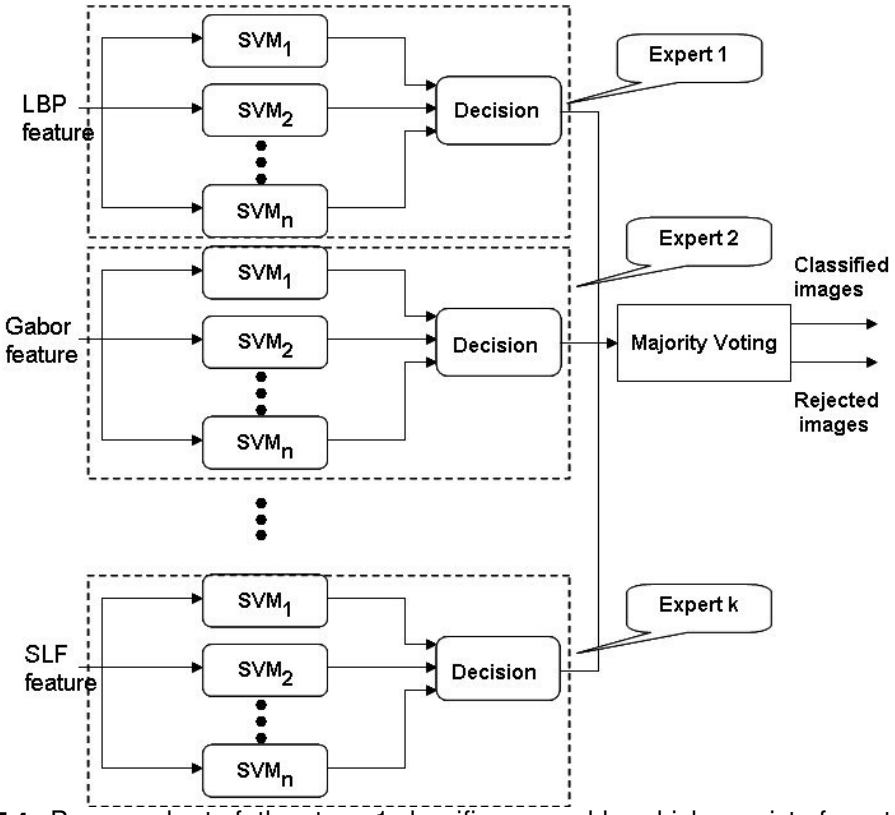


FIGURE 4: Process chart of the stage 1 classifier ensemble, which consist of a set of binary SVMs with high rejection rate.

The set of rejected patterns found by the first stage classifier ensemble will be handled by next stage ensemble, which is a multiple classifier combination with the aim of overcoming the limitations of individual classifiers. In our design, diversity is achieved by choosing classifiers differing in feature representation, architecture and learning algorithm in order to bring complementary classification behavior. In stage 2, the multi-class classification is handled directly by three individual classifiers, including neural network (NN), support vector machine (SVM), and Random Forest classifier [34], which are simultaneously trained with stage 1 ensemble. The three classifiers are of different types: NN classifier is weight-based, SVM classifier is distance or margin based, and Random Forest is rule based. Using different types of classifiers as the constituent classifiers in classifier fusion is one of our design strategies in obtaining necessary diversity, thus achieving improved performance.

The neural network classifier is a 2-layer feed-forward network. It has one hidden layer with a few hidden neurons and has 10 output nodes, each representing a class label. The activation functions for hidden and output nodes are logistic sigmoid function and linear function, respectively. Support Vector Machines (SVM) is a developed learning system originated from the statistical learning theory [30]. One distinction between SVM and many other learning systems is that its decision surface is an optimal hyperplane in a high dimensional feature space. The optimal hyperplane is defined as the one with the maximal margin of separation between positive and negative examples. Designing SVM classifiers includes selecting the proper kernel function and the corresponding kernel parameters and choosing proper C value.

The histogram intersection, $k_{HI}(h_a, h_b) = \sum_{i=1}^{n} \min(h_a(i), h_b(i)),$ is often used as a measurement of similarity between histograms h_a and h_b, and because it is positive definite, it can be used as a kernel for discriminative classification using SVMs. Recently, intersection kernel SVMs have been shown to be successful for detection and recognition [33].

Traditional decision tree classifiers are presented in a binary tree structure constructed by repeatedly splitting the data subsets into two descendant subsets. Each terminal subset is assigned a class label and the resulting partition of the dataset corresponds to the classifier. A random forest (RF) classifier [34] consists of many decision trees and outputs the class that is the mode of the classes output by individual trees. The RF algorithm combines ``bagging'' idea to construct a collection of decision trees with controlled variations. There are a number of advantages of RF classifiers, including: (1). it can efficiently handle high dimensional data; (2) it can simultaneously estimates the importance of variables in determining classification; (3). It maintains accuracy when a large proportion of the data are missing.

The last step of the second ensemble is to combine the above base models to give final decision. There are different types of voting systems, the frequently used ones are simple voting and weighted voting [29]. Simple voting, also called majority voting and select all majority (SAM) , considers each component classifier as an equally weighted vote. The classifier that has the largest amount of votes is chosen as the final classification scheme. In weighted voting schemes, each vote receives a weight, which is usually proportional to the estimated generalization performance of the corresponding component classifier. Weighted voting schemes usually give better performance than simple voting. In our study, however, we only experimented with the simple voting.

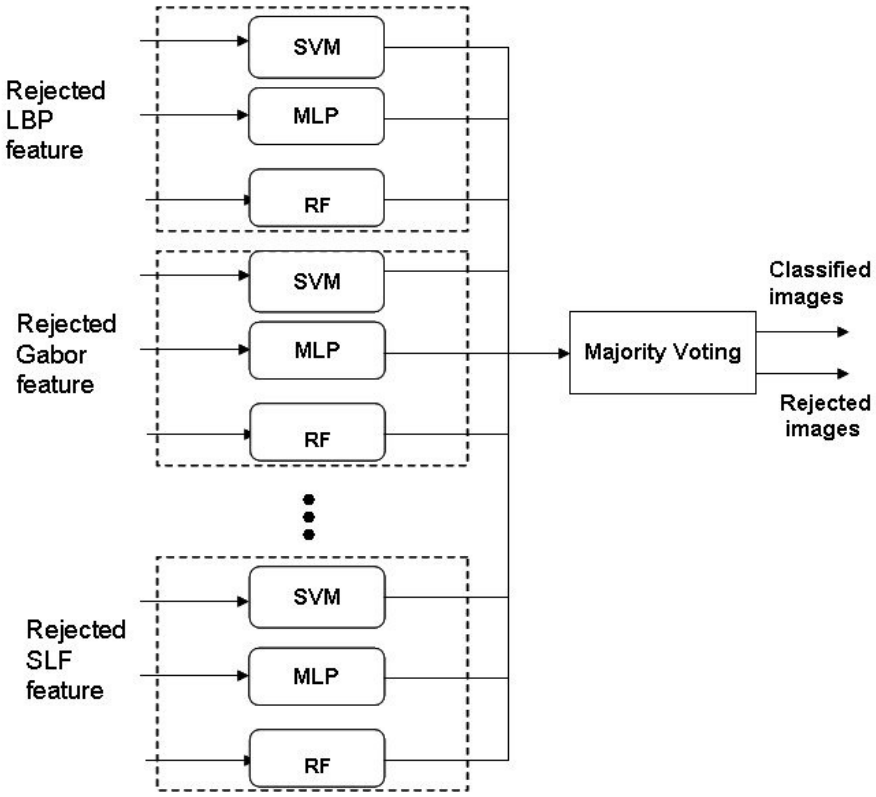


FIGURE 5: Illustration of the stage 2 classifier ensemble which consist of a set of binary SVMs with high rejection rate.

4. EXPERIMENTS

The dataset used for evaluating the system is the 2D HeLa dataset, a collections of HeLa cell immunofluorescence images containing 10 distinct subcellular location patterns [5,6]. The subcellular location patterns in these collections include endoplasmic reticulum (ER), the Golgi complex, lysosomes, mitochondria, nucleoli, actin microfilaments, endosomes, microtubules, and nuclear DNA. The 2D HeLa image dataset is composed of 862 single-cell images, each with size 382x512. Sample images for each class are illustrated in Figure 6. The 2D HeLa image datasets have been used as benchmark for automatically identifying sub-cellular organelles [9-11]. A good verifiable performance for 2D HeLa image classification is currently 91.5% [8], by including a set of multi-resolution features. The best published accuracy 97.5% was recently reported in [9], for which we could not confirm from our own experiments.

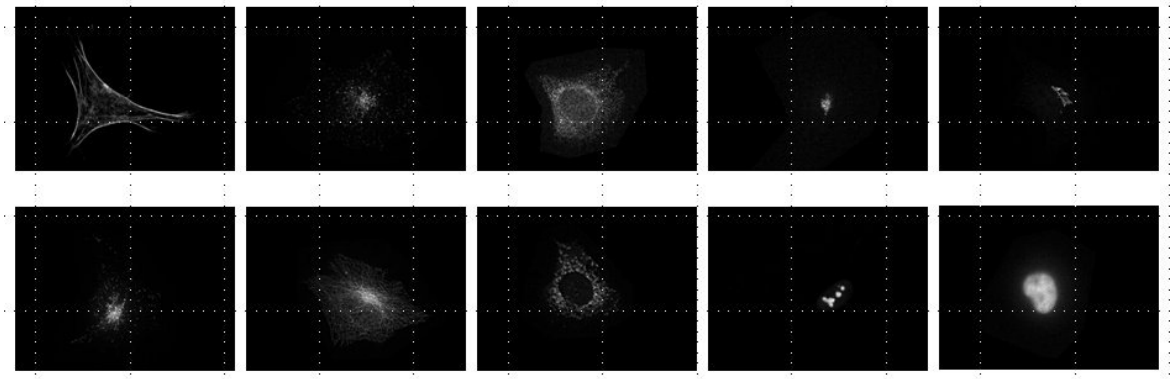


FIGURE 6: Sample 2D HeLa images

As elaborated in Section 2, we are interested in those numerical features that are generally applied in computer vision to describe the pattern in the images. Regarding the LBP feature, a 59-label $LBP^{u}_{8,1}$ operator was used as most of the texture information is contained in the uniform patterns. Specifically, $LBP^{u}_{8,1}$ operator is applied to non-overlapping image subregions to form a concatenated histogram. The performance of LBP representation is not sensitive to the subregion divisions, which do not need to be of the same size or cover the whole image. It is also quite robust with respect to the selection of parameters when looking for the optimal window size. Changes in the parameters may cause big differences in the length of the feature vector, but the overall performance is not necessarily affected significantly. Therefore, in all the experiments we fixed a subregion (window) size 95x102 for the HeLa images, yielding LBP feature vector with length 4x5x59 =1180 . As a comparison, we also applied a newly published variant of LBP operator, called Complete LBP (CLBP for short) [16]. A problem of CLBP is its much higher dimension, which is 2400 with a much larger subregion (size 125 x128) and parameters radius=3 and neighborhood =8.

The Gabor feature vector contains pairs for all the scales and orientations of the wavelets. From a number of experiments we found that a filter bank with six orientations and four scales gave the best classification performance for the classifiers used, which means 24x2 component features will be extracted for a given image patch. Therefore, the figuration is applied to 6x8 non-overlapping image subregions each with the size 60x64, yielding overall feature vector with length 4x5x48=960 for each image. For GLCM feature case, 16 gray co-occurrence matrices were created for each image with an offset that specifies four orientations 0, π/4, π/2 and 3π/4 and 4 distances (1,2,3 and 4 pixels) for each direction. Then for each normalized co-occurrence matrix P(i,j), 12 different type of statistic measurements were estimated, including correlation, variance, contrast, energy, difference variance, entropy, and homogeneity, as described in Section 2. Thus the dimension of GLCM feature is 16x12 = 192. To normalize for the differences in range, each of the LBP, CLBP, Gabor and GLCM feature components is scaled to have a mean of zero and a standard deviation of one across the dataset.

As first set of experiment, we compared the classification performance from the three base classifiers, *i.e.*, random forest, SVM and three-layer perceptron (MLP) neural network, for each of the features (LBP, CLBP, Gabor, GLCM and SLF). The experiment settings for all the classifiers are summarized as follows. For MLP, we experimented with a three-layer network. Specifically, the number of inputs is the same as the number of features, one hidden layer with 20 units and a single linear unit representing the class label. The network is trained using the Conjugate Gradient learning algorithm for 500 epochs. To prevent saturation, the target values are scaled to 0.9 for positive cases and to 0.1 for negative cases.

The popular library for support vector machines LIBSVM (*www.csie.ntu.edu.tw/~cjlin/libsvm*) was sued in the experiment. The parameter γthat defines the spread of the radial function was set to be 5.0 and parameter C that defines the trade-off between the classifier accuracy and the margin (the generation) to be 3.0. We use the radial based function kernel for the SVM classifier when Gabor, GLCM and SLF features were applied and the histogram intersection kernel for LBP/CLBP histograms. With the random forest classifier, the number of trees was chosen as 300 and the number of variables to be randomly selected from the available set of variables was selected as 20. For the 2D HeLa data set, we randomly split it into training and testing sets, each time with 20% of each class's images reserved for testing while the rest for training. The classification accuracy results reported in Table 1 are the average accuracies from 100 runs, such that each run used a random split of the data to training and testing sets.

Classifier	Gabor	LBP	CLBP	GLCM	SLF
RF	73%	72%	85.3%	72%	84%
SVM	82.4%	71.9%	71.6%	78.4%	83.8%
MLP	80%	65.2%	58.5%	86.5%	85.4%

TABLE 1: Performances of three classifiers using different features.

Then we proceeded the experiment with the proposed two-stage hybrid classifier system. The first stage consists of five SVM ensembles which use different sets of features (Gabor, LBP, CLBP, GLCM and SLF). Each base SVM classifier ensemble is trained using the entire training set of the corresponding feature, for example, an LBP feature is used to train 10 binary SVMs. Each binary SVM classifier in a feature specific ensemble was trained to act as a subcellular location detector, outputting a high posterior probability if its corresponding feature is present and a low posterior probability otherwise. During classification, a test instance feature is sent to the 10 base SVM classifiers that estimate the posterior probabilities, with the largest one among the base SVMs indicating the class label. Then 3-out-of-5 majority voting is applied to the output labels from the five SVM ensemble to decide a class label if there is a consensus or reject otherwise. Here the ``consensus'' criterion k=3 acts like a threshold to split the instances into two partitions. In another words, the SVM classifier ensemble collectively labels the multiple feature instances for a give testing HeLa image as belonging or not to any of the 10 categories, while it rejects them from the remaining categories, i.e. no decision is taken about these latter categories. Using a holdout experiment with 80% of data were used for training while the remaining for testing, the first stage accuracy approximates 98% with rejection rate 48%.

The second stage of classifier ensemble consists of 5x3 = 15 multi-class classifiers, which are neural network (NN) classifier, multi-class support vector machine (SVM), and Random Forest classifier, with the five different features. All the base classifiers are simultaneously trained with stage 1 ensemble. During classification, the rejected instances from stage 1 ensemble is passed to the stage 2. Similar to stage 1, k-out-of-15 majority voting is applied to the output labels from the 15 classifiers to decide a class label if there is a consensus or reject otherwise, while k can be controlled to yield varying rejection rate. The overall classification accuracy is defined as the number of correctly classified samples from both stage 1 and stage 2 over the total number of samples tested. From the same holdout experiment with 80% of data for training while the remaining for testing, the second stage accuracy is above 96% with rejection rate 21%, as shown in Figure 7. We also compared different rejection rates between 6% and 42% from stage 2 by varying k in the k-out-of-15 majority voting, yielding the classification accuracies as illustrated in Figure 8. It seems that rejection rate larger than 35% will not bring any more improvement for the classification performance. The corresponding box plot for the comparison is given in Figure 9.

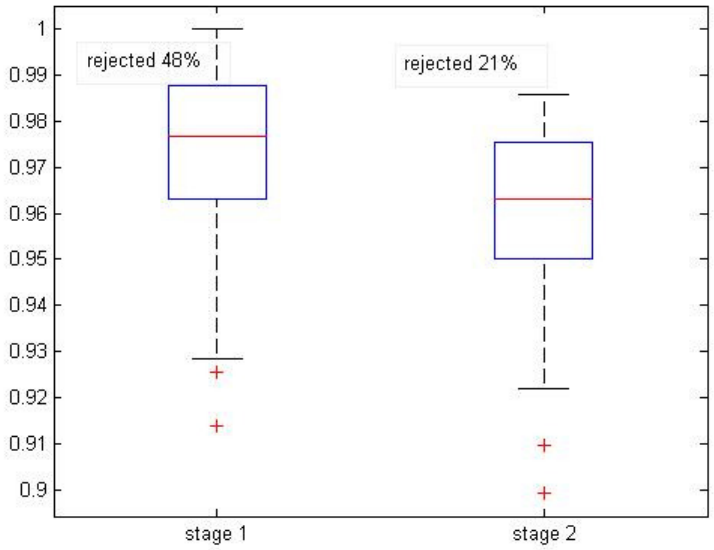


FIGURE 7: Comparison of the final accuracy from stage 2 with overall rejection rate 21% and the first stage accuracy with rejection rate 48%. resulting from holdout experiment with 80% of data were used for training while the remaining for testing. The results were from the average of 100 tests.

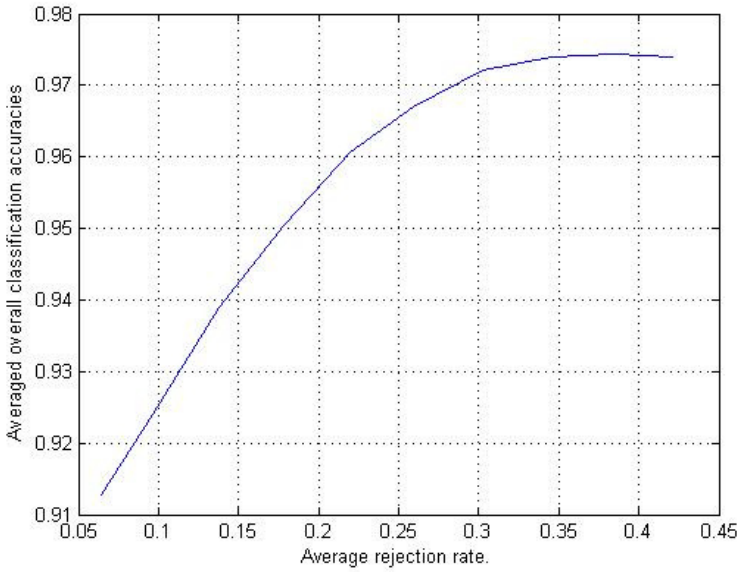


FIGURE 8: Overall accuracies with 10 varying rejection rates in the second stage

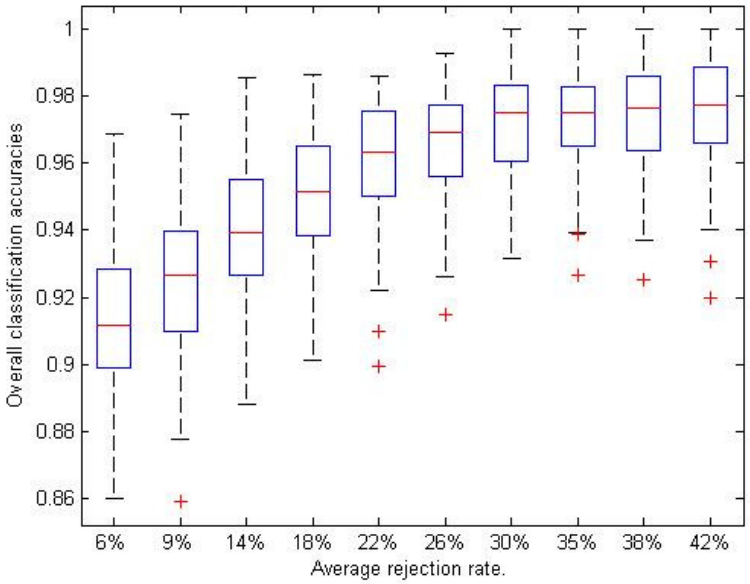


FIGURE 9: Boxplots of classification performances from the different classifiers resulting from holdout experiment with 80% of data were used for training while the remaining for testing.

The confusion matrices that summarize the details of a special situation with rejection rate 6\% is given the following Table 2. For the total number of 187 testing samples, the 10-by-10 matrix displays the number of correct and incorrect predictions made by the hybrid classification system compared with the actual classifications in the test data. It is obvious that among the 10 classes, Actin Filaments type is the easiest to be correctly classified while the Endosome and Golgi_gpp are the difficult categories. This is consistent with previous observations regarding the different degree of difficulties to distinguish the 10 type of subcellular locations [6-7] .

%	1	2	3	4	5	6	7	8	9	10	Accuracy
1	20	0	0	0	0	0	0	0	0	0	100%
2	0	12	1	0	0	2	1	1	0	0	70.6%
3	0	1	15	0	1	0	2	1	0	1	71.4%
4	0	0	0	16	2	1	0	0	0	1	80%
5	0	1	0	1	14	1	1	1	1	0	70%
6	0	3	0	0	1	14	0	0	0	0	77.8%
7	0	1	1	0	0	0	16	1	0	0	84.2%
8	0	1	1	0	0	1	1	11	0	0	73.3%
9	0	0	0	0	1	1	0	0	16	0	88.9%
10	0	0	1	0	0	0	0	1	0	17	89.5%

TABLE 2: Confusion matrix for test set with overall rejection rate 6%. (1: ActinFilaments, 2: Endosome, 3: ER, 4: Golgi_gia, 5: Golgi_gpp, 6: Lysosome, 7: Microtubules, 8: Mitochondria, 9: Nucleolus, 10: Nucleus)

5. CONSLUSION & FUTURE WORK

Automated identification of sub-cellular organelles is important when characterizing newly discovered genes or genes with an unknown function. In this paper, a two-stage multiple classifier system was proposed with rejection strategies for subcellular phenotype images classification. Rather than simply pursuing classification accuracy, we emphasized reject option in order to minimize the cost of misclassifications while secure high classification reliability. The two-stage method used a serial approach where the second classifier ensemble is only responsible for the patterns rejected by the first classifier ensemble. The first stage ensemble consits of binary SVMs with different features, including texture features local binary patterns (LBP), Gabor filtering and Gray Level Co-occurrence Matrix (GLCM), together with Subcellular Location Features (SLF). The first stage ensemble was trained in parallel with the second which is composed of multiple layer perceptron, multi-class support vector machine (SVM), and the Random Forest classifier. During classification, the cascade of classifier ensembles receives a plurality of samples corresponding to different features. The first stage classifier ensemble generates classifications for each of the samples as well as a confidence score associated with the classifications. If the confidence score for a received sample is above a threshold associated with the ensemble, then it absorbs the sample. Otherwise, the classifier ensemble rejects the sample, and such sample is directed to a subsequent classifier ensemble within the cascade. A high classification accuracy 96% is obtained with rejection rate 21% for the 2D HeLa cells from the exploitation of the complementary strengths of feature construction and classifiers decision fusion.

6. ACKNOWLEDGMENTS

The project is funded by China Jiangsu Provincial Natural Science Foundation Intelligent Bioimages Analysis, Retrieval and Management (BK2009146).

7. REFERENCES

1. J. Davis, M. Kakar, and C. Lim. “Controlling protein compartmentalization to overcome disease”. *Pharm Res.* **24(1)**: pp.17—27, 2007..

2. N. Orlov, J. Johnston, T. Macura, L. Shamir and I.Goldberg, “Computer Vision for Microscopy Applications. Vision Systems: Segmentation and Pattern Recognition”, Edited by: Goro Obinata and Ashish Dutta, pp.546, I-Tech, Vienna, Austria, June 2007

3. H. Peng, “Bioimage informatics: a new area of engineering biology”. *Bioinformatics*, **24(17)**: pp. 1827—36, 2008.

4. E.J. Roques and R.F. Murphy RF. “Objective Evaluation of Differences in Protein Subcellular Distribution”, *Traffic*, **3**, Pages 61 – 65, 2002.

5. M.V. Boland and R.F. Murphy, “A neural network classifier capable of recognizing the patterns of all major subcellular structures in fluorescence microscope images of HeLa cells”. *Bioinformatics*, **17**(12): pp.1213—23, 2001.

6. M.V. Boland, M. Markey and R.F. Murphy, “Automated Recognition of Patterns Characteristic of Subcellular Structures in Fluorescence Microscopy Images”. *Cytometry*, **33**: pp. 366-375, 1998.

7. K. Huang and R.F. Murphy,. “Boosting accuracy of automated classification of fluorescence microscope images for location proteomics”. *BMC Bioinformatics*, **5**: 78, 2004.

8. A. Chebira, Y. Barbotin, C. Jackson, T. Merryman, G. Srinivasa, RF., Murphy and J. Kovacevic, “A multiresolution approach to automated classification of protein subcellular location images”. *Bioinformatics*, **8**: 210, 2007.

9. L.Nanni, A. Lumini, Y. Lin, C. Hsu, and C. Lin, "Fusion of systems for automated cell phenotype image classification". *Expert Systems with Applications*, **37**: pp. 1556-1562, 2010.

10. N.A. Hamilton, R.S. Pantelic, K. Hanson and R.D.Teasdale. "Fast automated cell phenotype image classification". *Bioinformatics*, **8**: pp. 110, 2007.

11. B. Zhang, "Classification of Subcellular Phenotype Images by Decision Templates for Classifier Ensemble". *International Conference on Computational Models for Life Sciences* (CMLS-09), AIP Conf. Proc. **1210**, pp.13-22, 2009.

12. T. Ojala, M. Pietikainen and T. Maenpaa, "Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **24**(7): pp.971-987, 2002.

13. L. Wolf, T. Hassner and Y. Taigman, "Descriptor Based Methods in the Wild". *Faces in Real-Life Images workshop at the European Conference on Computer Vision (ECCV)*, Oct 2008.

14. T. Ahonen, A. Hadid and M. Pietikainen, "Face Description with Local Binary Patterns: Application to Face Recognition". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **28**(12): pp.2037-2041, 2006.

15. G. Zhang, X. Huang, S.Z. Li, Y. Wang, and X. Wu, "Boosting Local Binary Pattern (LBP)-Based Face Recognition". In *Proc. Advances in Biometric Person Authentication: 5th Chinese Conference on Biometric Recognition, SINOBIOMETRICS* 2004, Guangzhou, China. pp. 179-186, 2005.

16. Z. Guo, L. Zhang and D. Zhang "A Completed Modeling of Local Binary Pattern Operator for Texture Classification". accepted for *IEEE Trans Image Process*., preprint, 2010.

17. B. Manjunath and W. Ma, "Texture Features for Browsing and Retrieval of Image Data". *IEEE Trans. on Pattern Analysis and Machine Intelligence*, **18**(8):, pp.837—842, 1996.

18. R. Haralick "Statistical and Structural Approaches to Texture",. *Proceedings of the IEEE*, **67**(5)} pp. 786-804, 1979.

19. L.I. Kuncheva, "Combining Pattern Classifiers: Methods and Algorithms". Wiley-Interscience., (2004).

20. R.M. Nosofsky, T.J. Palmeri and S.C. McKinley, "Rule-Plus-Exception Model of Classification Learning". *Psychological Review*, **101**, pp.53-79, 1994.

21. R. Rifkin and A. Klautau, "In Defense of One-Vs-All Classification". *Journal of Machine Learning Research*, **5**: pp. 101-141, 2004.

22. N. Holden and A.A. Freitas "A Hybrid PSO/ACO Algorithm for Discovering Classification Rules in Data Mining", *Journal of Artificial Evolution and Applications*, **2008**, Article ID 316145, 11 pages, 2008.

23. D.M.J Tax and R.P.W. Duin, "Growing a multi-class classifier with a reject option", *Pattern Recognition Letters*, **29**: pp. 1565-1570, 2008.

24. C.K.Chow, "On optimum recognition error and reject tradeoff". *IEEE Trans. Inf. Theory*, **IT-16** (1), 41–46, 1970.

25. N.Giusti, F. Masulli, F., Sperduti, “A Theoretical and Experimental Analysis of a Two-Stage System for Classification”. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, **24**: pp. 893–904,2002.

26. P. Pudil, J. Novovicova, S. Blaha, J. Kittler, “Multistage Pattern Recognition with Reject Option”. In: *Proc. 11th IAPR Int. Conf. on Pattern Recognition*, **2**: pp.92-95, 1992.

27. G. Fumera and F. Roli. Support Vector Machines with Embedded Reject Option, *Int. Workshop on Pattern Recognition with Support Vector Machines* (SVM2002), Springer, Niagara Falls, Canada, p.68-82, 2002.

28. R.P.W. Duin and D.M.J. Tax, “Classifier conditional posterior probabilities”. In: Amin, A., Dori, D., Pudil, P., Freeman, H. (eds.): Advances in Pattern Recognition. Lecture Notes in Computer Science 1451, Springer, Berlin, 611-619, 1998.

29. L. Lam and C.Y. Suen, “Application of Majority Voting to Pattern Recognition: An Analysis of Its Behavior and Performance”, *IEEE Transactions on Systems, Man, and Cybernetics -Part A: Systems and Human*, **27**: pp.553-568, 1997.

30. J. Shawe-Taylor and N. Cristianini, “Kernel methods for pattern analysis”. Cambridge University Press (2004).

31. C.-W. Hsu and C.-J. Lin, “A comparison on methods for multi-class support vector machines”. *IEEE Transactions on Neural Networks*, **13**: pp.415—425, 2002.

32. R.O. Duda, P.E. Hart and D.G. Stork,D.G. “Pattern classification”, Second Edition, John Wiley and Sons, New York, (2001).

33. S. Maji, A.C. Berg, and J. Malik, . “Classification Using Intersection Kernel Support Vector Machines is efficient”. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2008)* ,Anchorage, Alaska, pp. 1-8, 2008.

34. L. Breiman, “Random Forests”. *Machine Learning*, **45,** pp. 5–32, 2001.

CALL FOR PAPERS

Journal: International Journal of Biometrics and Bioinformatics (IJBB)
Volume: 4 **Issue:** 6
ISSN:1985-2347
URL: http://www.cscjournals.org/csc/description.php?JCode=IJBB

About IJBB

The International Journal of Biometric and Bioinformatics (IJBB) brings together both of these aspects of biology and creates a platform for exploration and progress of these, relatively new disciplines by facilitating the exchange of information in the fields of computational molecular biology and post-genome bioinformatics and the role of statistics and mathematics in the biological sciences. Bioinformatics and Biometrics are expected to have a substantial impact on the scientific, engineering and economic development of the world. Together they are a comprehensive application of mathematics, statistics, science and computer science with an aim to understand living systems.

We invite specialists, researchers and scientists from the fields of biology, computer science, mathematics, statistics, physics and such related sciences to share their understanding and contributions towards scientific applications that set scientific or policy objectives, motivate method development and demonstrate the operation of new methods in the fields of Biometrics and Bioinformatics.

To build its International reputation, we are disseminating the publication information through Google Books, Google Scholar, Directory of Open Access Journals (DOAJ), Open J Gate, ScientificCommons, Docstoc and many more. Our International Editors are working on establishing ISI listing and a good impact factor for IJBB.

IJBB List of Topics

The realm of International Journal of Biometrics and Bioinformatics (IJBB) extends, but not limited, to the following:

- Bio-grid
- Bioinformatic databases
- Biomedical image processing (registration)
- Biomedical modelling and computer simulation
- Computational intelligence
- Computational structural biology
- Bio-ontology and data mining
- Biomedical image processing (fusion)
- Biomedical image processing (segmentation)
- Computational genomics
- Computational proteomics
- Data visualisation

- DNA assembly, clustering, and mapping
- Fuzzy logic
- Gene identification and annotation
- Hidden Markov models
- Molecular evolution and phylogeny
- Molecular sequence analysis
- E-health
- Gene expression and microarrays
- Genetic algorithms
- High performance computing
- Molecular modelling and simulation
- Neural networks

IMPORTANT DATES

Volume: 4
Issue: 6
Paper Submission: November 31, 2010
Author Notification: January 01, 2011
Issue Publication: January /February 2011

CALL FOR EDITORS/REVIEWERS

CSC Journals is in process of appointing Editorial Board Members for ***International Journal of Biometrics and Bioinformatics***. CSC Journals would like to invite interested candidates to join **IJBB** network of professionals/researchers for the positions of Editor-in-Chief, Associate Editor-in-Chief, Editorial Board Members and Reviewers.

The invitation encourages interested professionals to contribute into CSC research network by joining as a part of editorial board members and reviewers for scientific peer-reviewed journals. All journals use an online, electronic submission process. The Editor is responsible for the timely and substantive output of the journal, including the solicitation of manuscripts, supervision of the peer review process and the final selection of articles for publication. Responsibilities also include implementing the journal's editorial policies, maintaining high professional standards for published content, ensuring the integrity of the journal, guiding manuscripts through the review process, overseeing revisions, and planning special issues along with the editorial team.

A complete list of journals can be found at http://www.cscjournals.org/csc/byjournal.php. Interested candidates may apply for the following positions through http://www.cscjournals.org/csc/login.php.

Please remember that it is through the effort of volunteers such as yourself that CSC Journals continues to grow and flourish. Your help with reviewing the issues written by prospective authors would be very much appreciated.

Feel free to contact us at coordinator@cscjournals.org if you have any queries.

Contact Information

Computer Science Journals Sdn BhD
M-3-19, Plaza Damas Sri Hartamas
50480, Kuala Lumpur MALAYSIA

Phone: +603 6207 1607
+603 2782 6991
Fax: +603 6207 1697

BRANCH OFFICE 1
Suite 5.04 Level 5, 365 Little Collins Street,
MELBOURNE 3000, Victoria, AUSTRALIA

Fax: +613 8677 1132

BRANCH OFFICE 2
Office no. 8, Saad Arcad, DHA Main Bulevard
Lahore, PAKISTAN

EMAIL SUPPORT
Head CSC Press: coordinator@cscjournals.org
CSC Press: cscpress@cscjournals.org
Info: info@cscjournals.org

COMPUTER SCIENCE JOURNALS SDN BHD
M-3-19, PLAZA DAMAS
SRI HARTAMAS
50480, KUALA LUMPUR
MALAYSIA