DOI: http://dx.doi.org/10.26483/ijarcs.v9i5.6326

Volume 9, No. 5, September-October 2018

ISSN No. 0976-5697

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Higher order Analyzes of ASD genetic Data Using Prefix span and PCA methods

Dr. B. Lavanya
Department of Computer Science
University of Madras
Chennai – 600 025, India

T. Madhumitha
Department of Computer Science
University of Madras
Chennai – 600 025, India

Abstract: The most important aim of data mining is to extract useful information from the datasets. Data mining can extract meaningful patterns from large datasets and it can analyze the dataset to predict and classify the dataset based on user specification. This paper deals with medical database called Gene Expression Omnibus from NCBI database, analysed using data mining techniques. The Microarray data of Autism Spectrum Disorder (ASD), contains 100 genes from 21 ASD children, analysed using unsupervised pattern mining algorithm called PREFIXSPAN to find the sequence pattern and dimensionality reduction as Principal Component Analysis (PCA) algorithm, to find the positively and negatively correlated genes for ASD. From the comparison of algorithms, it infers the genes that are Highly Influence by Autism Spectrum Disorder from the 100 genes.

Keywords: Pattern Mining; Prefixspan; Positively Correlated; Negatively Correlated; Data Mining

I. INTRODUCTION

The microarray data contain huge number of genes and number of samples. And from the data the disease prediction and gene analyzing is done. Pattern mining discovers the most useful and interesting patterns from the database. Principal Component Analysis (PCA) reduces the dimensionality of data. In data mining there are many numbers of variables in data base from which the highly correlated variables are identified using principal component analysis. The visual representation of PCA, shows the pattern in the dataset. This PCA used in to compare the genes as to analyse the gene expression. By using PCA, prefix span analyse to find the positively, negatively, and poorly correlated variables.

II. LITRATURE REVIEW

Yin Li, Yan Cong, Yun Zhao (2016), [1], describes the network motif for coronary artery disease. Differential integrated gene and protein-protein interaction gene are analyzed to interaction pattern is identified by screening of differential network. The network is to find the top 20 network, which is used to identify the coronary artery disease. For screening the network the R package global ancova software where used. The main advantage of screened network motif is, to give the accurate result to identify the coronary artery disease. This network motif method gives the accurate result. **Yin Wang, Rudong Li, Yuhua Zhou, Zongxin Ling, Xiaokui Guo, Lu Xie and Lei Liu (2016), [2],** classify the disease based on microbial meta-genome. These classifications are done by the method Phylogenetic tree based motif finding algorithm (PMF). The PMF algorithm has three parts that is motif finding, motif sorting and model evaluations. This PMF classifies two diseases, pneumonia and dental caries based on the microbial meta-genome. The main advantage of using PMF is to find the motifs in the training data, from which disease is classified. **S.Padmavathi, Ramanujam. E (2015), [5],** use the method Multivarient maximal time series motifs to identify the frequently occurring patterns and then it uses a Naive bases classifier to classify the normal and abnormalities signal, the accuracy is 93.33% and 98% of precision rate. This method is used in the application of Electrocardiogram (ECG) to classify the abnormality in ECG signals. **Duc-Hau Le, Vu-Tung Dang, Springer Berlin Heidelberg, (2016), [10],** in this the network motifs is used for disease prediction. The Random walk restart on heterogeneous network (RWRH) algorithm is used in network motifs, which identify the similarity of network for Alzhemer's disease based on the network it gives the better functionality among the disease. Ontology is used to predict the network similarity **Shameek Ghosh , HungNguyen and jinyan Li, (2016), [6],** which deals to detect the critical patient events like hypotension and septic shock based on the method, order sequential contrast pattern based classification in the time series sequence for detecting patient event. SVM and HMM is used to classify the disease and this use the arterial pressure series. And this will give the better prediction in ICU outcomes which is the application of this system. **Kai Shi, Lin Gao, Lin Gao, Bingbo Wang (2016), [7],** used the method called network motifs, the centrality for analysing the shortest path between the nodes. The highest the centrality scores the more significant motifs. This is the application based on colorectal cancer disease. The pathway in the disease, it is a significant pathway which enriches the gene reported related to cancer development. **Adnan Ferdous Ashrafi, A.K.M Iqtidar Newaz, Rasif Ajwad Moin (2015), [8],** which will find the motifs in DNA sequence by Integer Matching using Hash table indexing, and rank the motifs then calculate the fitness in DNA sequence. The main advantage is the DNA sequence will be accurate and effective.

III. PROPOSED WORK

The proposed work analyzes the genes in Microarray of Autism Spectrum Disorder; the dataset was collected from Gene Expression Omnibus in NCBI database. The dataset was analyzed using unsupervised algorithms called the pattern mining that is PREFIXSPAN and dimensionality reduction algorithm called Principal Component Analysis (PCA).The Dataset Contains 21 ASD children as samples and their respective genes as attributes. And this data is collected from peripheral blood leucocytes associated with gene expression. RNA was prepared from the venous blood of 21 ASD children

Each algorithm was implemented in the dataset to analyze the genes that are influenced by ASD. And then finally compare the result of the algorithms to infer the genes that are Highly Influenced by ASD from the 100 genes.

The dataset format is in the following Table 1

Table 1: ASD Dataset

GENE NUMBER	GENE SYMBOL
1	FAM174B
2	AP3S2
3	SV2B
4	RBPMS2
5	AVEN
10 . . .	ATMIN . .
25	ORC6
.	.
.	.
.	.
50	HN1
.	.
.	.
100	SFRP1

IV. ALGORITHM TO ANALYZE THE GENES

A frequent pattern mining is a set of item that frequently repeated and form as pattern. This frequent pattern is based on the user specified threshold value. The association, correlation are mine using frequent items in the dataset. Association rules or frequent patterns techniques is used in bioinformatics to analyze, predict the disease.

A. PREFIXSPAN Algorithm

Prefix-projected Sequential pattern mining algorithm helps to identify sequential pattern in data. Prefixspan identifies the combination of various sequence patterns from the dataset. The sequence patterns are visible, that are not less than Min_support value. Prefix Span in microarray data analyze the genes, and then identify the genes, that form as pattern for ASD. This patterns are mine using the Min_support as threshold value.

The Table 2 identifies the genes mostly repeated in many patterns, and those patterns of genes are influenced by Autism Spectrum Disorder.

Table 2: Genes repeated in number of patterns

GENE SYMBOL	No. of patterns repeated
NPRL3	6
VSP39, ZNF614	5
FAM174B, ZCCHC14, ARL13B	4
RBPMS2, NELL2, CMIP, EXOSC6, HMOX2, HN1 HELZ, SMG6, MAPK7, YO5B, C18orF21	3
ZSCAN29, CHP,UNKL,MMP15, PRM1,ESRP2,DECR2 ORC6, MTHFSD, KAT8,VKORC1,EXOC7 LPO,SKQL1,SERPINF1,SPNS3, KRT38, KRTAP3-2, OR1E1,ZNF614	2
APEN, SV2B, AVEN, CASC5, ATMIN, HSBP1, GRIN2A, ZNF598, OC6, IL34 ZB7B4, SLC25A39, PNPLA2, STAT5B, GIT1, AIPL1 KRTAP9-3, KRT34, CTDP1 VSIG1OL, SFRP1	1

Gene's pattern in Prefixspan

From the Fig 3, show the genes that form as patterns, and those genes are repeated in number of pattern. Only 57 genes from 100 genes are occurring as different pattern and the remaining genes are not repeated and form as pattern because those genes are not support by the Minimum support threshold value.

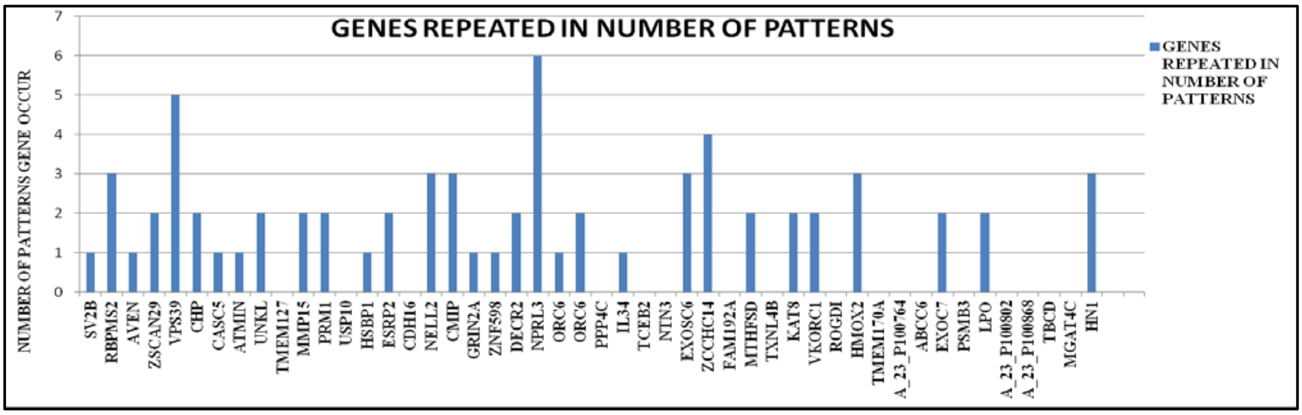

Fig 1 (i) Gene as pattern

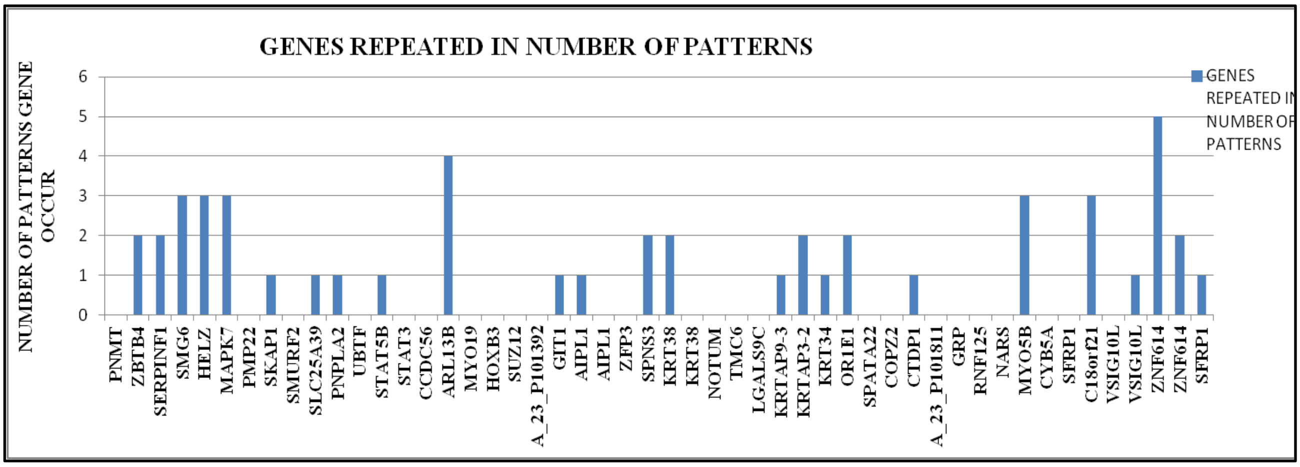

Fig 1(ii) Genes in number of pattern

B. Principal Component Analysis

Principal component analysis (PCA) has been used in data analysis to reduce the dimensionality of the data in order to simplify analysis. PCA uses mathematical technique to reduce the dimension of data. The standard deviation, covariance, eigenvectors and eigenvalues are used in PCA to analyze correlation in variables. PCA mainly concentrated with identifying correlation in data. Values close to +1 indicate positive correlation, and values near to -1 are negative correlation. Values close to zero is poor correlation and 0 indicates no correlation at all. From the ASD microarray data, PCA analyze from the 100 genes of ASD, only 62 genes are highly correlates with each and those genes are influenced by ASD and it is plotted as two principal components.

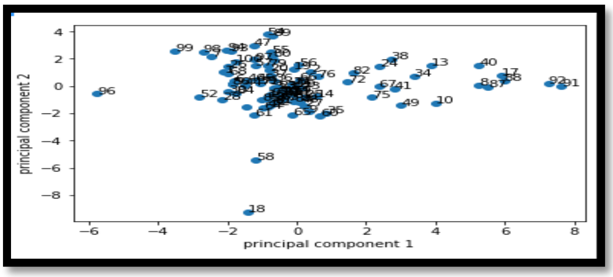

Fig 2: Principal Component Analysis

The 62 genes are correlated and influenced by the ASD.

Positively and Negatively Correlated Genes

The genes are positively correlated and negatively correlated gene. From 100 genes positively correlated genes are 17 and negatively correlated genes are 45 and poorly correlation genes are 38. This was described in the Fig 3. The negatively correlated gene matches with pattern mining algorithm.

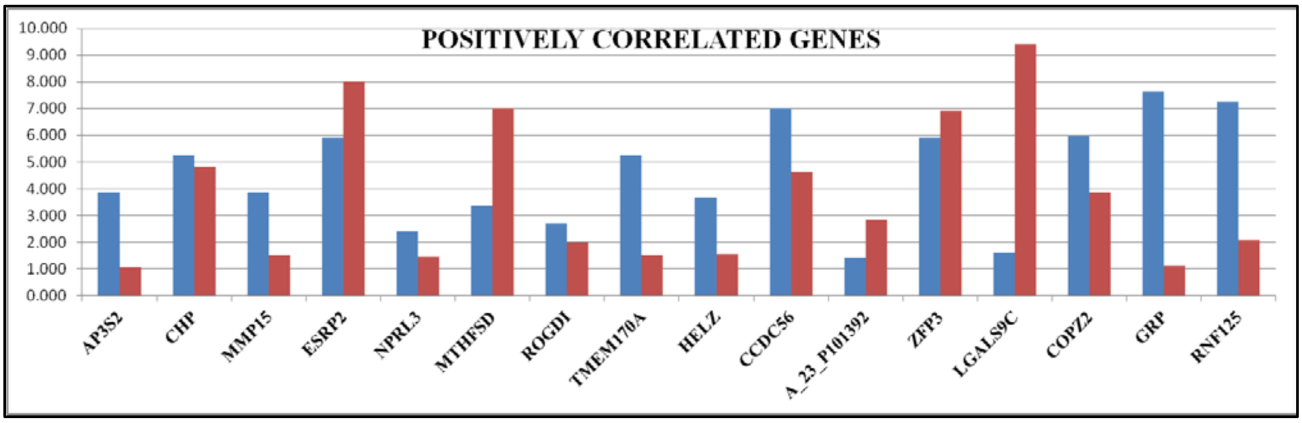

Fig 3(i): Positively Correlated Genes

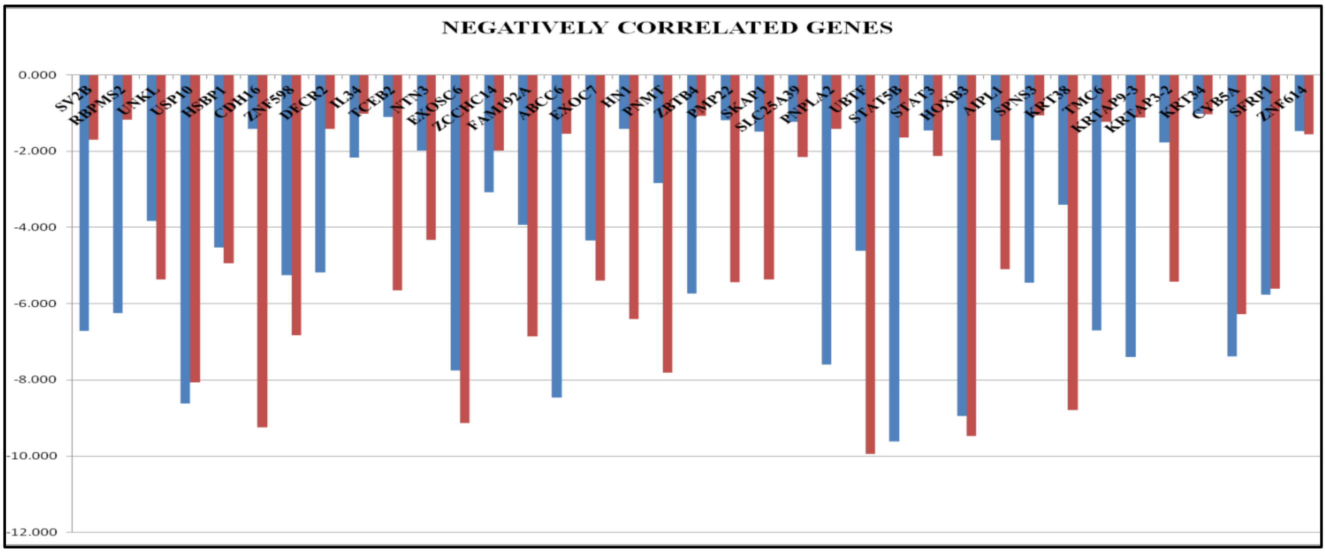

Fig 3(ii): Negatively correlated gene

V. RESULT

Some of the genes are influenced by the algorithms Prefixspan and PCA. From the comparison table it can clearly visualize only 16 genes are highly influenced by Autism Spectrum Disorder Out of 100 genes and those genes are highlighted in the table 3.

Table 3: Comparison of algorithm for to identify Highly Influenced Genes

GENE NO	PREFIXSPAN ALGORITHM	PCA	HIGHLY INFLUENCED GENES FOR ASD
1	**FAM174B**	FAM174B	*
2	AP3S2	**AP3S2**	*
3	SV2B	**SV2B**	*
4	**RBPMS2**	**RBPMS2**	**RBPMS2**
5	**AVEN**	AVEN	*
6	**ZSCAN29**	ZSCAN29	*
7	**VPS39**	VPS39	*
8	**CHP**	**CHP**	**CHP**
9	**CASC5**	CASC5	*
10	**ATMIN**	ATMIN	*
11	UNKL	**UNKL**	*
12	TMEM127	**TMEM127**	*
13	**MMP15**	MMP15	*
14	**PRM1**	PRM1	*
15	USP10	**USP10**	*
16	**HSBP1**	**HSBP1**	**HSBP1**
17	ESRP2	**ESRP2**	*
18	CDH16	**CDH16**	*
19	**CMIP**	CMIP	*
20	**ZNF598**	**ZNF598**	**ZNF598**
21	**DECR2**	**DECR2**	**DECR2**
22	**NPRL3**	**NPRL3**	**NPRL3**
23	**ORC6**	ORC6	*
24	IL34	**IL34**	*
25	**TCEB2**	**TCEB2**	*
26	NTN3	**NTN3**	*
27	**ZCCHC14**	**ZCCHC14**	**ZCCHC14**
28	FAM192A	**FAM192A**	*
29	**MTHFSD**	**MTHFSD**	*
30	**KAT8**	KAT8	*
31	**VKORC1**	VKORC1	*

32	ROGDI	**ROGDI**	*
33	**HMOX2**	HMOX2	*
34	TMEM170A	**TMEM170A**	*
35	ABCC6	**ABCC6**	*
36	**EXOC7**	**EXOC7**	**EXOC7**
37	**LPO**	LPO	*
38	**MGAT4C**	MGAT4C	*
39	HN1	**HN1**	*
	PNMT	**PNMT**	*
41	**ZBTB4**	**ZBTB4**	*
42	**SERPINF1**	SERPINF1	*
43	**HELZ**	**HELZ**	**HELZ**
44	**MAPK7**	MAPK7	*
45	PMP22	**PMP22**	*
46	SKAP1	**SKAP1**	*
47	SMURF2	SMURF2	*
48	**SLC25A39**	**SLC25A39**	**SLC25A39**
49	**PNPLA2**	**PNPLA2**	**PNPLA2**
50	UBTF	**UBTF**	*
511	STAT5B	**STAT5B**	*
52	STAT3	**STAT3**	*
53	CCDC56	**CCDC56**	*
54	**ARL13B**	ARL13B	*
55	HOXB3	**HOXB3**	*
56	A_23P101392	**A_23P101392**	*
57	**GIT1**	GIT1	*
58	AIPL1	**AIPL1**	*
59	ZFP3	**ZFP3**	*
60	**SPNS3**	**SPNS3**	**SPNS3**
61	**KRT38**	**KRT38**	**KRT38**
62	TMC6	**TMC6**	*
63	LGALS9C	**LGALS9C**	*
64	**KRTAP9x3**	**KRTAP9x3**	**KRTAP9x3**
65	KRTAP3x2	**KRTAP3x2**	*
66	**KRT34**	**KRT34**	**KRT34**
67	**OR1E1**	OR1E1	*
68	COPZ2	**COPZ2**	*
69	**CTDP1**	CTDP1	*
70	GRP	**GRP**	*
71	RNF125	**RNF125**	*
72	**MYO5B**	MYO5B	*
73	CYB5A	**CYB5A**	*
74	SFRP1	**SFRP1**	*
75	**C18orf21**	C18orf21	*
76	**VSIG10L**	VSIG10L	*
77	**ZNF614**	**ZNF614**	**ZNF614**
78	**SFRP1**	SFRP1	*

VI. CONCLUTION

The Autism Spectrum Disorder Microarray dataset is analysed to identify the genes which influenced by ASD using the algorithms, sequence pattern mining that is Prefixspan algorithm and Dimensionality reduction algorithm called Principal Component Analysis (PCA) to find the positively and negatively correlated genes. From the comparison table to visualize the genes that are Highly Influenced by ASD

Future Scope

In this paper only 100 genes are analyzed, but there are more genes in dataset, for future many genes can analyze by the algorithms to identify the genes that are highly influenced by the ASD.

VII. REFERNCE

[1] Yin Li, Yan Cong, Yun Zhao, “Network motif-based for identifying coronary artery disease”, Experimental and therapeutic medicine (12)(1): 257-261, Jul; 2016.[online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4907106/. [Received 2015 May 22; Accepted 2016 Apr 1]. doi: 10.3892/etm.2016.3299 .

[2] Yin Wang, Rudong Li, Yuhua Zhou, Zongxin Ling, Xiaokui Guo, Lu Xie and Lei Liu , “Motif-Based Text Mining of Microbial Metagenome Redundancy Profiling data for disease Classification”, BioMed Research International Volume , Hindawi publishing corporation, 2016, ArticleID 6598307, 11pages, [Recevied 28 October 2015; Accepted 12 January 2016], http://dx.doi.org/10.1155/2016/6598307.

[3] Ian Fox, Lynn Ang, Mamta Jaiswal, Rodica, Pop-Busui, Jenna wiens, “Contextual Motifs- Increasing the utility of Motifs using Contextual Data”, in KDD ’17 Proceeding of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada – August 13 - 17, 2017, Pages 155 – 164.

[4] Jiawei Lu, Di Dai,Buwen Cao, Ying Yin, “Inferring human miRNA functional similarity based on gene ontology annotation”, IEEE, 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), Changsha, China, 13-15 Aug. 2016.

[5] Padmavathi. S, Ramanujam. E , “Naïve Bayes Classifier for ECG abnormalities using Multivariate Maximal Time Series Motif”, Elsevier Procedia Computer Science 47:222 - 228, December 2015, DOI: 101016/j.procs2015.03.201

[6] Shameek Ghosh, HungNguyen and jinyan Li, “Predicting short-term ICU outcomes using a sequential contrast motif based classification framework”, IEEE 38th Annual International Conference of the IEEE Engineering in Medical and Biology Society (EMBC), Orlando, FL, USA, 16-20 Aug. 2016.

[7] Kai Shi, Lin Gao, Bingbo Wang, “Systematic tracking of coordinate differential network motifs identifies novel disease-related genes by integrating multiple data”, Elsevier Science Publishers B.V. Amsterdam, The Netherlands, Neurocomputing, Volume 206 Issue c, September 2016 Page 3-12.

[8] Adnan Ferdous Ashrafi, A.K.M Iqtidar Newaz, Rasif Ajwad Moin, Mahmud Tanvee, M.A Mottalib, “A Modified Algorithm for DNA Motif Finding and Ranking Considering Variable Length Motif and Mutation” Conference: Recent Trends in Information Systems, Kolkata, India, 2015.

[9] J.Sivaranjani, A.Neela Madheswari, “A Novel Technique of Motif Discovery for Medical Big data using Hadoop”

2017 Conference on Emerging Devices and Smart Systems (ICEDSS), Tiruchengode, India.

[10] Duc-Hau Le, Vu-Tung Dang, Springer Berlin Heidelberg, "Ontology-based disease similarity network for disease gene prediction", Vietnam Journal of Computer science , Volume 3 Issue 3, August 2016.