Context rank based hierarchical clustering algorithm on medical databases (CRBHCA).

BANO, S. and RAO, K.R.

2015

Journal of Theoretical and Applied Information Technology
20th May 2015. Vol.75. No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

CONTEXT RANK BASED HIERARCHICAL CLUSTERING ALGORITHM ON MEDICAL DATABASES (CRBHCA)

[1]SHAHANA BANO, [2]K.RAJASEKHARA RAO
[1]Research Scholar, Department of Computer Science and Engineering, KL University
[2]Professor, Department of Computer Science and Engineering, Sri Prakash College of Engineering
E-mail: [1]shahanabano_cse@kluniversity.in, [2] krr_it@yahoo.co.in

ABSTRACT

In this paper we proposed a method which avoids the choice of natural language processing tools such as pos taggers and parsers reduce the processing overhead. Moreover, we suggest a structure to immediately create a large-scale corpus annotated along with disease names, which can be applied to train our probabilistic model. In this proposed work context rank based hierarchical clustering method is applied on different datasets namely colon, Leukemia, MLL, Lymphoma medical diseases. Optimal rule filtering algorithm is applied on these datasets to remove unwanted special characters for gene/protein identification. Finally, experimental results show that proposed method outperformed existing methods in terms of time and clusters space.

Keyword: *Biomedical, Machine Learning, Gene/Protein, Clustering, Medline, Pubmed.*

1. INTRODUCTION

Life science studies are portrayed by the development of extensive and heterogeneous examples of biological study, including protein or gene series. Subsequently, various routines based upon content mining have been utilized to enhance the distinguish protein and gene names in medicinal writings.. Machine learning means the advancement and investigation of frameworks that could gain from information. It is a gigantic field with many algorithms for tending to diverse issues. Machine learning gives testing issues as far as algorithmic methodology, information representation, computational viability, and nature of the subsequent system. Biomedical information alongside its overhauls are spared in natural language style. Because of the improved measure of biomedical sources, it is getting to be more difficult to discover valuable and significant data with respect to a particular subject. All exploration innovations come and enter the repository at high-rate, making the procedure of discovering and dispersing quality data an exceptionally troublesome undertaking. Manual evaluation of such huge amount of information will presumably be exceptionally troublesome and lengthy. The issue is further amplified by the utilization of extensive assessment measures, and datasets that contain basically distinctive annotation arrangements and assignment definitions.

Medical text archives consistently conceal profitable organized information. An application of systems biology is to reveal the bio-courses of action fundamental the examples of a cell. Connections inside genes encode the greater part of this information and are incidentally found and symbolized as key items. Understanding these connections is a to a great degree testing issue as even the easiest living beings contain assortment genes that connect in intricate mixes to manage biological circumstances. An alternate entangling component is present high throughput method intended to focus the action level of genes is amazingly uproarious [1]. As there exists not very many well comprehended genetic activities, unsupervised clustering is a typical first step to to understand these information.

The clustering procedure is a fundamental tool to arrange an accumulation of articles inside a metric space into a set of littler segments called clusters. By utilizing clusters, the representation of the article pool can be made less demanding and the processing cost of information administration can be lessened. The made clusters can be utilized to present guidelines of top levels describing the common characteristics of information articles. . In the case of grammar induction structures, the rules of punctuation are expressed on word groupings as the words inside the same classification are changed also. If word categories are known,

Journal of Theoretical and Applied Information Technology
20th May 2015. Vol.75. No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

sentence structure standards may be investigated in a superior manner.

The rest of the paper is organized as follows: section 2 formally defines the previous Related work done and introduces notation. The Section 3 describes the algorithmic details to implement them. The experimentally evaluate of algorithms in section 4 and 5 with comparative Analysis. In section 6 the review related work and conclude in section 7.

2. RELATED WORK:

In the Related work context rank based hierarchical clustering method is applied on distinctive datasets specifically Leukemia, therapeutic ailments. Optimal rule filtering algorithm is applied on these datasets to remove undesirable special characters for gene /protein identification. This work conquers a percentage of the limits in the writing, for example: noise removal in medical datasets, strength, high disease forecast rate, high quality cluster result with less inquiry space and high genuine positive rate. At last, exploratory results demonstrate that proposed strategy outflanked well regarding time and clusters inquiry space are concerned. In future this work can be stretched out to execute comparable disease clusters on online medicinal records like medline, pubmed and so forth.

3. PROPOSED APPROACH

Following are the limitations of the related work discussed in this section.

Eliminate the non-functional characters

• Apply heuristic policies to remove non-functional symbols
• remove and replace the following symbols with gaps: #â€œ? $&*ã³@|~!\
• remove the subsequent characters if they are followed by a space: ;: .,
• eliminate the following pairs of brackets if the open bracket is preceded by a space and the closed bracket is followed by a space: [] ()
• eliminate the single quotation symbol if it is associated with by a space or if it is preceded by a space.
• remove s and t if they are associated with by a space
• eliminate slash / if it is associated with by a space.

Our proposed work overcomes all these limitations. We take three biomedical disease datasets namely colon, leukemia, mll medical diseases offline to extract hidden patterns using feature extraction and hierarchical clustering approaches. Each dataset is preprocessed to remove non-functional characters to identify disease names by using gene/protein database. Hierarchical methods for supervised

And unsupervised datamining give multilevel indexing of data. It can be relevant for several applications associated to data extraction, patterns retrieval and data organization.

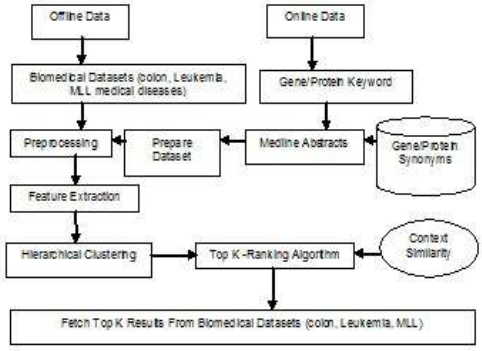

Fig No: 1 Proposed Method For Eliminating The Non-Functional Characters

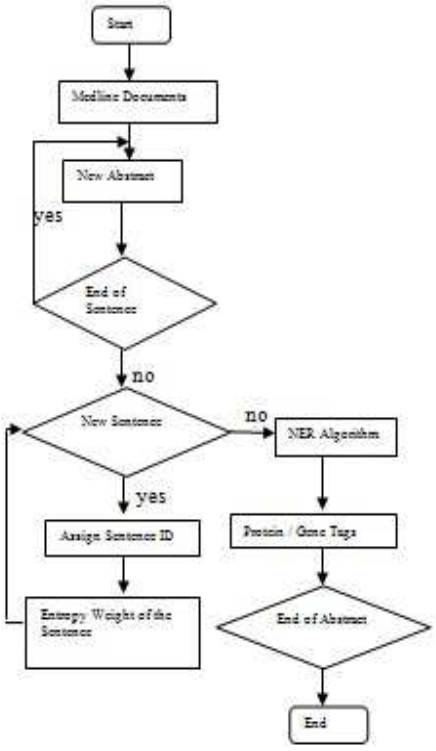

Fig No: 2 Proposed Method Flow Chart For Eliminating The Non-Functional Characters

Journal of Theoretical and Applied Information Technology
20th May 2015. Vol.75. No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Hierarchical Clustering Algorithm:

Input : Name entity Gene/Protein tags Tgp using NER approach,Gene/Protein DB, Probability P,Classes Positive pos, Negative neg, Tokenset Tk, Sentenceset Sen .

Read k, Threshold , Entropy weight ;

Output: Quality k- abstracts.

Tgp=Get(Name _Entity_Gene/ Protein_Tags)
for each tg in Tgp
For each in Tk
Calculate tag probability
List.add(tg)
List.add()
count=count+1
end
end.
For each token t in Tk
For each sen in Sentenceset
If((t ∈Sen)&& (t Tgp)&&(>getProb(t))
List Data ← Sentence_id,token, Pmid, Entropy_weight,Synonyms,Data,Title,PositiveClass
Else
List Data ← Sentence_id,token, Pmid, Entropy_weight,Synonyms,Data,Title,NegativeClass
End
End
For each pair of objects in Data
Calculate distance between two objects as

$$D(c1_i, c2_j) = (1 - r_{ij}) * 0.5$$

$$r_{ij} = (\sum_{i=0}^{d} (c1_{ij} - \overline{c1_i})(c1_{ji} - \overline{c1_j})) / \sqrt{\sum_{i=1}^{d} (c1_{ij} - \overline{c1_i})^2 \sum_{i=1}^{d} (c1_{ji} - \overline{c1_j})}$$

6.

a. Start with the disjoint clustering that have level as 0 and sequence_number m = 0.

b. Rank the pairs from smallest distance (similarities in common) to the maximal distance.

c. Calculate and count pairs, say n pairs.

If n >=0
do,

c.1 Explore the median as root hierarchical node.

c.2 Split the pairs as left and right side branches based on the median.

c.3 Explore the smallest unlike pair of clusters in the left side and right side current clustering, say pair rs, ls according to d[(rs),(ls)] = min r[(i),(j)] in which the minimum value is taken over all pairs of clusters in the current clustering.

c.4 If left side and right side have at least one similar object. In this case merge it collectively in one cluster, and look up smallest value over all pairs of clusters in the current clustering.

Else

c.5 Find the maximal dissimilar pair of clusters in the left side and right side current clustering, say pair rs, ls according to d[(rs),(ls)] = max r[(i),(j)] in which the m value is taken over all pairs of clusters in the current clustering.

d. Increment the sequence number: m = m +1. (In both left and right sides) Merge clusters (r) and (s) into a single-cluster to form the subsequent cluster m. Place the level of this cluster to L(m) = r[(r),(s)]

e. Revise the tree, T, by eliminating the nodes corresponding to clusters (p) and (q) and adding a node corresponding to the newly composed cluster. The neighborhood between the new cluster, denoted (p,q) and old cluster (m) is stated in this way:

d[(m), (p,q)] = min r[(m),(p)], d[(m),(q)].

If d <0
Then
Minimum Variance

f. If all objects are in one cluster, stop. Else, go to step b.

Algorithm2:

Input : Hierarchical clusters from top to bottom

Output: Top K Disease Results.

6.1 For each cluster in Cluster-set

6.1.1 t1=gene/protein search keyword.

6.1.2 For each synonym in the cluster
t2=synonym.
Find context similarity between t1 and t2.

Context Similarity Score:

$$\sum_{\substack{t1 \in cluster \\ t2 \in keyword}} Cos(t1, t2) / \prod_{i=1}^{m} sizeof(cluster_i)$$

End for

6.2 Sort <t1,t2> according to context similarity score.

6.3 Get abstracts from biomedical databases according to tag pair score.

4. RESULTS

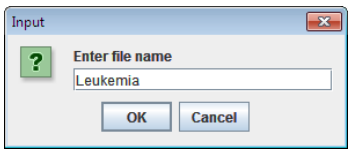

Fig 3: Loading Leukemia Disease Data

Partial Context Similarity of Gene/Proteins in leukemia:

Context Similarity %5.3f=>0.2526455026455026
<= U19107_rna1_at => synonyms are ZNF127 (ZNF127) gene
Context Similarity %5.3f=>0.3436507936507936

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

<= U19142_at => synonyms are GAGE1 G antigen 1 (GAGE-1)
Context Similarity %5.3f=>0.4829059829059829 <= U19180_at => synonyms are BAGE B melanoma antigen
Context Similarity %5.3f=>0.4363929146537842 <= U19261_at => synonyms are Epstein-Barr virus-induced protein mRNA
Context Similarity %5.3f=>0.2578347578347578 <= U19345_at => synonyms are AR1 protein (AR) mRNA
Context Similarity %5.3f=>0.43915343915343913 <= U19487_at => synonyms are Prostaglandin E2 receptor mRNA
Context Similarity%5.3f=>0.26296296296296295 <= U19517_at => synonyms are (apoargC) long mRNA
Context Similarity %5.3f=>0.38791423001949316 <= U19523_at => synonyms are GCH1 GTP cyclohydrolase 1 (dopa-responsive dystonia) {alternative products}
Context Similarity %5.3f=>0.41629629629629633 <= U19718_at => synonyms are MFAP2 Microfibrillar-associated protein 2
Context Similarity %5.3f=>0.3785004516711834 <= U19796_at => synonyms are Melanoma antigen p15 mRNA
Context Similarity %5.3f=>0.43407407407407406 <= U19878_at => synonyms are Transmembrane protein mRNA
Context Similarity %5.3f=>0.43304843304843305 <= U19906_at => synonyms are VASOPRESSIN V1A RECEPTOR
Context Similarity %5.3f=>0.0 <= U19948_at => synonyms are Protein disulfide isomerase (PDIp) mRNA
Context Similarity %5.3f=>0.2578347578347578 <= U19977_at => synonyms are Preprocarboxypeptidase A2 (proCPA2) mRNA
Context Similarity %5.3f=>0.42407407407407405 <= U20158_at => synonyms are 76 kDa tyrosine phosphoprotein SLP-76 mRNA
Context Similarity %5.3f=>0.42328042328042326 <= U20230_at => synonyms are "GB DEF = Guanyl cyclase C gene, partial cds"
Context Similarity %5.3f=>0.37777777777777777 <= U20240_at => synonyms are "CEBPG CCAAT/enhancer binding protein (C/EBP), gamma"
Context Similarity %5.3f=>0.4199860237596087 <= U20285_at => synonyms are Gps1 (GPS1) mRNA
Context Similarity %5.3f=>0.0 <= U20325_at => synonyms are Cocaine and amphetamine regulated transcript CART (hCART) mRNA
Context Similarity %5.3f=>0.41816009557945044 <= U20350_at => synonyms are CMKRL1 Chemokine receptor-like 1
Context Similarity %5.3f=>0.38078703703703703 <= U20362_at => synonyms are Tg737 mRNA
Context Similarity %5.3f=>0.4037037037037037 <= U20391_rna6_at => synonyms are Folate receptor (FOLR1) gene
Context Similarity %5.3f=>0.32936507936507936 <= U20428_at => synonyms are SNC19 mRNA sequence
Context Similarity %5.3f=>0.0<= U20530_at => synonyms are GB DEF = Bone phosphoprotein spp-24 precursor mRNA
Context Similarity %5.3f=>0.37703703703703706

Correlation Distance Metric:
Correlation Distances:0.5246662304909925
Correlation Distances:0.5619362422999764
Correlation Distances:0.6513947712224407
Correlation Distances:0.48759512587181975
Correlation Distances:0.5319049159237761
Correlation Distances:0.5246662304909925
Correlation Distances:0.5619362422999764
Correlation Distances:0.6513947712224407
Correlation Distances:0.5319049159237761
Correlation Distances:0.5246662304909925
Correlation Distances:0.5619362422999764
Correlation Distances:0.6513947712224407
Correlation Distances:0.5319049159237761
Correlation Distances:0.5619362422999764
Correlation Distances:0.6221864849879517
Correlation Distances:0.6058234775837336

COLON DATASET: Gene Clustering Results
KEY IS M55265 Value is "Human casein kinase II alpha subunit mRNA, complete cds.
KEY IS T55871 Value is "HEXOKINASE, TYPE I (Homo sapiens)
KEY IS T56350 Value is NUCLEOLIN (Rattus norvegicus)
KEY IS R36644 Value is ACTIVIN RECEPTOR TYPE IIB PRECURSOR (Xenopus laevis)
KEY IS M83664 Value is "Human MHC class II lymphocyte antigen (HLA-DP) beta chain mRNA, complete cds.
KEY IS L26050 Value is "Human mitochondrial 2,4-dienoyl-CoA reductase mRNA, complete cds.
KEY IS X15943 Value is Huamn calcitonin/alpha-CGRP gene.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

KEY IS R60168 Value is HYPOTHETICAL 64.3 KD GTP-BINDING PROTEIN C02F5.3 IN CHROMOSOME III (Caenorhabditis elegans)
KEY IS U10868 Value is "Human aldehyde dehydrogenase ALDH7 mRNA, complete cds.
KEY IS X80230 Value is H.sapiens mRNA (clone C-2k) mRNA for serine/threonine protein kinase.
KEY IS U14631 Value is "Human 11 beta-hydroxysteroid dehydrogenase type II mRNA, complete cds.
KEY IS R88749 Value is TRANSCRIPTION FACTOR BTF3 (Homo sapiens)
KEY IS H06245 Value is PHOSPHOLIPASE ADRAB-B PRECURSOR (Oryctolagus cuniculus)
KEY IS X07290 Value is Human HF.12 gene mRNA.
KEY IS D38537 Value is Human mRNA for protoporphyrinogen oxidase.
KEY IS R41561 Value is INSULIN-LIKE GROWTH FACTOR BINDING PROTEIN COMPLEX ACID LABILE CHAIN PRECURSOR (Rattus norvegicus)
KEY IS M87434 Value is "Human 71 kDa 2'5' oligoadenylate synthetase (p69 2-5A synthetase) mRNA, complete cds.
KEY IS H06524 Value is "GELSOLIN PRECURSOR, PLASMA (HUMAN);.
KEY IS R22953 Value is "CASEIN KINASE I, ALPHA ISOFORM (Bos taurus)
KEY IS L06111 Value is "Human L-type voltage-gated calcium channel B subunit mRNA for isoform b, complete cds.
KEY IS T52806 Value is ANTIGEN KI-67 (Homo sapiens)
KEY IS X77130 Value is H.sapiens mRNA for ORL1 receptor.
KEY IS H41528 Value is STAGE V SPORULATION PROTEIN E (Bacillus subtilis)
KEY IS V00520 Value is Human germ line gene for growth hormone (presomatotropin).
KEY IS M22760 Value is "Homo sapiens nuclear-encoded mitochondrial cytochrome c oxidase Va subunit mRNA, complete cds.
KEY IS X83299 Value is H.sapiens SMA3 mRNA.
KEY IS X12892 Value is H.sapiens mRNA for protein S.
KEY IS M64673 Value is "Human heat shock factor 1 (TCF5) mRNA, complete cds.
KEY IS Y00661 Value is Human bcr mRNA (break point cluster gene).
KEY IS D38521 Value is "Human mRNA (KIAA0077) for ORF (novel product), partial cds.
KEY IS X80692 Value is H.sapiens ERK3 mRNA.
KEY IS U37690 Value is "Human RNA polymerase II subunit (hsRPB10) mRNA, complete cds.
KEY IS H74265 Value is LEUKOCYTE COMMON ANTIGEN PRECURSOR (Homo sapiens)
KEY IS H06061 Value is VOLTAGE-DEPENDENT ANION-SELECTIVE CHANNEL PROTEIN 1 (Homo sapiens)
KEY IS L34059 Value is NEURAL-CADHERIN PRECURSOR (HUMAN);
KEY IS H72850 Value is "5-AMINOLEVULINIC ACID SYNTHASE MITOCHONDRIAL PRECURSOR, (HUMAN);
KEY IS X60708 Value is Human pcHDP7 mRNA for liver dipeptidyl peptidase IV.
KEY IS X89985 Value is H.sapiens mRNA for BCL7B protein.
KEY IS T91121 Value is APOLIPOPROTEIN B-100 PRECURSOR (Homo sapiens)
KEY IS R67072 Value is GAP JUNCTION ALPHA-1 PROTEIN (Homo sapiens)
KEY IS T67173 Value is RETINOIC ACID RECEPTOR RXR-BETA ISOFORM 2 (Homo sapiens)
KEY IS U32519 Value is "Human GAP SH3 binding protein mRNA, complete cds.
KEY IS R71092 Value is EBNA-2 NUCLEAR PROTEIN (Epstein-barr virus)
KEY IS D26535 Value is "Human gene for dihydrolipoamide succinyltransferase, complete cds (exon 115).
KEY IS H55916 Value is "PEPTIDYL-PROLYL CIS-TRANS ISOMERASE, MITOCHONDRIAL PRECURSOR (HUMAN);.
KEY IS M84490 Value is "Human extracellular signal-regulated kinase 1 mRNA, 3' end.
KEY IS X69295 Value is H.sapiens MSX2 mRNA for transcription factor.
KEY IS L07648 Value is "Human MXI1 mRNA, complete cds.
KEY IS L13385 Value is "Homo sapiens(clone 71) Miller-Dieker lissencephaly protein (LIS1) mRNA, complete cds.
KEY IS H49515 Value is SIGNAL RECOGNITION PARTICLE 68 KD PROTEIN (Canis familiaris)
KEY IS K03192 Value is "Human cytochrome P-450 mRNA, partial.
KEY IS U33849 Value is "Human lymphoma proprotein convertase (LPC) mRNA, complete cds.
KEY IS L04953 Value is "Human x11 protein (x11) mRNA, 3' end.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

KEY IS M93009 Value is "Human L-isoaspartyl/D-aspartyl protein carboxyl methyltransferase isozyme I, mRNA, 3' end.
KEY IS R09502 Value is LAMININ BETA-1 CHAIN PRECURSOR (HUMAN);.
KEY IS H88876 Value is CD9 ANTIGEN (Homo sapiens)
KEY IS H13292 Value is RNA-BINDING PROTEIN FUS/TLS (Homo sapiens)
KEY IS R44057 Value is "PROTEIN PHOSPHATASE PP2A, 72 KD REGULATORY SUBUNIT (Homo sapiens)
KEY IS R44301 Value is MINERALOCORTICOID RECEPTOR (Homo sapiens)
KEY IS M85085 Value is "Human cleavage stimulation factor, complete cds.
KEY IS M84721 Value is "Human AMP deaminase (AMPD3) mRNA, complete cds.
KEY IS M58050 Value is "Human membrane cofactor protein (MCP) mRNA, complete cds.
KEY IS H02611 Value is ATP SYNTHASE A CHAIN (Trypanosoma brucei brucei)
KEY IS H21042 Value is CYCLIC-AMP-DEPENDENT TRANSCRIPTION FACTOR ATF-3 (Homo sapiens)
KEY IS M19156 Value is "Human acidic keratin-10 mRNA, complete cds.
KEY IS H29546 Value is NEUROTENSIN RECEPTOR (Homo sapiens)
KEY IS U25138 Value is "Human MaxiK potassium channel beta subunit mRNA, complete cds.
KEY IS L34657 Value is "Homo sapiens platelet/endothelial cell adhesion molecule-1 (PECAM-1) gene, exon 16 and complete cds.
KEY IS X07767 Value is Human mRNA for cAMP-dependent protein kinase catalytic subunit type alpha (EC 2.7.1.37).
KEY IS R99907 Value is INTERFERON REGULATORY FACTOR 2 (Homo sapiens)
KEY IS M76558 Value is "Human neuronal DHP-sensitive, voltage-dependent, calcium channel alpha-1D subunit mRNA, complete cds.
KEY IS U19969 Value is "Human two-handed zinc finger protein ZEB mRNA, partial cds.
KEY IS X64229 Value is H.sapiens dek mRNA.
KEY IS U37673 Value is "Human neuron-specific vesicle coat protein and cerebellar degeneration antigen (beta-NAP) mRNA, complete cds.
KEY IS X56597 Value is Human humFib mRNA for fibrillarin.
KEY IS R53455 Value is SERINE CARBOXYPEPTIDASE I PRECURSOR (Hordeum vulgare)
KEY IS U01038 Value is "Human pLK mRNA, complete cds.
KEY IS D43947 Value is "Human mRNA (KIAA0100) for ORF (human counterpart of mouse e1 gene), complete cds.
KEY IS X54941 Value is H.sapiens ckshs1 mRNA for Cks1 protein homologue.
KEY IS R23907 Value is "Human mRNA for PIG-F (phosphatidyl-inositol-glycan class F), complete cds.
KEY IS R90908 Value is PUTATIVE SERINE/THREONINE-PROTEIN KINASE T17E9.1 IN CHROMOSOME III (Caenorhabditis elegans)
KEY IS U12134 Value is "Human DNA damage repair and recombination protein RAD52 mRNA, complete cds.
KEY IS L25851 Value is INTEGRIN ALPHA-E PRECURSOR (HUMAN);contains Alu repetitive element;.
KEY IS U10886 Value is "Human density enhanced phosphatase-1 mRNA, complete cds.
KEY IS R45442 Value is STE6 PROTEIN (Schizosaccharomyces pombe)
KEY IS R59583 Value is PRE-MRNA SPLICING FACTOR SRP75 (Homo sapiens)
KEY IS R16156 Value is "RED CELL ACID PHOSPHATASE 1, ISOZYME F (Homo sapiens)
KEY IS R74066 Value is DNA DAMAGE RESPONSE PROTEIN KINASE DUN1 (Saccharomyces cerevisiae)
KEY IS U26710 Value is "Human cbl-b mRNA, complete cds.
KEY IS H87193 Value is HETEROGENEOUS NUCLEAR RIBONUCLEOPROTEIN K (Homo sapiens)
KEY IS T41204 Value is "P14780 92 KD TYPE V COLLAGENASE PRECURSOR ,.
KEY IS M91463 Value is "Human glucose transporter (GLUT4) gene, complete cds.
KEY IS X77548 Value is H. sapiens cDNA for RFG.
KEY IS J00146 Value is Human dihydrofolate reductase pseudogene (psi-hd1).
KEY IS J04102 Value is "Human erythroblastosis virus oncogene homolog 2 (ets-2) mRNA, complete cds.
KEY IS M20543 Value is "Human skeletal alpha-actin gene, complete cds.
KEY IS R20666 Value is PROBABLE G PROTEIN-COUPLED RECEPTOR EDG-1 (Homo sapiens)
KEY IS R42029 Value is "DIHYDROPRYRIDINE-SENSITIVE L-TYPE,

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

SKELETAL MUSCLE CALCIUM CHANNEL BETA SUBUNIT (Oryctolagus cuniculus)
KEY IS H64807 Value is PLACENTAL FOLATE TRANSPORTER (Homo sapiens)
KEY IS M25108 Value is "Human integrin beta-3 subunit mRNA, 3' end.
KEY IS M25809 Value is "VACUOLAR ATP SYNTHASE SUBUNIT B, KIDNEY ISOFORM (HUMAN);.
KEY IS U36621 Value is "Human Y-chromosome RNA recognition motif protein (YRRM) gene, exon 12, partial cds, subclone 7S2.
KEY IS R78595 Value is TUBULIN ALPHA CHAIN (Lytechinus pictus)
KEY IS H20503 Value is CELL ADHESION PROTEIN (HUMAN);.
KEY IS D28137 Value is "Human mRNA for BST-2, complete cds.
KEY IS M81695 Value is "LEUKOCYTE ADHESION GLYCOPROTEIN P150,95 ALPHA CHAIN (HUMAN);contains Alu repetitive element;contains element MER22 repetitive element ;.
KEY IS L40380 Value is "Homo sapiens thyroid receptor interactor (TRIP11) mRNA, 3' end of cds.
KEY IS T77537 Value is PLASMINOGEN (Sus scrofa)
KEY IS X76057 Value is MANNOSE-6-PHOSPHATE ISOMERASE (HUMAN);.
KEY IS X62048 Value is H.sapiens Wee1 hu gene.
KEY IS X85750 Value is H.sapiens mRNA for transcript associated with monocyte to macrophage differentiation.
KEY IS H74178 Value is RAPAMYCIN-SELECTIVE 25 KD IMMUNOPHILIN (Homo sapiens)
KEY IS U31248 Value is "Human zinc finger protein (ZNF174) mRNA, complete cds.
KEY IS X70297 Value is "NEURONAL ACETYLCHOLINE RECEPTOR PROTEIN, ALPHA-7 CHAIN (HUMAN);.
KEY IS M81840 Value is "Human NRL gene product mRNA, complete cds.
KEY IS D90041 Value is Human liver arylamine N-acetyltransferase (EC 2.3.1.5) gene.
KEY IS R67999 Value is PROBABLE ATP-DEPENDENT RNA HELICASE PRH1 (Schizosaccharomyces pombe)
KEY IS D29808 Value is "Human mRNA for T-cell acute lymphoblastic leukemia associated antigen 1 (TALLA-1), complete cds.
KEY IS Z33642 Value is H.sapiens V7 mRNA for leukocyte surface protein.
KEY IS M69181 Value is "MYOSIN HEAVY CHAIN, NONMUSCLE TYPE B (HUMAN);.
KEY IS L38951 Value is "Homo sapiens importin beta subunit mRNA, complete cds.
KEY IS R44112 Value is 60S RIBOSOMAL PROTEIN L35A (HUMAN);.
KEY IS X56411 Value is "H.sapiens ADH4 gene for class II alcohol dehydrogenase (pi subunit), exon 1.
KEY IS Y00796 Value is Human mRNA for leukocyte-associated molecule-1 alpha subunit (LFA-1 alpha subunit).
KEY IS M14603 Value is "Human myoglobin gene, exon 3.
KEY IS X73882 Value is H.sapiens E-MAP-115 mRNA.
KEY IS T78323 Value is PROCOLLAGEN ALPHA 1(IV) CHAIN PRECURSOR (HUMAN);.
KEY IS M80359 Value is PUTATIVE SERINE/THREONINE-PROTEIN KINASE P78 (HUMAN);contains MSR1 repetitive element ;.
KEY IS M35531 Value is "Human GDP-L-fucose:beta-D-galactoside 2-alpha-l-fucosyltransferase mRNA, complete cds.
KEY IS R35903 Value is "INTERLEUKIN-1 RECEPTOR, TYPE II PRECURSOR (Mus musculus)
KEY IS X67325 Value is H.sapiens p27 mRNA.
KEY IS U25435 Value is "Human transcriptional repressor (CTCF) mRNA, complete cds.
KEY IS D37931 Value is Human mRNA for RNase 4.
KEY IS M15841 Value is U2 SMALL NUCLEAR RIBONUCLEOPROTEIN B' (HUMAN);.
KEY IS D59253 Value is Human mRNA for NCBP interacting protein 1.
KEY IS H77510 Value is COAGULATION FACTOR VIII PRECURSOR (Mus musculus)
KEY IS R78142 Value is PUTATIVE SERINE/THREONINE-PROTEIN KINASE B0464.5 IN CHROMOSOME III (Caenorhabditis elegans)
KEY IS R62945 Value is COMPLEMENT DECAY-ACCELERATING FACTOR 1 PRECURSOR (Homo sapiens)
KEY IS X70070 Value is H.sapiens mRNA for neurotensin receptor.
KEY IS U05040 Value is "Human FUSE binding protein mRNA, complete cds.
KEY IS U33429 Value is "human K+ channel beta 2 subunit mRNA, complete cds.
KEY IS T60778 Value is MATRIX GLA-PROTEIN PRECURSOR (Rattus norvegicus)
KEY IS X04011 Value is Human mRNA of X-CGD gene involved in chronic granulomatous disease located on chromosome X.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

KEY IS L12350 Value is THROMBOSPONDIN 2 PRECURSOR (HUMAN);.
KEY IS R06749 Value is ERYTHROID KRUEPPEL-LIKE TRANSCRIPTION FACTOR (Mus musculus)
KEY IS U24105 Value is "Human coatomer protein (HEPCOP) mRNA, complete cds.
KEY IS T57468 Value is FIBRILLARIN (HUMAN).
KEY IS M55422 Value is "Human Krueppel-related zinc finger protein (H-plk) mRNA, complete cds.
KEY IS M64110 Value is "Human caldesmon mRNA, complete cds.
KEY IS U14577 Value is "Human microtubule-associated protein 1A (MAP1A) mRNA, complete cds.
KEY IS K03474 Value is "Human Mullerian inhibiting substance gene, complete cds.
Correlation Distances:0.38095074890042807
Correlation Distances:0.36577387497875224
Correlation Distances:0.3439398818311977
Correlation Distances:0.3251439700491894
Correlation Distances:0.39444770140553403
Correlation Distances:0.33925300493694815
Correlation Distances:0.31967752515416586
Correlation Distances:0.3632866502755317
Correlation Distances:0.3499089442399207
Correlation Distances:0.404147302046919
Correlation Distances:0.3935131330235525
Correlation Distances:0.306499184888537
Correlation Distances:0.3845016757054823
Correlation Distances:0.41852476492710133
Correlation Distances:0.42788819215800344
Correlation Distances:0.40142741109302194
Correlation Distances:0.411835509172185
Correlation Distances:0.2785849249459654
Correlation Distances:0.3586472254202513
Correlation Distances:0.3558281265517634
Correlation Distances:0.2795434747558908
Correlation Distances:0.28236954402569503
Correlation Distances:0.35418876246207254
Correlation Distances:0.29620726286842414
Correlation Distances:0.29207826408674215
Correlation Distances:0.37230741829164443
Correlation Distances:0.3689437046157218
Correlation Distances:0.33790026018342795
Correlation Distances:0.35645611821679
Correlation Distances:0.386541798555382
Correlation Distances:0.3139985821171466
Correlation Distances:0.3203503245281585
Correlation Distances:0.3330626180322059
Correlation Distances:0.32295384258936416
Correlation Distances:0.43415243065794834
Correlation Distances:0.3555228509599469
Correlation Distances:0.4197713585381581
Correlation Distances:0.4259477914677696
Correlation Distances:0.31590251245730916
Correlation Distances:0.3271682206145489
Correlation Distances:0.3951900458830304
Correlation Distances:0.31344594772797907
Correlation Distances:0.3156485234851386
Correlation Distances:0.3661652632264971
Correlation Distances:0.3752734021166865
Correlation Distances:0.3384104177054316
Correlation Distances:0.3035002399703193
Correlation Distances:0.3490252045429707
Correlation Distances:0.3854818580373486
Correlation Distances:0.46282088550520406
Correlation Distances:0.390448653385745
Correlation Distances:0.4240499084875071
Correlation Distances:0.3137535520836849
Correlation Distances:0.31297914865718446
Correlation Distances:0.3471814256362601
Correlation Distances:0.31114527584019736
Correlation Distances:0.32601500018974056
Correlation Distances:0.3300983784420102
Correlation Distances:0.3855974176835045
Correlation Distances:0.32828456931899425
Correlation Distances:0.3714625545786509
Correlation Distances:0.38391393004152585
Correlation Distances:0.4072211425925476
Correlation Distances:0.36866559530743886
Correlation Distances:0.3914034819128684
Correlation Distances:0.4412827367731554
Correlation Distances:0.29880178053530493
Correlation Distances:0.31006931127374476
Correlation Distances:0.35509285298332766
Correlation Distances:0.31114527584019736
Correlation Distances:0.29196336411262136
Correlation Distances:0.4116879519391564
Correlation Distances:0.35666151811662233
Correlation Distances:0.38095074890042807
Correlation Distances:0.36577387497875224
Correlation Distances:0.3439398818311977
Correlation Distances:0.3251439700491894
Correlation Distances:0.3583634818013692

MLL DATASET

32403_at"Cluster Incl. U86813:Homo sapiens serotonin-7 receptor pseudogene, complete sequence /cds=(0,464) /gb=U86813 /gi=3138916 /ug=Hs.234784 /len=1326"
32404_at"Cluster Incl. AF065314:Homo sapiens cone photoreceptor cGMP-gated channel alpha subunit (CNGA3) mRNA, complete cds /cds=(39,2123) /gb=AF065314 /gi=3153886 /ug=Hs.234785 /len=3469"

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

32405_at"Cluster Incl. AB014607:Homo sapiens mRNA for KIAA0707 protein, partial cds /cds=(0,1894) /gb=AB014607 /gi=3327227 /ug=Hs.234786 /len=6359"
32406_at"Cluster Incl. AB020696:Homo sapiens mRNA for KIAA0889 protein, complete cds /cds=(121,1677) /gb=AB020696 /gi=4240266 /ug=Hs.234791 /len=4122"
32407_f_at"Cluster Incl. U92818:Homo sapiens c33.28 unnamed HERV-H protein mRNA, partial cds /cds=(0,298) /gb=U92818 /gi=2465329 /ug=Hs.239501 /len=432"
32408_s_at"Cluster Incl. AL022101:dJ845O24.4 (Heterogenous Nuclear Ribonucleoprotein HNRNP C1 LIKE protein) /cds=(100,981) /gb=AL022101 /gi=3171895 /ug=Hs.239530 /len=1299"
32409_at"Cluster Incl. AC004472:Homo sapiens chromosome 9, P1 clone 11659 /cds=(0,2642) /gb=AC004472 /gi=2984582 /ug=Hs.239950 /len=2643"
32410_at"Cluster Incl. X17651:Human Myf-4 mRNA for myogenic determination factor /cds=(52,792) /gb=X17651 /gi=34831 /ug=Hs.2830 /len=1418"
32411_at"Cluster Incl. X68561:H.sapiens SPR-1 mRNA for GT box binding protein /cds=(181,2535) /gb=X68561 /gi=38419 /ug=Hs.2982 /len=2986"
32412_at"Cluster Incl. M13934:Human ribosomal protein S14 gene, complete cds /cds=(2,457) /gb=M13934 /gi=337498 /ug=Hs.3491 /len=503"
32413_at"Cluster Incl. M13934:Human ribosomal protein S14 gene, complete cds /cds=(0,494) /gb=M13934 /gi=337498 /ug=Hs.3491 /len=495"
32414_at"Cluster Incl. U05589:Human ribosomal protein S1 homolog mRNA, partial cds /cds=(0,1220) /gb=U05589 /gi=497001 /ug=Hs.371 /len=1478"
32415_at"Cluster Incl. V00541:Messenger RNA for human leukocyte interferon (one of eight) /cds=(0,401) /gb=V00541 /gi=32718 /ug=Hs.37113 /len=763"
32416_at"Cluster Incl. L48728:Homo sapiens T cell receptor beta (TCRBV10S1) gene, complete cds /cds=(0,83) /gb=L48728 /gi=1054550 /ug=Hs.37163 /len=348"
32417_at"Cluster Incl. D17427:Human mRNA for desmocollin type 4 /cds=(67,2757) /gb=D17427 /gi=639672 /ug=Hs.41690 /len=3552"
32418_at"Cluster Incl. U40371:Human 3,5 cyclic nucleotide phosphodiesterase (HSPDE1C1A) mRNA, complete cds /cds=(176,2080) /gb=U40371 /gi=1151110 /ug=Hs.41718 /len=2694"
32419_at"Cluster Incl. U40372:Human 3,5 cyclic nucleotide phosphodiesterase (HSPDE1C3A) mRNA, partial cds /cds=(0,1695) /gb=U40372 /gi=1151112 /ug=Hs.41718 /len=2076"
32420_at"Cluster Incl. U18549:Human GPR6 G protein-coupled receptor gene, complete cds /cds=(18,1106) /gb=U18549 /gi=604501 /ug=Hs.46332 /len=1614"
32421_at"Cluster Incl. M90359:Human cAMP-dpendent protein kinase (AKAP 79) mRNA, complete cds /cds=(1297,2580) /gb=M90359 /gi=178323 /ug=Hs.48714 /len=2604"
32422_at"Cluster Incl. D70830:Homo sapiens mRNA for Doc2 beta, complete cds /cds=(160,1398) /gb=D70830 /gi=1235721 /ug=Hs.54402 /len=2043"
32423_at"Cluster Incl. U48408:Human kidney water channel (hKID) mRNA, complete cds /cds=(342,1190) /gb=U48408 /gi=1293545 /ug=Hs.54505 /len=1347"
32424_at"Cluster Incl. D84424:Homo sapiens mRNA for hyaluronan synthase, complete cds /cds=(148,1779) /gb=D84424 /gi=1401033 /ug=Hs.57697 /len=2096"
KEY IS 36629_at Value is "Cluster Incl. AI635895:tz82a07.x1 Homo sapiens cDNA, 3 end /clone=IMAGE-2295060 /clone_end=3 /gb=AI635895 /gi=4687225 /ug=Hs.75450 /len=1082"
KEY IS 36630_at Value is "Cluster Incl. Z50781:H.sapiens mRNA for leucine zipper protein /cds=(135,368) /gb=Z50781 /gi=1834506 /ug=Hs.75450 /len=420"
KEY IS 36631_at Value is "Cluster Incl. D49396:Human mRNA for Apo1_Human (MER5(Aop1-Mouse)-like protein), complete cds /cds=(6,776) /gb=D49396 /gi=682747 /ug=Hs.75454 /len=1531"
KEY IS 36632_at Value is Cluster Incl. U00957:Human clone KDB1.2 (CAC)n/(GTG)n repeat-containing mRNA /cds=UNKNOWN /gb=U00957 /gi=405059 /ug=Hs.75456 /len=2197
KEY IS 36633_at Value is "Cluster Incl. AA114830:zk88e06.s1 Homo sapiens cDNA, 3 end /clone=IMAGE-489922 /clone_end=3 /gb=AA114830 /gi=1669952 /ug=Hs.75456 /len=626"
KEY IS 36634_at Value is "Cluster Incl. U72649:Human BTG2 (BTG2) mRNA, complete cds /cds=(71,547) /gb=U72649 /gi=1703500 /ug=Hs.75462 /len=2717"
KEY IS 36635_at Value is "Cluster Incl. AB023173:Homo sapiens mRNA for KIAA0956 protein, partial cds /cds=(0,2020) /gb=AB023173 /gi=4589555 /ug=Hs.75478 /len=5542"
KEY IS 36636_at Value is "Cluster Incl. M12267:Human ornithine aminotransferase

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

mRNA, complete cds /cds=(54,1373) /gb=M12267 /gi=189328 /ug=Hs.75485 /len=2013"
KEY IS 36637_at Value is "Cluster Incl. L19605:Homo sapiens 56K autoantigen annexin XI gene mRNA, complete cds /cds=(178,1695) /gb=L19605 /gi=457128 /ug=Hs.75510 /len=1958"
KEY IS 36638_at Value is "Cluster Incl. X78947:H.sapiens mRNA for connective tissue growth factor /cds=(145,1194) /gb=X78947 /gi=474933 /ug=Hs.75511 /len=2312"
KEY IS 36639_at Value is "Cluster Incl. AF067853:Homo sapiens adenylosuccinate lyase (ADSL) mRNA, alternatively spliced, complete cds /cds=(55,1509) /gb=AF067853 /gi=3211981 /ug=Hs.75527 /len=1734"
KEY IS 36640_at Value is "Cluster Incl. X66141:H.sapiens mRNA for cardiac ventricular myosin light chain-2 /cds=(30,530) /gb=X66141 /gi=34845 /ug=Hs.75535 /len=784"
KEY IS 36641_at Value is "Cluster Incl. U03851:Human capping protein alpha mRNA, partial cds /cds=(16,870) /gb=U03851 /gi=433307 /ug=Hs.75546 /len=2263"
KEY IS 36642_at Value is "Cluster Incl. J00287:Human pepsinogen gene /cds=(55,1221) /gb=J00287 /gi=189798 /ug=Hs.75558 /len=1381"
KEY IS 36643_at Value is "Cluster Incl. L20817:Homo sapiens tyrosine protein kinase (CAK) gene, complete cds /cds=(192,2933) /gb=L20817 /gi=306474 /ug=Hs.75562 /len=3774"
KEY IS 36644_at Value is "Cluster Incl. D29963:Homo sapiens mRNA for CD151, complete cds /cds=(84,845) /gb=D29963 /gi=2073384 /ug=Hs.75564 /len=1486"

Context Similarity %5.3f=>0.3581215560490172 <= AFFX-HUMISGF3A/M97935_5_at => synonyms are "M97935 Homo sapiens transcription factor ISGF-3 mRNA, complete cds (_5, _MA, MB, _3 represent transcript regions 5 prime, MiddleA, MiddleB, and 3 prime respectively)"
Context Similarity %5.3f=>0.5174961432446462 <= AFFX-HUMISGF3A/M97935_MA_at => synonyms are "M97935 Homo sapiens transcription factor ISGF-3 mRNA, complete cds (_5, _MA, MB, _3 represent transcript regions 5 prime, MiddleA, MiddleB, and 3 prime respectively)"
Context Similarity %5.3f=>0.4916516174001204 <= AFFX-HUMISGF3A/M97935_MB_at => synonyms are "M97935 Homo sapiens transcription factor ISGF-3 mRNA, complete cds (_5, _MA, MB, _3 represent transcript regions 5 prime, MiddleA, MiddleB, and 3 prime respectively)"
Context Similarity %5.3f=>0.49733529573848934 <= AFFX-HUMISGF3A/M97935_3_at => synonyms are "M97935 Homo sapiens transcription factor ISGF-3 mRNA, complete cds (_5, _MA, MB, _3 represent transcript regions 5 prime, MiddleA, MiddleB, and 3 prime respectively)"
Context Similarity %5.3f=>0.5016231273716304 <= AFFX-HUMRGE/M10098_5_at => synonyms are "M10098 Human 18S rRNA gene, complete (_5, _M, _3 represent transcript regions 5 prime, Middle, and 3 prime respectively)"
Context Similarity %5.3f=>0.4361526886829809 <= AFFX-HUMRGE/M10098_M_at => synonyms are "M10098 Human 18S rRNA gene, complete (_5, _M, _3 represent transcript regions 5 prime, Middle, and 3 prime respectively)"
Context Similarity %5.3f=>0.4316261073210468 <= AFFX-HUMRGE/M10098_3_at => synonyms are "M10098 Human 18S rRNA gene, complete (_5, _M, _3 represent transcript regions 5 prime, Middle, and 3 prime respectively)"
Context Similarity %5.3f=>0.4361526886829809 <= AFFX-HUMGAPDH/M33197_5_at => synonyms are "M33197 Human glyceraldehyde-3-phosphate dehydrogenase (GAPDH) mRNA, complete cds (_5, _M, _3 represent transcript regions 5 prime, Middle, and 3 prime respectively)"
Context Similarity %5.3f=>0.4734939759036145 <= AFFX-HUMGAPDH/M33197_M_at => synonyms are "M33197 Human glyceraldehyde-3-phosphate dehydrogenase (GAPDH) mRNA, complete cds (_5, _M, _3 represent transcript regions 5 prime, Middle, and 3 prime respectively)"
Context Similarity %5.3f=>0.46692454026632846 <= AFFX-HUMGAPDH/M33197_3_at => synonyms are "M33197 Human glyceraldehyde-3-phosphate dehydrogenase (GAPDH) mRNA, complete cds (_5, _M, _3 represent transcript regions 5 prime, Middle, and 3 prime respectively)"
Context Similarity %5.3f=>0.49016064257028114 <= AFFX-HSAC07/X00351_5_at => synonyms are "X00351 Human mRNA for beta-actin (_5, _M, _3 represent transcript regions 5 prime, Middle, and 3 prime respectively)"
Context Similarity %5.3f=>0.4329227637996982 <= AFFX-HSAC07/X00351_M_at => synonyms are "X00351 Human mRNA for beta-actin (_5, _M, _3 represent transcript regions 5 prime, Middle, and 3 prime respectively)"
Context Similarity %5.3f=>0.4329227637996982 <= AFFX-HSAC07/X00351_3_at => synonyms are "X00351 Human mRNA for beta-actin (_5, _M, _3 represent transcript regions 5 prime, Middle, and 3 prime respectively)"
Context Similarity %5.3f=>0.4329227637996982

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

<= AFFX-HUMTFRR/M11507_5_at => synonyms are "M11507 Human transferrin receptor mRNA, complete cds (_5, _M, _3 represent transcript regions 5 prime, Middle, and 3 prime respectively)"
Context Similarity %5.3f=>0.4223429951690821
<= AFFX-HUMTFRR/M11507_M_at => synonyms are "M11507 Human transferrin receptor mRNA, complete cds (_5, _M, _3 represent transcript regions 5 prime, Middle, and 3 prime respectively)"
Context Similarity %5.3f=>0.4187370600414078
<= AFFX-HUMTFRR/M11507_3_at => synonyms are "M11507 Human transferrin receptor mRNA, complete cds (_5, _M, _3 represent transcript regions 5 prime, Middle, and 3 prime respectively)"
Context Similarity %5.3f=>0.4223429951690821
<= AFFX-M27830_5_at => synonyms are "M27830 Human 28S ribosomal RNA gene, complete cds (_5, _M, _3 represent transcript regions 5 prime, Middle, and 3 prime respectively)"
Context Similarity %5.3f=>0.4740740740740741
<= AFFX-M27830_M_at => synonyms are "M27830 Human 28S ribosomal RNA gene, complete cds (_5, _M, _3 represent transcript regions 5 prime, Middle, and 3 prime respectively)"
Context Similarity %5.3f=>0.4740740740740741
<= AFFX-M27830_3_at => synonyms are "M27830 Human 28S ribosomal RNA gene, complete cds (_5, _M, _3 represent transcript regions 5 prime, Middle, and 3 prime respectively)"
Context Similarity %5.3f=>0.4740740740740741
<= AFFX-HSAC07/X00351_3_st => synonyms are "X00351 Human mRNA for beta-actin (_5, _M, _3 represent transcript regions 5 prime, Middle, and 3 prime respectively)"
Context Similarity %5.3f=>0.4329227637996982
<= AFFX-HUMGAPDH/M33197_5_st => synonyms are "M33197 Human glyceraldehyde-3-phosphate dehydrogenase (GAPDH) mRNA, complete cds (_5, _M, _3 represent transcript regions 5 prime, Middle, and 3 prime respectively)"
Context Similarity %5.3f=>0.4734939759036145
<= AFFX-HUMGAPDH/M33197_M_st => synonyms are "M33197 Human glyceraldehyde-3-phosphate dehydrogenase (GAPDH) mRNA, complete cds (_5, _M, _3 represent transcript regions 5 prime, Middle, and 3 prime respectively)"
Context Similarity %5.3f=>0.46692454026632846
<= AFFX-HUMGAPDH/M33197_3_st => synonyms are "M33197 Human glyceraldehyde-3-phosphate dehydrogenase (GAPDH) mRNA, complete cds (_5, _M, _3 represent transcript regions 5 prime, Middle, and 3 prime respectively)"
Context Similarity %5.3f=>0.49016064257028114
<= AFFX-HSAC07/X00351_5_st => synonyms are "X00351 Human mRNA for beta-actin (_5, _M, _3 represent transcript regions 5 prime, Middle, and 3 prime respectively)"
Context Similarity %5.3f=>0.40911323999017446
<= AFFX-HSAC07/X00351_M_st => synonyms are "X00351 Human mRNA for beta-actin (_5, _M, _3 represent transcript regions 5 prime, Middle, and 3 prime respectively)"
Context Similarity %5.3f=>0.4329227637996982

5. COMPARATIVE ANALYSIS

Comparative Analysis of the Gene Synonyms Detection Rate and Medical Data Gene Similarity Rate using NNGE, IBL, RBHCA (proposed) Algorithms on Leukemia, Colon, MLL, Lymphoma Diseases. Table:1 shows the Gene Synonyms Detection Rate. Table:2 shows the Medical Data Gene Similarity Rate.

TABLE:1 Gene Synonyms Detection Rate

Algorithms	*Leukemia*	*Colon*	*MLL*	*Lymphoma*
NNGE	0.68	0.72	0.78	0.82
IBL	0.69	0.63	0.83	0.79
RBHCA (Proposed)	0.94	0.96	0.89	0.92

TABLE:2 Medical Data Gene Similarity Rate

Algorithms	*Leukemia*	*Colon*	*MLL*	*Lymphoma*
NNGE	0.81	0.78	0.81	0.86
IBL	0.77	0.84	0.9	0.79
RBHCA (Proposed)	0.98	0.975	0.92	0.88

6. PERFORMANCE ANALYSIS DISCUSSION

In this work, each medical attribute along with synonyms are considered to find the most efficient clusters representation. In the traditional algorithms like NNGE and IBL, attribute synonyms are not consider to find the relevant gene/protein attribute relationships. Proposed approach takes less search space compare to traditional techniques due to filtering and synonyms identification process. Proposed approach was executed on different medical datasets like leukemia, colon, mll and lymphoma to find the highest similarity and

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

gene\protein synonym detection. Experimental result on different datasets gives high true positive cluster rate compare to traditional NNGE and IBL approaches.

PERFORMANCE ANALYSIS:

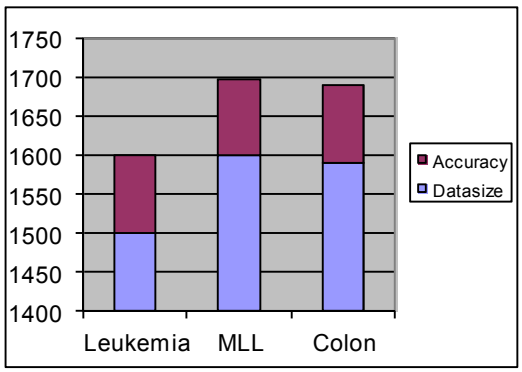

Fig 4: Comparison Between Data size And Accuracy In Different Datasets

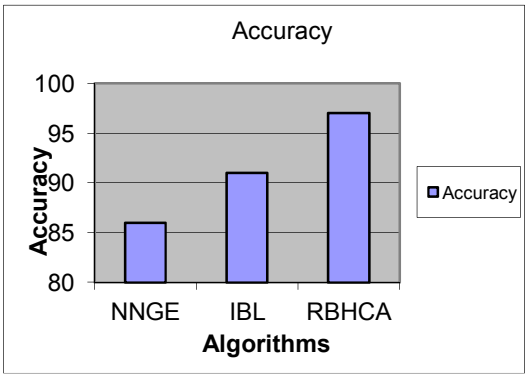

Fig 5: Comparison between proposed and traditional algorithms for leukemia dataset

7. CONCLUSION

In this proposed work context rank based hierarchical clustering method is applied on different datasets namely colon, Leukemia, MLL, Lymphoma medical diseases. The experimental Results show that proposed method outperformed well in terms of time and clusters search space is concerned. In future this work can be extended to implement visualizing biomedical terms from abstracts using gene/protein features. Clustering each gene/protein(s) based on protein/gene id, synonyms, name, category, patterns and its description. Easy to get all relations of gene/protein(s) using graph based techniques.

8. REFRENCES:

[1] T.S. Bhatti, R.C. Bansal, and D.P. Kothari, "Reactive Power Control of Isolated Hybrid Power Systems", *Proceedings of International Conference on Computer Application in Electrical Engineering Recent Advances (CERA)*, Indian Institute of Technology Roorkee (India), February 21-23, 2002, pp. 626-632.

[1] B. F. Momin, S. Mitra, and R. D. Gupta, "Reduce Generation and Classification of Gene Expression Data," in Proceedings of the 2006 *International Conference on Hybri Information Technology*, pp. 699-708, 2006.

[2] Jung-HsienChiang, Senior Member, IEEE, and Shing-HuaHo,"A Combination of Rough-Based Feature Selection and RBF Neural Network for Classification Using Gene Expression Data", *Ieee Transactions OnNanobioscience*, Vol.7, No.1, March 2008, pp:91-99.

[3] Masser, M.B., White, M. Katherine, Hyde and K. Melissa *et al*., "Predicting blood donation intentions and behavior among Australian blood donors: Testing an extended theory of planned Behavior model" ,Transfusion, 49(2),2009, pp. 320-329.

[4] S. Gopal, A. Haake, R. P. Jones *et al.*, *Bioinformatics: a computing perspective*, Int.Ed. ed.: McGraw-Hill Higher Educ, 2009.

[5] Anil Rajput, RameshPrasadAharwal, Nidhi Chandel, Devenra Singh Solanki and Ritu Soni,"Approaches of Classifications to Policy of Analysis of Medical Data" *IJCSNS International Journal of Computer Science and Network Security",* VOL.9 No.11, November 2009,pp. 01-09.

[6] T. Santhanam and Shyam Sundaram, "Application of CART Algorithm in Blood Donors Classification**" ,** *Journal of Computer Science* 2010 ISSN 1549-3636 Vol6 (5): PP 548-552,.

[7] Rossen Dimov et al., Weka: Practical machine Learning Tools and Techniques -April 30, 2010.

[8] ZhiwenYu, Hau-SanWongb, JaneYou, QinminYang, and Hongying Liao, *"Knowledge Based Cluster Ensemble for Cancer Discovery from Biomolecular Data", Ieee Transactions On Nanobioscience*, Vol.10, No.2, pp:76-85,June2011.

[9] Devchand J Chaudhari, Mamta Ramteke and Manoj G Lade. Article: Data Mining in Blood Platelets Transfusion using Classification Rule. *IJCA Proceedings on Emerging Trends in Computer Science and Information*

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Technology(ETCSIT2012)etcsit1001 Etcsit(2) pp:14-17, April 2012.

[10] Shahana Bano,Dr. K. Rajasekhara Rao "Key Word Based Word Sense Extraction In A Index For Text Files: Design Approach", *Ciit International Journal Of Data Mining And Knowledge Engineering* JAN'12.

[11] Shahana Bano, Dr. K. Rajasekhara Rao "Key Word Based Word Sense Extraction In Text: Design Approach", *International Journal Of Computer Science And Communication* March'12, pp:95-99.

[12] Shahana Bano, Dr. K. Rajasekhara Rao "Pattern Based Gene/Protein Synonyms Identification From Biological Databases*", International Journal Of Applied Engineering Research (IJAER)* Volume 9, No 12 (2014), pp:1815-1827.