A Frame Work For Identification Of Relationship Between Gene And Disease Causing Mutation Using Biological Text Mining

A.MURALI KRISHNA
Research scholar, Dept. of Computer Sceince
Rayalaseema University
Kurnool-518007.(A.P),INDIA.

Dr. S. JYOTHI
Professor, Dept. of Computer Science
Sri Pamavathi Mahila Visvavidyalayam
Tirupati-517502. AP, INDIA.

Abstract—**We have gone through various papers describing the mutations in between them and associated disease in a rapid pace. The articles of previous studies show that there is a need to acquire knowledge of gene mutation causing diseases and its association. The need cannot be solved manually, but it has to be automated, so our study is based to develop a framework which gathers information of disease association mutation for knowledge sharing to doctors and researchers. Our work is done using texting mining for extraction of disease causing mutation and its associated NLP from previous abstracts. Our proposed system extracts mutation causing gene using NLP.DMLtool consists of modules of NLP that process text input using semantic and synaptic patterns to gain disease mutation. DML developed gives recall and precision high with F-score 0.87 , 0.89 and 0.91, which were evaluated on 3 various datasets related to associated disease mutations. In DML we used a special module which extracts mentioned mutation and its gene text associated with it.**

Various types of datasets have been evaluated on our framework and its performance has been check with performance metric.The obtained results shows better performance compared to the existing on association of disease-mutation and also solves problems of low precision and their approaches.LMA is applied to large data sets of different type of abstracts in Pubmed, it extracts associated disease-mutations and its related information of patients, population of data and its type size.

The gained result from our work is stored in a database, which can be acquired by query processing. In our work we conclude that using text mining method, we can increase high throughput, this gives potential to the research and also assist the research in identifying mutation causing disease and it's associated with.

Keywords: **Mutation;Extraction;NLP;Text Mining;Pubmed;Gene;**

I. INTRODUCTION

Development of technology rapidly has evolved the rise of number of research articles on genomic association and its associated diseases. To collect this information manually it is tedious and very difficult. Nearly 10,000 articles related to genomic are published every year on disease and disease causing mutation.

Various type of literature related should not be kept as side; it has to be updated for further research and usage. To eliminate the manual process, and identify the various studies done in identification of disease causing mutation gene various types of text mining method are notified. Various studies show certain failure in detection of disease causing mutation gene. Previous studies show that various mining techniques have used and identified certain limitation in finding the disease causing mutation.With the utility of regular expression in finding mutation, this uses grammar in identification of meaning of mutation grammar. One of the such method in scanning the grammar of mutation associate phonotype. Basic search method is used finding the relationship between disease causing mutation gene and its associated diseases from the abstracts of pubmed[31].

Works has shown that, the combination of existing text mining and expression search method are associate in find high rate of mutation causing genes related to disease with the metric of low recall and precision high[32].Text mining method [16] is used to extract text of mutations from pub med abstracts, connect them to the gene associated and finds the disease associated with it.

Our study proposes text-mining approach in extraction of mutations from pubmed abstracts for identification of diseases associated with it. In relate to the co-occurrence of mutation associated with disease and available with the existing relationship in the abstract, we applied an extraction method in extracting the relationship between the mutations of disease and stored in a database for further analysis.

Existing and co-existing based approaches are use to search the frequency of occurrence of mutation causing gene and its associated is mined from various abstracts of articles. The basic process used is select a single research article, prepare a structure of sentence words together with the textual relate to disease causing mutation with the structure.The proposed system is developed using NLP, which syntactically process the text input in detecting various

types of mutations. We developed an algorithm identify the gene associated with mutation, which also contains an extraction module to process semantic patterns and lexical analysis to acquire relationships between diseases which are caused using mutations. patterns and lexical analysis to acquire relationships between diseases which are caused using mutations.diseases which are caused using mutations.

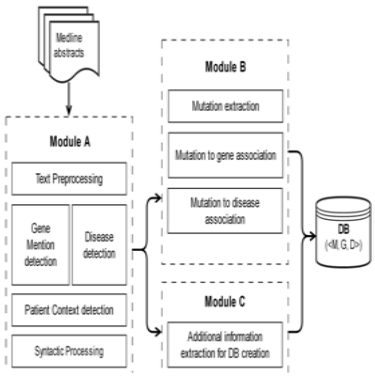

FIGURE 1: TEXTING MINING FOR BIOLOGICAL RELATIONSHIP FOR DISEASE IDENTIFICATION

Additional rules are also framedWe applied our system on roughly 10,000 abstracts. The extraction results are stored in a database, which can be queried and downloaded for further analysis. The database is accessible via a simple interface that we developed (http://biotm.cis.udel.edu/dimex/). Some sample queries showing the possible uses of the system are presented in the Results section of this article. We conducted an evaluation of a sample of extractions from the database and found it to be consistent with our evaluations mentioned above. This exercise illustrates the scalability and robustness of DiMeX.

Materials and Methods

The architecture of the system is shown in figure 1. The architecture is divided into 3 modules, namely A, B AND C. Module A is used to process text from fetching data from the abstracts of pubmed. It uses syntactic pre-processor for processing the text.In B module, mutation related to text is identified and its associated disease causing gene is associated.At last in Module C, required additional information related to association of the database and its

Module A

The basic components included in this module is preprocessing of text , tagging of disease related to gene, context patient detection sentence, pattern matching of syntactic processing.

Processing of Text

To process text we use Dimex , which accepts input from pubmed and process it and result the abstract input individually. In process, first we extract the title, text of abstract and key terms used in our mutation repository. The process used in splitting the sentence is in house splitter to split the sentences of abstract individually. Next, acronym identifier is used for identifying all possible abbreviation related to the gene and mutation tagging of diseases in steps.

Gene and disease tagging

In this we use a detector named[35] Mention Gene which is named as in house. To identify the mentioned detection, we identified a text mining tool which assists the user with biological entity tags in identification of words.Few enhancements cannot be performed to this tool which result in disease discarded acronym. "AR" is detected as a disease but the full form "Androgen Receptor" hints that it is a gene. Additionally, "tumor" or "cancer" might be mentioned in general throughout the abstract, without specifying the actual disease every time. These general occurrences, which are annotated by Pubtator as disease terms, are mapped to the closest mentioned disease, which often appears within the same noun phrase.

We detect mentioned patient sentence for monitoring the information related to the patients and their study. Likewise total number of persons and their information of demographic. In common sentences mentioned to the disease and their study. It is very important to show the relationship between the mutation and the disease causing gene is known explicitly by scanning the sentences associated with it.Next we identify the starting patients information from the abstract and the type of information related to. The sentence related to the context of patient is first extracted from the abstract .Ex : Total number of patients related to breast cancer 453 and age range 28- 45 and related sex Male and Female matched from Andhra Pradesh were analyzed". (PMID . 15330212)

Syntactic Processing

It is a process of tokeniz ing the terms used for parsing[36] , the tokens identify the phrases in sentences. We use NLP of BIONex detection method for identifying the phrases of nouns and group of verbs, these verbs are connection with conjunction, proposition or punctuation which are marked.We assign such sequence of NPs to perform a long single NP. Likewise we group the inter-related words based on punctuations and verb groups together.

Example-2: "The PON1 102V allele appears to be associated with an increased risk for pros-tate cancer." (PMID:12783936)

Next after relating each of them into verbs, verb groups then we merge NP and its associated based on pattern matching and extraction method. In this process, we use terms like NP merge or VG merge which mainly emphasize on relationship between

them and form a connection together i.e , we do the process of pattern matching. In this section , we describe the pattern used for matching the information related to text, first we prepare a pattern as a sequence of components such as noun phrase (NP), verb phase(VG) or word group. NP represent the pattern group (head, lexical item) or NP type, here in our work NP is related to two type of information namely NP<disease> and NP<mutation>, similarly VG can be represented in two ways NP{head: lexical item}, NP{contains: lexical item} or NP<type>. NP{head: lexical item} indicates NPs whose head word is the same as the specified lexical item or one of its textual variants. NP{contains: lexical item} describes NPs that contain .

For the matching of NP<type>, recall that a NP<type> can be of two types: NP<mutation> or NP<disease>. To match the former, the NP must contain a reference to some mutation. This can be either in the form of a specific mutation (e.g., R1699W) or a more generic description, i.e., the head word of one of its constituent NPs must indicate a mutation (e.g., mutation, polymorphism, variant, SNP etc.). To match NP<disease> we require the noun phrase to contain a mention of a disease. As an example, consider the following pattern:

NP f head: associated of NP <mutation> with NP <disease>

Example-3 refers to a sentence which corresponds to the above pattern. Note that although the entire text is a merged NP, in order to match the pattern, we need to break the merged NP into its constituent NPs. The first base NP matches NP{head: associate} since its head word is "Association". The second constituent NP, "the BRCA1 missense variant R1699W", matchesNP<mutation> since it contains a specific mutation. The third constituent NP, "a malignant phyllodes tumor of the breast", matches NP<disease> because it includes a mention of adisease. Note that even after we split the longer merged NP into three NPs, NP<disease> is still a merged NP.

Example-3: "Association of the BRCA1 missense variant R1699W with a malignant phyl-lodes tumor of the breast." (PMID:17574969)

Module B

This module consists of extraction of mutation, mutation to gene association and disease to mutation associated. Each of these components are portable, means they are used in the process of gene identification dynamically.

Extraction of Mutations

It is a process of extracting the mutation content based on regular express pattern detection, we extract three components from mutation extraction.The expression used to symbol ate the 3 letters of amino acids as [A,C,G,T] for DNA

Examples of the mutations that are detected using such regular expressions are listed below.

1. Protein level mutations: "Ala282Val", "Asp 327—>Asn", "T877A", "Phe153——Ala" etc.
2. DNA level mutations: "A3537G", "4304G> A", "1066-6T> G", "-79C/T" etc.

In addition, regular expressions are also used to capture mutations that correspond to insertion, deletion and SNP IDs. Examples of these kinds are listed below.

1. Insertions: "5382insC", "IVS9-5insT" etc.
2. Deletions:"9631delC", "6886delGAAAA", "IVS19+2delT" etc.
3. SNP IDs: "rs1800795", "ss984046046" etc.

In some cases, conjunctions are part of mutation mentions. For example, in PMID:9466928, "Ala16>Cys, Thr, Met, Arg, His and Tyr" is mentioned. We detect the conjunctions in this case and generate six mutations: Ala-16-Cys, Ala-16-Thr, Ala-16-Met, Ala-16-Arg, Ala-16-His and Ala-16-Tyr.

We also include extraction of some patterns that are beyond the scope of the above regular expressions. These correspond to mutations that are mentioned in regular text rather than special formats used for mutations as recommended by the Human Genome Variation Society (HGVS) [37]. These extractions are triggered by detection of a pair of amino acids or nucleotides [A,C,G,T]. These are considered as wild-type and mutant-type symbols if an associated mutant position is found. If the mutant position is not mentioned in the same phrase as the wild and mutant-type symbols, then it is usually attached to the phrase with a prepositional phrase (See examples below). We search for specific words, such as codon, position, residue etc. to locate the mutant position. Some examples of a range of mutations extracted using this technique are listed below.

1. "A–> C transversion in codon 135"
2. "T to C transition at positions 409 and 412"
3. "Ser—>Leu change at amino acid 217"
4. "termination at codon 3110"
5. "guanine-adenine point mutation at nucleotide 2185"

We employ a normalization technique to normalize the mutations into one standard format by matching the wild-type, mutant-type and position. We use "WildType-pos-MutantType" as the standard format for normalization.For example, G5557A and 5557G>A (PMID:22200742) normalize to the same mutation G5557A.

Often, informative sentences refer to a mutation-disease association in terms of alleles or genotypes rather than mutations. For this reason, we detect these mentions and match them to the corresponding mutation that might be mentioned elsewhere in the abstract. In Example-4, the association with gastric cancer is referred using the allele 194Trp, whose corresponding mutation is Arg194Trp found in the abstract.

Example-4: "XRCC1 194Trp allele significantly increased the risk of gastric cancer and also associated with risk of gastric cardia carcinoma and promoted distant metastasis of gastric cancer." (PMID:20863780)

Mutation to gene association

Once the mutations are detected, we associate them with their relevant genes. In many cases, this is straightforward as the mutations and the corresponding genes are mentioned close to each other. For example, when both the mutation and the gene appear in the same merged NP, we associate them with high confidence. Example-6 presents one such case where "C-2123G", "G-1969A", and "T715P" are associated with SELP, and "Met62Ile" is associated with PSGL-1.Please note that we do not detect the generic references to genetic variations as mutation men-tions, such as the phrase "VNTR variants" in this case.

Example-6: "Our aim was to evaluate the contribution to CHD of the following SNPs: C-2123G, G-1969A and T715P in SELP, Met62Ile and the VNTR variants in PSGL-1 gene in a North African population from Tunisia." (PMID:20376705)

Even in situations when a particular mutation occurrence does not have an accompanying gene in the same sentence, we have noticed that, often, the gene is mentioned in the same NP or same merged NP with the mutation at least once in some other sentence in the abstract. We propagate the gene detected in these latter cases to all occurrences of the mutation in the rest of the abstract.

There are cases, however, of mutations that do not appear together with their corresponding genes in any sentence of the abstract. If a gene is mentioned anywhere in the whole abstract and it is the only gene mentioned, then we associate it with the mutation. However, if multiple genes are mentioned in the abstract, we look for ones that occur together with a mutation-spe-cific term, such as "variant", "mutant", "variation", "mutation", "polymorphism", "alteration" or "SNP" in the same merged NP. We call this occurrence of the gene and the mutation specific term a gene-mutation pairing. For any detected mutation, we associate it to the gene mentioned in the closest gene-mutation pairing that occurred previously in text, either in the same sentence or any sentence before. Once a mutation has been associated with a gene, the association is propagated to every occurrence of the mutation in that abstract. In Example-7a, the gene ELAC2/HPC2 is detected as having a gene-mutation pairing because of the phrase "mutations of the ELAC2/HPC2 gene". The immediately following sentence, which is shown in Example-7b, has a mutation Glu622Val that does no co-appear with a gene. Applying our rule, Glu622-Val is associated with ELAC2/HPC2.

Mutation to disease association

Once the mutations are extracted and paired with genes, the next step is to find the association of mutations with diseases to complete the extraction of the mutation, gene and disease triplet.

An association between a mutation and a disease can be conveyed in different ways in a sentence, either explicitly or implicitly. Based on our preliminary studies, we have observed that there are six types of sentence structures that are commonly used to specify such associations. For each of these cases, we first describe the type of sentence structure and the patterns used to identify them. Next, we describe how the mutations and the diseases, which will be associated, are identified. Since the technique of extraction for the mutation and the disease may be specific to the sentence structure types, we will discuss them after describing each sentence structure type.

(i) Association sentence type. There are several lexico-syntactic structures that are used to denote an association between two entities. We call these structures as Association Sentence Type. Based on our preliminary study, we identified a few common ways that are employed to describe associations between a mutation and a disease and capture them by defining a set of lexico-syntactic patterns. Matching a sentence against these patterns allows us to identify sentences that are Association Sentence Type. We associate each pattern with a trigger word, where the trigger word appears in the lexical instantiation within the pattern. For example the following four patterns are defined for the lexical trigger "associate".

NP1 VG passivefhead:associateg with NP2

NP1 VG activefhead:associateg with NP2

NPfhead:associateg of NP1 with=and NP2

NPfhead:associateg between NP1 and NP2

As noted earlier, Example-2 matches the first of these rules, where NP1 matches the base NP "the PON1 102V allele" which will allow us to identify the mutation and NP2 matches the merged phrase "an increased risk for prostate cancer" which is used to identify the disease.

In addition to "associate", we use several other trig-

gers words "contribute", "correlate", "relationship", and "effect" and their associated patterns. Many of these words do not necessarily indicate associations but could indicate closely related concepts such as causality. Also, many of these words impose strong subcategorization requirements on prepositions that appear in their arguments. Instead of using multi-word triggers, the prepositions are mentioned in the patterns themselves. For example, one such pattern is NP{head: effect} of NP1 in/ on NP2 which is matched by Example-8.

Example-8: "Synergistic effect of stromelysin-1 (matrix metalloproteinase-3) promoter (-1171 5A->6A) polymorphism in oral submucous fibrosis and head and neck lesions." (PMID:20630073)

Module C

Our work is to extract available information and store in database, when stored in a database these can be easily searched and can be used for manipulation. In module c section, we describe details of database and its information.

Creation of Database

DML is an extraction tool used to identify and extract mutations, diseases and gene (<M,D,G> from sentences of abstract with additional information update and search for accurate results. The information of abstract is extracted from survey literature; the document is linked with the database entities for extraction and linkage of information.Since database is an electronic process, we create database for extraction of information in various ways using scenarios. These scenarios are as

a.Provided a gene, we have to search for a mutation possibility and the gene associated with and its various diseases
b.Provided a mutation, we have to find the disease associated with and its possibilities.
c.Then provided a disease, we have to identify possible pair of mutation gene and its disease associated and effected
d.Provided a mutation, gene pair , we have to find the disease associated and also conduct practical approach on the mutation studies
e.Provided a disease, mutation pair, we have to find disease associated with the existing articles and perform experimental study on given gene.
f.Provided disease, mutation pair , we have to find in how many way disease is effected with mutation and is overall symptoms and outcomes
g. Provided mutation, gene pair, identify the disease associated with available article based on review for analysis.

The scenarios used I to iii should perform query on database, where as iv to vii these are complex , it requires additional information like literature paper of abstract for identification of <M,D,G> , it is a theoretical approach of extraction. v it identifies the manner of which associated disease is caused using mutation, vi it is used to provide analysis on review study vii) it is used to identify the pair of identity across related gene and its mutation causing diseases.. From all the above, result gained are listed using sorting or filtering operation based on user needs.

Table 1. Illustrates the sections of Abstract of PubMEd PMID

Section	Example sentence
Title	Gene detection of cancer patient
Background	Types of cancer and its subtypes
Aim	Extraction of mutation regions
Results	Identified rare variant in brain of patients related to tumor
Conclusion	We have identify gene causing cancer other than breast cancer

Table 1, gives the information of the abstract and its format for extraction of information for the pubmed datasets.

Theoretical extraction of datasets: information can be extracted theoretically using text mining based on the information available from abstract or review papers related to gene and disease causing gene mutation.In this approach we sentence the abstract into 5 category types namely title, background or introduction, aim or methods, conclusion and results.We use different type of approaches in classifying the paper into various sections. Each of the section information is ranked based on order as specified above.

Table 2. Executive Information of Datasets used Evaluation metrics

Name of dataset	Used for tasks of	Used for evaluation of	Size
Biomutac	<M,D,G>&<M,G>	DiMex	62 abstracts , 119
PCA	<M,D,G>&<M,G>	DiMex& EMU	97 abstracts , 170
BCA	<M,D,G>&<M,G>	DiMex& EMU	132 abstracts , 216
MF	M	DiMex	508 abstracts , 910
Variome	M	DiMex	10 full articles 118
Tm var	M	DiMex	166 abstracts 464

We counted true positives (TP), false positives (FP), and false negatives (FN), and used the standard information retrieval metrics of Precision (P), Recall (R), and F-measure (F) for performance evaluation, where P = TP/(TP+FP), R = TP/(TP+FN) and F = 2PR/(P+R).

Table 3 Performance Measure of mutation disease and Mutation – Gene

Dataset	DML Performance in <M,D,G>			DML performance in <M,G> extraction		
	P	R	F	P	R	F
BioMutate	0.89	0.89	0.87	0.91	0.94	0.92

Table 3, show the performance measure of information extracted based on metric evaluation from table 6 and 7.

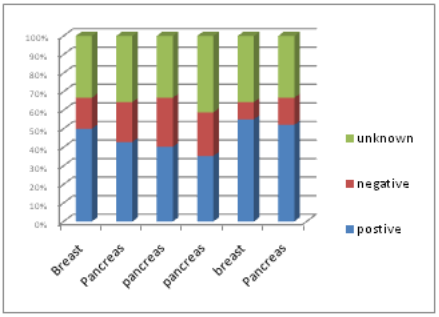

Figure 2. Results of Data sets extracts and their identity of Mutation Disease

Table 4 Measure of Performance associated with Mutation disease and Related Gene Mutation

Dataset	DML Performance in <M,D,G>			DML performance in <M,G> extraction		
	P	R	F	P	R	F
Pancreas Ca	0.95	0.89	0.9	0.76	0.76	0.77
Breast Can	0.94	0.86	0.88	0.65	0.72	0.68

Table 4, show the performance measure of data sets extracts from figure 2 of datasets shown from table 7.

Table 5 Measure of Performance associated with Mutation –Gene and comparison

Data set	DML Performance in <M,G>			EM performance in <M,G> extraction		
	P	R	F	P	R	F
Pancreas Cancer	0.97	0.92	0.93	0.78	0.74	0.78
Breast Cancer	0.95	0.93	0.96	0.66	0.74	0.69

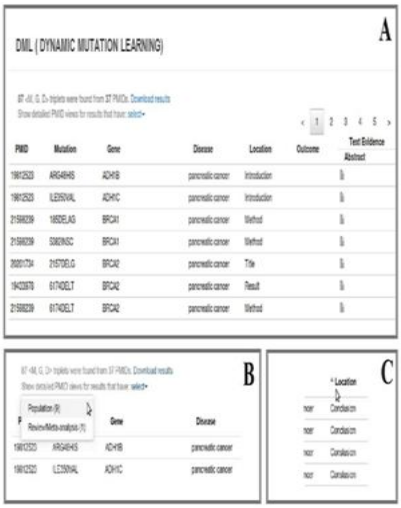

Figure3.Query of Data base for Learning Pancreatic Cancer and Breast Cancer.

The above figure 2 gives the extraction of information from various datasets using DMLex tool of the mutation-disease association.

Table 6. Abstracts dataset extracted using the tool

Genes	Total Abstracts	No of Patients	Cancer type	positive	negative	unknown
biomutac	62	9000	Breast	4500	1500	3000
PCA	97	7000	Pancreas	3000	1500	2500
BCA	132	12000	pancreas	4825	3175	4000
MF	508	17000	pancreas	6000	4000	7000
Variome	10	600	breast	385	65	250
Tmvar	166	1200	Pancreas	625	175	400

Table 7. Details Report of Extracted information of Datasets from Abstracts

Genes	Cancer type	Positive		Negative	
		Male	Female	Male	Female
biomutac	Breast	0	4500	0	3000
PCA	Pancreas	1700	1300	1500	1000
BCA	pancreas	3000	1825	1800	1200
MF	pancreas	4800	1200	4000	3000
Variome	breast	0	385	0	250
Tmvar	Pancreas	425	200	245	155

II. CONCLUSIONS

In this paper, we have described the text-mining system DML, which extracts mutations and identifies their association with diseases in Medline abstracts. We have employed NLP techniques to capture the relationship from text. The system achieved state-of-the-art performances for all three tasks, namely mutation detection, mutation-gene and mutation-disease associations. The evaluation results on

three different test sets showed that our system outperforms the EMU system. The two separate modules for mutation detection and mutation-gene association are portable and they can be used in the context of any other problem definition, irrespective

III. REFERENCES

[1]. Zhang J, Chiodini R, Badr A, Zhang G (2011) The impact of next-generation sequencing on genomics. J Genet Genomics 38: 95–109. doi: 10.1016/j.jgg.2011.02.003 PMID: 21477781

[2]. Capriotti E, Nehrt NL, Kann MG, Bromberg Y (2012) Bioinformatics for personal genome interpretation. Brief Bioinform 13: 495–512. PMID: 22247263

[3]. Burger JD, Doughty E, Khare R, Wei C-H, Mishra R, Aberdeen J, et al. (2014) Hybrid curation of gene-mutation relations combining automated extraction and crowdsourcing. Database 2014. Available: doi: 10.1093/database/bau094

[4]. The UniProt Consortium (2014) Activities at the Universal Protein Resource (UniProt). Nucleic Acids Res 42:D191–D198. doi: 10.1093/nar/gkt1140 PMID: 24253303

[5]. Forbes SA, Bindal N, Bamford S, Cole C, Kok CY, Beare D, et al. (2011) COSMIC: mining completecancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic Acids Res 39: D945–50.doi: 10.1093/nar/gkq929 PMID: 20952405

[6]. Wu T-J, Shamsaddini A, Pan Y, Smith K, Crichton DJ, Simonyan V, et al. (2014) A framework for orga-nizing cancer-related variations from existing databases, publications and NGS data using a High-performance Integrated Virtual Environment (HIVE). Database 2014: bau022. doi: 10.1093/database/ bau022 PMID: 24667251

[7]. Amberger J, Bocchini CA, Scott AF, Hamosh A (2009) McKusick's Online Mendelian Inheritance in Man (OMIM). Nucleic Acids Res 37: D793–6. doi: 10.1093/nar/gkn665 PMID: 18842627

[8]. Stenson PD, Mort M, Ball EV, Howells K, Phillips AD, Thomas NS, et al. (2009) The Human Gene Muta-tion Database: 2008 update. Genome Med 1: 13. doi: 10.1186/gm13 PMID: 19348700

[9]. Béroud C, Hamroun D, Collod-Béroud G, Boileau C, Soussi T, Claustres M (2005) UMD (Universal Mutation Database): 2005 update. Hum Mutat 26: 184–191. PMID: 16086365

[10]. Thorisson GA, Lancaster O, Free RC, Hastings RK, Sarmah P, Dash D, et al. (2009) HGVbaseG2P: central genetic association database. Nucleic Acids Res 37: D797–802. doi: 10.1093/nar/gkn748PMID: 18948288

[11]. Singh A, Olowoyeye A, Baenziger PH, Dantzer J, Kann MG, Radivojac P, et al. (2008) MutDB: updateon development of tools for the biochemical analysis of genetic variation. Nucleic Acids Res 36: D815–9. PMID: 17827212

[12]. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. (2001) dbSNP: the NCBIdatabase of genetic variation. Nucleic Acids Res 29: 308–311. PMID: 11125122

[13]. Whirl-Carrillo M, McDonagh EM, Hebert JM, Gong L, Sangkuhl K, Thorn CF, et al. (2012) Pharmacoge-nomics knowledge for personalized medicine. ClinPharmacolTher 92: 414–417. doi: 10.1038/clpt. 2012.96 PMID: 22992668

[14]. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, et al. (2014) ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res 42: D980–5. doi: 10.1093/nar/gkt1113 PMID: 24234437

[15]. Plazzer JP, Sijmons RH, Woods MO, Peltomäki P, Thompson B, DenDunnen JT, et al. (2013) The InSiGHT database: utilizing 100 years of insights into Lynch syndrome. Fam Cancer 12: 175–180. doi: 10.1007/s10689-013-9616-0 PMID: 23443670

[16]. Doughty E, Kertesz-Farkas A, Bodenreider O, Thompson G, Adadey A, Peterson T, et al. (2011) Toward an automatic method for extracting cancer- and other disease-related point mutations from the biomedical literature. Bioinformatics 27: 408–415. doi: 10.1093/bioinformatics/btq667 PMID: 21138947

[17]. Rebholz-Schuhmann D, Marcel S, Albert S, Tolle R, Casari G, Kirsch H (2004) Automatic extraction of mutations from Medline and cross-validation with OMIM. Nucleic Acids Res 32: 135–142. PMID: 14704350

[18]. Horn F, Lau AL, Cohen FE (2004) Automated extraction of mutation data from the literature: applicationof MuteXt to G protein-coupled receptors and nuclear hormone receptors. Bioinformatics 20: 557–568.

