Visualization Technique for Mutation Functional Analysis

S Pacheco\textsuperscript{a} and H Al-Mubaid\textsuperscript{b}

Computer Science Department, University of Houston, Clear Lake, Houston, Texas 77058, USA

\textsuperscript{a}PachecoS7542@uhcl.edu; \textsuperscript{b}hisham@uhcl.edu

Abstract. A great body of research have been devoted in the past two decades to understanding gene functions, gene mutations, and mutation-disease relationships. In this paper, we present a method for exploring and understanding mutation functions using visualization and graphics representation. Functional annotation of human gene mutations is an important step in mutation pathogenicity prediction and understanding diseases progress and mechanisms. The presented visualization method is based on identifying the most significant and most specific sets of functions for a given set of mutations under a target disease or medical condition. The presented visualization allows for easy and effective understanding of the details and differences among the various sets of functions of the given mutations under certain disease or medical condition.

1. Introduction
A mutation is basically a genetic variation and a change in the DNA sequence [1]-[5]. If the DNA change is common in the population then it is called polymorphism [6]-[8]. The single nucleotide polymorphism (SNP) is a change in only one nucleotide in the DNA sequence and is considered the most common type of genetic variation [6]. Any change in the DNA sequence may affect one or more genes and can be harmful or neutral. If the mutation affects the gene functionality in the negative way then it can cause diseases, conditions or genetic disorders [6]-[7], [9]. Diseases and medical conditions resulting from changes in the DNA sequence are collectively called genetic disorders [6]-[7]. The common genetic diseases are cystic fibrosis, sickle cell anaemia, and hemochromatosis [6]. In mutation research, the interests are in predicting mutation functions [2], [9]-[11], predicting pathogenicity of mutations [2],[4], prediction of mutation-disease associations [9], and discovering new gene mutations [4],[12]. In this paper, we present a technique for analysing disease mutations using graphics representation and visualization. The proposed technique relies mainly on visualizations of mutation functions for the analysis of mutation-disease relationships. Therefore, this paper contributes into mutation functional analysis, visualization, and mutation pathogenicity prediction. We utilize the functional gene annotations (GOA) from the Gene Ontology (GO) to identify specific and significant functions related to mutations. In other words, this work explores and examines the relationship between mutations and functional annotations in the context of a specific disease or medical condition as shown in Figure 1 and Figure 2. Then, we identify the most significant functions for the mutations from the Biological Process (bp) aspect of the GO. The experiments are conducted with the aim to identify the most significant bp functional annotations from GO for gene mutations. The discovered functions for a given set of mutations are then grouped into several significance levels for visualization and to explore these functions with graphics representation based on their significance levels.
2. Background and Related Work
The deep and complete understanding of human diseases, disease mutations, and medical conditions is still in need of research and investigations to identify and explain the consequences of variations and mutations in the genetic sequences and their functions [13]. There is also a demand to explore and analyse mutations from various aspects, and more importantly, regarding their functions [13]. The work in this paper contributes into this direction of exploring mutation and mutation functions with visualization for better understanding of diseases and medical conditions. In their recent research, López-Urrutia et al. (2018) reported that many mutations are still in need to be analysed for their functions and it is a ‘very complex task [13]’ The area of human-gene-mutation research flourished after completing human genome and date back to more than two decades ago [1]-[3], [7], [14]. The work in this area of human-gene-mutation research can be categorized as follows. (1) extracting mutation-gene-disease associations from medical texts and literature [5], [12]. (2) prediction and classification of mutation pathogenicity [2], [4]. (3) analysing mutations in the context of specific diseases, e.g. Alzheimer, and Breast Cancer [4], [9], [13], [15]. (4) mutation-disease association extraction from biomedical literature, mainly with NLP and machine learning methods [12], [14], [16]. (5) prediction and annotation of mutation functions and other analyses of genetic variations [2],[9]-[11], [17]. Recent research concluded that most mutations tend to have more than just a single impact or single biological sequence as it has commonly believed [15]. Calabrese et al. (2009) developed tools, namely SNPs&GO, to predict and analyse mutations using functional GO annotations [10]. Moreover, in the past two decades, great body of research and projects were devoted for mutation-disease associations analysis and extraction of mutation-gene-disease triplets from the biomedical literature [1]-[3], [15], [18].

Among the most common public archives of mutation is ClinVar [2] database which is freely available archive of human gene variations and phenotypes containing ~200,000 submitted interpretations covering more than 125,000 variants [2]. The interpretations in ClinVar include large number of genes (~more than 26,000 genes) and including structural variants that may include many genes; for variants that affect a single gene, almost 4800 genes are represented in ClinVar [2]. Another one of most comprehensive resources is COSMIC, the Catalogue Of Somatic Mutations In Cancer [9]. In fact, COSMIC is the largest resource in the world for somatic mutations in human cancer [9].

![Figure 1](image_url)

**Figure 1.** In this diagram there is a mutation $m_1$, two genes $g_1, g_2$, and five GO-bp function terms $f_1, \ldots, f_5$. Mutation $m_1$ is associated with two genes $g_1, g_2$. Gene $g_1$ is annotated with three terms $f_1, f_2, \text{ and } f_3$; while $f_4$ is associated with both $g_1$ and $g_2$. 
3. Methods and Techniques

Mutation pathogenicity prediction is an important area of study in mutation research [14]. In this area, besides pathogenicity prediction, we are also interested in identifying and discovering mutation functions or what is called mutation functional annotations [14]. The work in this paper is in this direction. Specifically, we are interested in examining the functional annotations of mutations and visualizing the mutations as overlapping functional sets from the biological process (bp) aspect of GO. The GO is the main resource of functional annotation for all genomics data including genes, RNA, and proteins; see Figure 3 [11], [17], [19]. We would like to explore and analyze the mutations and investigate mutation functions using graphic representation and visualization based on the significance levels of the functions. The proposed method works on a given set of mutations for some target disease or medical condition. Given a set \( M \) of \( n \) mutations \( M = \{ m_1, m_2, ..., m_n \} \) with a common disease/condition, e.g., Alzheimer disease (AD). We extract the functional bp terms from the GO for functionally annotating these \( n \) mutations at various levels of significance as follows. We retrieved all the genes associated with each mutation \( m_i \in M \). Next, we use the pb annotations from the gene ontology annotation database (GOA_human [19]) for each gene \( g_{ij} \) associated with each mutation \( m_i \).

Figure 3 shows a part of the GO from the bp aspect.

Next, we would like to visually illustrate the functionality of this group \( M \) of mutations with various significance levels. The significance level is a number between 0 and 1, where 1 is very significant and 0 indicates no significance. The significance level will be associated with one or more functional terms in respect to the given mutation set \( M \) which eventually will be an indicator for the disease/condition we study. Typically, a mutation may be associated with one or more genes as shown in Figures 1 and 2. In Figure 1, the mutation \( m_1 \) is associated with two genes \( g_1, g_2 \), and gene \( g_1 \) is annotated with three terms \( f_1, f_2, \) and \( f_4 \). The term \( f_4 \) is associated with both \( g_1 \) and \( g_2 \). Figure 2 shows three mutations \( m_1 \ldots m_3 \), two genes \( g_1, g_2 \), and five GO annotation terms. Since the bp aspect of the GO is a tree-like directed acyclic graph (DAG), each function is represented as a node in this tree. The proposed visualization method is based on identifying the function nodes in the ontology that represent least common subsumer (LCS) for most function nodes associated with the mutations. We are interested in identifying the most significant LCS’s as these will be the main cause of the disease or condition associated with the mutations under investigation. We calculate the significance level of the LCS’s as follows:

\[
\text{Significance level of node } n: \quad s_n = \frac{\# \text{ of subsumed terms}}{\text{total } \# \text{ of function terms}}
\]  

From this equation, the root node has significance level of 1.0 as it subsumes all the nodes in the tree. For example, in Figure 4, there are 7 functions \( \{ f_1, f_2, ..., f_7 \} \) each one of these functions \( f_i \) is a bp annotation term from the gene ontology. These function terms \( f_i \) are divided into four significance levels \( \{1.0, 0.80, 0.60, 0.40\} \) according to equation (1) and illustrated with color codes where darker means more significant. For example, the function \( f_3 \) is relatively important among the target mutations under
4. Experiments and Results

One of the goals of this research is to find the most significant LCS nodes in the bp aspect of the GO, and gene function(s), related to a target disease or condition. We obtained mutations with their genes and associated conditions from the ClinVar database as shown in Table 1 [1]-[3]. We mainly focused on the pathogenic mutations (82321 mutations) involving more than 9200 genes; see Table 1. For the annotation data we used the Gene Ontology (GO) which is widely used as the main source of functional annotations of genetic data [11], [17], [19]; we used OMIM for gene-disease association information [20]; and for human gene functional annotations we relied on the GOA database (GOA_human) which includes ~ 600K annotations [19].

![Figure 3](image)

**Figure 3.** This diagram shows a part of the GO annotation terms from the biological process (bp) aspect (source: QuickGO EMBL)
Figure 4. Visualization of the functional annotations with four levels of significance.

Figure 5. General visualization and using graphics to explain the functionality of mutations.
**Table 1.** In *Clinvar* we studied 82321 mutations that are pathogenic with 9270 genes

| Type       | # of Mutations | # of Genes | # of Associated Conditions |
|------------|----------------|------------|----------------------------|
| pathogenic | 82321          | 9270       | 7473                       |
| benign     | 72070          | 9044       | 2572                       |

**Table 2.** The top genes associated with Alzheimer disease (*src: Morbid map OMIM*).

| Gene symbol | Gene Id | MIM Id  |
|-------------|---------|---------|
| HFE         | 3077    | 613609  |
| NOS3        | 4846    | 163729  |
| PLAU        | 5328    | 191840  |
| A2M         | 4353    | 103950  |
| MPO         | 2       | 606989  |
| APP         | 351     | 104760  |

*Mutation functional visualization for one disease:* we conducted analysis for the mutations and genes of *Alzheimer disease* (AD). There are 6 genes associated with *AD* in *OMIM* and in *morbid map* [21] as shown in Table 2. We extracted all the functional *bp* annotations of these 6 genes (total of 96 *bp* annotations) from *GOA_human* [19]. We found that the *bp* function {GO:0031325; positive regulation of cellular metabolic process} to be the most significant LCS for all these 96 *bp* annotations with significance level = 0.64 as shown in the visualization in Figure 6.

The conducted experiments examine mutation groups for functional visualization (see for example Figures 1 - 6) at various levels. Figure 7 provides an illustration of a simple case with 8 functions (nodes) from the *bp* aspect. As shown in part (a) the yellow nodes are the annotated functions for visualization. The visualization is shown in part (b) of the figure. These eight nodes are divided into three significance levels {1.0, 0.33, and 0.66}. The visualization technique is an effective way to investigate and explore the functionality of which will lead to clear understanding of the pathogenicity of mutations.

![Figure 6](image-url)
5. Discussion and Conclusions

Visualization is one of the best means to deliver data and results, and to explore the differences of your outcomes [5],[22]-[24]. The interest in delivering visualization in Data Science has been growing and increasing in the past decade [24]. Inspecting and examining results with visual illustrations and graphics can give better ideas and comprehensive outlook. We presented a visualization methodology for identifying the most important functions that mutations, specifically disease mutations, manifest themselves with. The proposed approach is based on the functional gene annotations from the publicly available databases like GO, GOA, OMIM, and ClinVar. From the experiments and results, the visualization method provides an easy and effective means to understand the functions of mutations and which functions are more significant. It also gives a clear understanding of the differences between the functions of mutations. This leads to better understanding mutation pathogenicity and disease mechanisms.

Acknowledgement

The work in this paper was supported by the University of Houston – Clear Lake under the Pathway to STEM Careers (PSC) program. We express our thanks to PSC program for the support.

References

[1] Stephens Z, Wang C, Iyer R K and Kocher J. Detection and visualization of complex structural variants from long reads. BMC Bioinformatics 2018, vol. 19 (Suppl 20) :508 https://doi.org/10.1186/s12859-018-2539-x, 2018.
[2] Landrum M J, Lee J M, Benson M, Brown G, Chao C, Chitipiralla S, et. al. (2016). ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Research, 44 (Database issue), D862–D868. http://doi.org/10.1093/nar/gkv1222
[3] Stenson P D, Ball E V, Mort M, Phillips A D, Shiel J A, Thomas N S T, Abeyinghe S, Krawczak M, and Cooper D N. Human Gene Mutation Database (HGMD): 2003 update. Human Mutation. No. 21:577–581; 2003.
[4] Bailey M H, Tokheim C, Porta-Pardo E et al. Comprehensive Characterization of Cancer Driver Genes and Mutations. Cell 173, 371–385. April 5, 2018.
[5] Krawczak M and Cooper N D. 1997. The Human Gene Mutation Database. Trends Genet 13:121–122; 1997.
[6] Genetic Home Reference GHR, US National Library of Medicine, NIH. Retrieved September 2018; https://ghr.nlm.nih.gov/
Genetic Alliance. The New York-Mid-Atlantic Consortium for Genetic and Newborn Screening Services. Understanding Genetics: A New York, Mid-Atlantic Guide for Patients and Health Professionals. Washington (DC): Genetic Alliance; 2009. Available from: https://www.ncbi.nlm.nih.gov/books/NBK115568/

GOTermFinder, URL: https://go.princeton.edu/cgi-bin/GOTermFinder.

Kordopati V, Salhi A, Razali R, Radovanovic A, Tifratene F, Uludag M, Bajic V B (2018). DES-Mutation: System for Exploring Links of Mutations and Diseases. Scientific Reports, 8, 13359. http://doi.org/10.1038/s41598-018-31439-w

Calabrese R, Capriotti E, Fariselli P, Martelli PL, Casadio R. Functional annotations improve the predictive score of human disease-related mutations in proteins. Human Mutation 2009; 30(8):1237-44. doi: 10.1002/humu.21047.

Capriotti E, Martelli P L, Fariselli P, and Casadio R (2017). Blind prediction of deleterious amino acid variations with SNPs&GO. Human Mutation, 38(9), 1064–1071. http://doi.org/10.1002/humu.23179

Opak K and Mulder N (2017). Recent advances in predicting gene–disease associations. F1000Research Journal, 6, 578. http://doi.org/10.12688/f1000research.10788.1

López-Urrutia E, Salazar-Rojas V, Brito-Elias L et al. BRCA mutations: is everything said? Breast Cancer Res Treat (2018). https://doi.org/10.1007/s10549-018-4986-5

Wang M and Wei L (2016). iFish: predicting the pathogenicity of human nonsynonymous variants using gene-specific/family-specific attributes and classifiers. Scientific Reports; 6:31321. DOI: 10.1038/srep31321 1

Butkiewicz M, Blue E, Leung Y, Jian X, Marcera E et.al. Functional annotation of genomic variants in studies of late-onset Alzheimer’s disease, Bioinformatics, Volume 34, Issue 16, 15 Aug. 2018, pp. 2724-2731. https://doi.org/10.1093/bioinformatics/bty177

Doughty T, Kertesz-Farkas A, Bodenreider O, Thompson G, Adadey A, Peterson T, and Kann MG (2011). Toward an automatic method for extracting cancer- and other disease-related point mutations from the biomedical literature. Bioinformatics, 27(3), 408–415. http://doi.org/10.1093/bioinformatics/btq667

Araujo F A, Barh D, Silva A, Guimaraes L and Juca Ramos R T. GO FEAT: a rapid web-based functional annotation tool for genomic and transcriptomic data. Scientific Reports. Vol. 8: 1794 (2018).

Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010; 38:e164.

Gene Ontology Annotation (GOA) Database: https://www.ebi.ac.uk/GOA

Online Mendelian Inheritance in Man, OMIM. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD), {2018}. URL: https://www.omim.org

Zhong L X, Kun S, Jing Q, Jing C, and Denise Y. Non-Syndromic Hearing Loss and High-Throughput Strateges to Decipher its Genetic Heterogeneity (2013). Journal of Otology, vol. 8 no.1.

Kreft L, Turan D, Hulstaert‡§ N, Botzki† A, Martens L, and Vandermarliere E. Scop3D: Online Visualization of Mutation Rates on Protein Structure. Journal of Proteome Research, 2019, 18 (2), pp 765–769. DOI: 10.1021/acs.jproteome.8b00681

Ng P K, Li J, Jeong K J, Shao S et al. Systematic Functional Annotation of Somatic Mutations in Cancer. Cancer Cell. 2018, 33(3):450-462.e10. doi: 10.1016/j.ccell.2018.01.021.

Ward M O, Grinstein G and Keim D. Interactive Data Visualization Foundations, Techniques, and Applications, Second Edition. Taylor and Francis, 2015. doi: doi.org/10.1201/b18379