International Journal of Bio-Science and Bio-Technology
Vol.7, No.6 (2015), pp.287-294
http://dx.doi.org/10.14257/ijbsbt.2015.7.6.29

Using Bioinformatics Tools for Identifying Disease-Causal Genetic Variants from Human Genomes

Youngmahn Han and Insung Ahn

Korea Institute of Science and Technology Information
245 Daehak-ro, Yuseong-gu, Daejeon 305-806, KOREA
{hans, isahn}@kisti.re.kr

Abstract

Identification of the disease-causal genomic variants that alter human phenotypes, particularly those that lead to diseases, is the central goal of human genetics studies. In the past decade, genome-wide studies have identified several hundreds of common variants associated with complex human diseases and traits. Despite these successes, most of the common variants only have a small individual contribution to the estimated heritability underlying common diseases and traits. Many explanations for these missing heritabilities have been suggested, including rare variants, structural variants, regulatory variants, and epigenetic variants. Recent advances in high-throughput technologies have provided an opportunity to construct comprehensive maps of genetic variation, including the several million single nucleotide variants, thousands of small insertion or deletion events, and thousands of structural variants, in both the protein-coding and noncoding regions of the human genome without time and cost limitations. The present review describes current bioinformatics tools for identifying deleterious variants in protein-coding regions based on the evolutionary and functional constraints of human proteins.

Keywords: *computational prediction, deleterious variants, human genome, prediction, bioinformatics, next-generation sequencing, protein evolution*

1. Introduction

Nonsynonymous single nucleotide polymorphisms (nsSNPs) are coding variants that cause amino acid changes in their corresponding proteins. An nsSNP alters the protein structure and function, resulting in drastic phenotypic consequences. Because most of the alterations in coding regions are deleterious, they are eventually eliminated through purifying selection. By contrast, beneficial mutations can sweep through the population and become fixed, thus evolutionarily contributing to species differentiation. Deleterious nsSNPs are associated with both Mendelian diseases and common complex diseases. Nonsynonymous substitutions account for approximately half of the genetic changes known to cause disease in databases such as Online Mendelian Inheritance in Man (OMIM) and Human Gene Mutation Database (HGMD) [1, 2]. Prediction methods for deleterious nonsynonymous substitutions have been successfully developed to distinguish between nonsynonymous changes that cause simple Mendelian diseases from neutral changes. Although most of the simple Mendelian diseases remain rare due to purifying selection, some have become relatively common in certain populations because of overdominant selection, which occurs when the heterozygote carrier has higher fitness than both the mutant and normal homozygotes. For example, the E6V substitution in β-globin is common in malaria-endemic populations, because heterozygous carriers are more resistant to malaria than normal homozygotes, whereas individuals that are homozygous for the rare allele have sickle-cell anemia. Another well-known example of

ISSN: 2233-7849 IJBSBT

over dominance involves methylenetetrahydrofolate reductase (MTHFR) alleles. Variants that reduce MTHFR activity can cause mental retardation and cardiovascular disease in carriers. Nevertheless, these protein function-damaging variants have become relatively common in human populations, because reduced MTHFR activity is considered to have been beneficial to an individual's overall fitness during recent human evolution. Because these overdominant nonsynonymous substitutions can severely affect protein function, they can readily be detected using some prediction methods [3].

Owing to emerging high-throughput sequencing techniques, such as next-generation sequencing, it is now feasible to detect large amounts of nsSNPs without time and cost limitations. Two general models, the common disease-common variant model and the common disease-rare variant model, have been proposed to explain the nature of the genetic variations underlying common complex diseases such as hypertension, diabetes, heart disease, and cancer [4-8]. However, for both rare variants with a large effect and common variants with a weak effect on the common phenotype, the use of prior knowledge could be crucial for the discovery of disease-causing genes [9-12]. The deleteriousness of nsSNPs can be used as prior knowledge to filter out the millions of benign variants. Several constraint-based methods for prediction of the deleteriousness of nsSNPs have been proposed (Table 1); these approaches generally presume that deleterious variants break the evolutionary and structural constraints for governing native protein functions.

In the present review, we provide a survey of the evolutionary and structural constraint-based approaches developed to date for detecting the deleteriousness of human genetic variants. We focus on the particular considerations and difficulties of these methods, and highlight their role and potential in improving our ability of readily detecting disease-causing variants.

2. Predicting Deleterious nsSNPs Based on Evolutionary Constraints

There are two major assumptions required for predicting deleterious nsSNPs from a phylogeny. First, nonsynonymous SNPs that destroy the stability and biochemical functions of their corresponding proteins, and thus cause medically detrimental phenotypes, are subject to purifying selection due to the reduction in evolutionary fitness. Second, a deleterious nsSNP in the current population is also assumed to be deleterious in homologous genes of different species. Homologous genes are considered to be orthologous if they separated at the time of species divergence; thus, the copies of the same gene in the two resulting species are said to be orthologous. Subsequently, the two orthologous protein-coding genes fulfilling the same biological function in the two diverging species start to accumulate mutations independently. Despite their independent evolutionary trajectories, the accumulation of mutations usually follows a similar pattern in the two diverging species, owing to similar structural and functional constraints of the proteins. For example, myoglobins of different species must fold into similar three-dimensional structures to fulfill their similar function. Therefore, the relative probability of mutations that are deleterious can be expected to be identical for the two orthologous genes. In the case of orthologous genes with well-established functions, the great majority of nonsynonymous substitutions will be deleterious. Under these assumptions, the functional effect of a nonsynonymous substitution can be predicted from the pattern of amino acids observed in a multiple sequence alignment of orthologous protein sequences. The statistical probability of a nonsynonymous substitution causing a genetic disease in the overall disease increases monotonically with an increase in the degree of evolutionary conservation of the mutation site [13, 14]. Besides the nonsynonymous substitutions at evolutionarily conserved sites of a multiple sequence alignment, the three following considerations are essential for predicting deleterious nsSNPs:

- Sequence conservation is not a predictor of deleteriousness; instead, conservation in excess of neutral expectations is used to infer constraint.
- The phylogenetic scope [15-17]
- Many protein sequence-based methods also exploit biochemical data, including amino acid properties (such as charge, solvent accessibility), sequence information (such as the presence of a binding site), and secondary structural information. The integration of these data with comparative sequence analysis can significantly improve predictions of deleteriousness [18-21].

Table 1. Computational Methods for Prediction of Deleterious Nonsynonymous Variants

Name	Description	URL
Align-GVGD	Phylogenetic method using phyisco-chemical amino acid properties	http://agvgd.iarc.fr/agvgd_input.php
LRT	Phylogenetic method using estimated evolutionary rate	No server
MAPP	Phylogenetic method using patterns of physico-chemical properties of amino acid substitutions	http://mendel.stanford.edu/sidowlab/downloads/MAPP/index.html
PMut	Phylogenetic and structural features combined with machine learning	http://mmb2.pcb.ub.es:8080/PMut/
PolyPhen-2	Phylogenetic and structural features combined with machine learning	http://genetics.bwh.harvard.edu/pph2
SIFT	Phylogenetic method using patterns of amino acid substitutions	http://sift.jcvi.org/
SNAP	Phylogenetic and structural features combined with machine learning	http://cubic.bioc.columbia.edu/services/SNAP/
MutationTaster	Phylogenetic and biochemical/structural features combined with machine learning	http://www.mutationtaster.org/

With these considerations, the typical method for predicting deleterious nsSNPs is as follows [22] (Figure 1). The first step is to choose appropriate homologous sequences and conduct a multiple sequence alignment. The choice of sequences is critical because very shallow alignments are uninformative, whereas deep alignments may include very distant sequences that will cause misleading predictions. Therefore, the most straightforward method of constructing an alignment would be to include only orthologous sequences; however, most existing methods also include paralogs [15]. This may be justified because the majority of damaging mutations affect the stability of the protein structure, which is

expected to be highly similar among paralogs. Limiting the analysis to orthologs would frequently result in shallow alignments given current methods of prediction. However, this may be resolved as a result of many new sequencing projects, and the development of new methods that could allow for the choice to limit the analysis to orthologs among closely related species only if sufficiently diverse and informative alignments can be generated [15-17]. The second step is to evaluate how well an allelic variant fits the amino acid pattern observed in the phylogeny. Existing methods for this purpose use positional conservation measures, probabilistic scoring functions, or both. MAPP [23] and Align-GVGD [24, 25] use a different approach based on the conservation of the physico-chemical properties of amino acids. Phylogenetic relationships among sequences are taken into account using sequence weights (SIFT, PMut, MAPP) [26, 27, 23], a pre-computed species tree (LRT) [28], or other heuristic algorithms such as PSIC (PolyPhen-2 and SNAP) [29, 30].

3. Structural and Functional Constraints in Deleterious nsSNPs

Most of the phylogeny-based methods for the prediction of nsSNPs also exploit the structural and functional constraints in protein evolution arising from the stability of the folding state, retaining essential conformational flexibility that mediates the protein's functions in the cell, and the need to avoid opportunistic interactions and the accumulation of amyloid fibrils formed from misfolded proteins. Missense mutations result in the loss of stability of the folding state and the native functions of the protein, and thus may cause genetic diseases. Hydrophobic interactions in the highly evolutionarily conserved solvent-inaccessible regions are crucial for maintaining the overall structural stability of a protein. Thus, introducing a charged residue into the protein's interior generally affects the entropy of the system due to the resulting change in solvent accessibility. This destabilizes the protein, resulting in a misfolded protein structure [31, 32]. Although the hydrophobic interactions in the buried non-polar regions are a major constraint in protein evolution, it is important to consider the average effect of both solvent accessibility and the type of secondary structure formed by hydrogen-bond interactions. The buried and hydrogen-bonded polar side may be relatively more conserved in the course of protein evolution than the buried and non-polar side that does not form any hydrogen bonds. This is because a buried polar residue that is satisfied in terms of side-chain hydrogen bonding stabilizes the protein structure, and hydrogen bonding further helps to increase the packing density in the protein's interior [33]. Therefore, the structural constraints on protein evolution could be evaluated based on the integrity of the effects arising from hydrophobic interactions, hydrogen bonding interactions, and other factors (such as electrostatic interactions) [23, 31, 32, 34].

The various functional constraints in protein evolution mainly result from interactions with other molecules such as substrates, ligands, nucleic acids, and other proteins; these are often components of interaction networks that are conserved throughout evolution [31,34]. Several masking models that exclude functional residues from multiple sequence alignment have been developed using various combinations of functional residues, and were compared with a non-masking model including functional residues in the calculation of substitution probabilities. The average probability of amino acid conservation for the non-masking model was reported to be ~1.36% higher than that of the masking model, although the difference was less distinct when the enzymes' active sites were omitted from masking [35]. Overall, this demonstrates that functional residues are under greater pressure to be conserved throughout evolution when they are crucial to the activity of proteins and to the interaction with partner molecules [31].

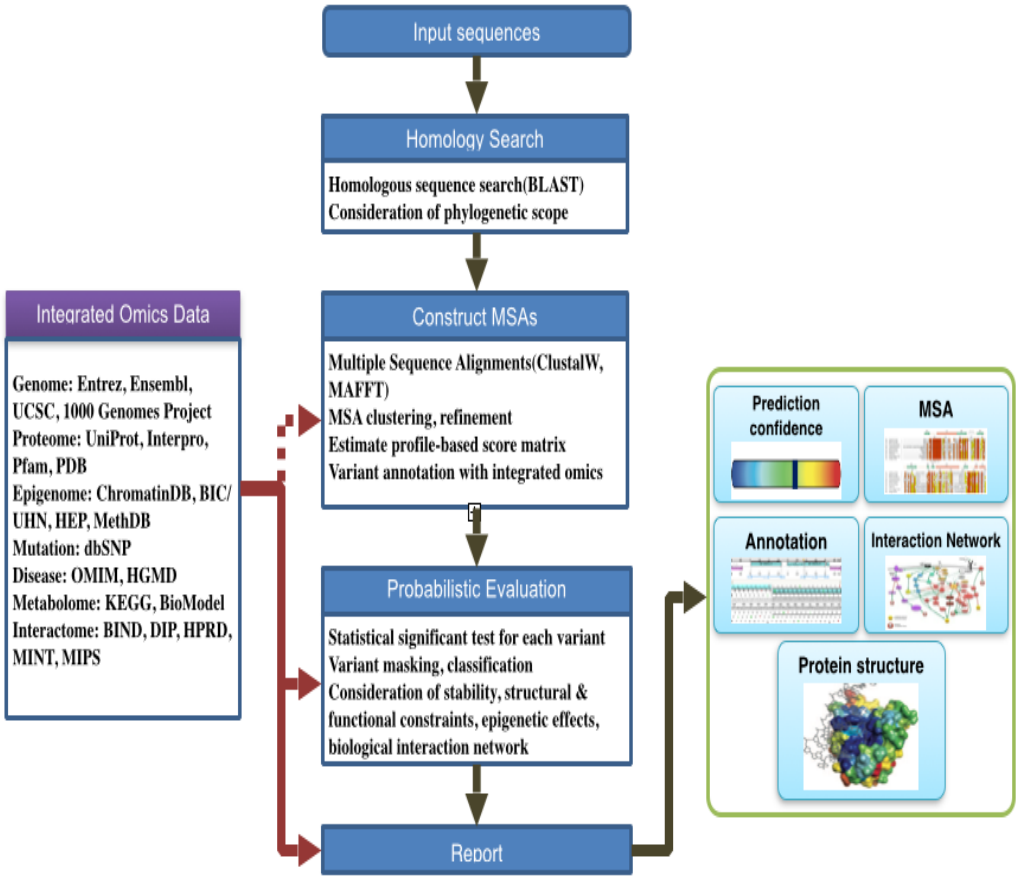

Figure 1. Typical Workflow for Prediction of Deleterious Non-synonymous Variants

4. Realizing the Identification of Disease-Causal Variants by Using Bioinformatics Tools

There is a need for integrated workflow combining the heterogeneous output data of bioinformatics tools in realizing the identification of disease-causal variants from the given input genome. Several bioinformatically-capable workflow management systems (WMSs) have been developed for allowing clinicians or researchers to construct complex workflows orchestrating heterogeneous output data of bioinformatics tools and to automate the execution of the workflows [36-39]. There are mainly required considerations to effectively implement a bioinformatically-capable WMS as follows:

- There are many bioinformatics tools and databases literally developed by geographically distributed organizations, research institutes, or related industries across the world. Some make their tools web accessible; some provide command line based standalone programs or software libraries. Standardization and extensible integration of distributed tools is necessary for providing seamless access to them.
- Bioinformatics tools are highly heterogeneous in their input/output data types. These heterogeneity leads to be difficult to make links among tool tasks according to data flow. A WMS should provide flexible integration methods to resolve data type heterogeneity.
- Workflow scalability is important to help in large-scale data analysis like NGS data analysis, protein interaction network analysis, and docking simulation through high performance computing resources, e.g., running a large number of parallel jobs on a

cluster computer. However, most research groups seem to be impossible to maintain such computing resources due to the high cost of computer hardware and the lack of professional human resources to manage and utilize them.

- Reproducibility of scientific analyses and processes is at the core of the scientific method, in that it enables researchers to evaluate the validity of each other‟s hypothesis and to repeat techniques and analysis methods to obtain scientifically similar results. In order to support reproducibility, WMS should capture and generate provenance information as a critical part of the workflow-generated data. Provenance information can be referred as a historical metadata that provides explanations on how a particular intermediate result data has been generated from the given input data

In order to meet these requirements, a bioinformatically-capable WMS called Bioworks has been developed as presented in our previous work [40, 41]. Bioworks is implemented in Java, and based on client/server system architecture as shown in Figure 2). The Bioworks client program provides the user-friendly graphical user interface (GUI) which enables users to easily compose workflows for complex bioinformatics analysis. Workflows are executed at the server side on high-performance cluster computing resources. Users can monitor the status of workflow execution through the client program anywhere, anytime. Especially, by adopting Java Web Start technology, it can be automatically installed and upgraded via web. Figure 2 shows the implementation of identifying disease-causal variants from the given input genome data by using Bioworks.

5. Conclusions

Nonsynonymous mutations in protein-coding regions result in drastic phenotypic consequences by altering the structure and function of proteins, leading to both Mendelian and common complex diseases. Several computational methods have been developed to estimate the functional effect of human nsSNPs, as shown in Table 1. These methods presume that most deleterious nsSNPs break down the evolutionary and structural constraints that govern the functions of native proteins. The typical workflow for these methods consists mainly of constructing a multiple sequence alignment from appropriate homologous sequences and estimating how strongly the nsSNPs break the evolutionary and structural constraints that mediate protein functions. There are essential considerations in the workflow, including sequence conservation as an evolutionary constraint, the phylogenetic scope, and the integration of structural and physico-chemical information with comparative sequence analysis. The structural and functional constraints in protein evolution can result from the integrative effects of hydrophobic interactions, hydrogen-bonding interactions, and interactions with other molecules such as substrates, ligands, nucleic acids, and other proteins.

In this review, we focused on computational methods for predicting deleterious variants in only protein-coding regions. However, the development of high-throughput DNA sequencing technologies now makes it feasible to identify whole maps of genetic variants in both the protein-coding and noncoding regions of individual human genomes. Indeed, current available tools allow for the identification of deleterious variants in both protein-coding and noncoding regions (for example, MutationTaster [42], VAAST [43], and CADD [44]). Bioinformatically-capable WMSs, such as Bioworks enable clinicians or researchers to realizing the identification of disease-causal variants from the given input genome data by allowing to construct workflows orchestrating heterogeneous output data of bioinformatics tools and to automate the execution of the workflows. We anticipate that this bioinformatics research area at the interface of molecular evolution, structural biology, and human genetics will increase in importance in the forthcoming personal genomics era.

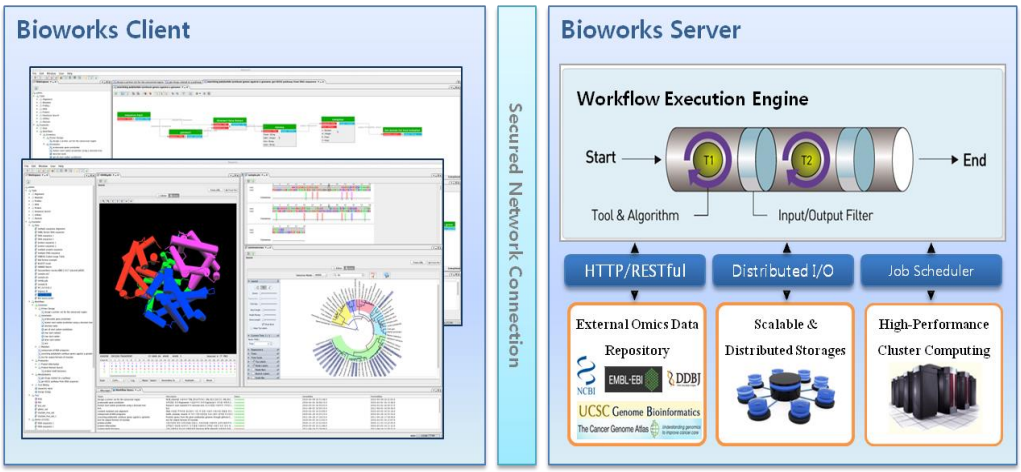

Figure 2. Client/Server System Architecture of Bioworks

Acknowledgment

The author would like to acknowledge the Biomedical Prediction Technology Team: Insung Ahn, Thai Quang Tung, and Jinhwa Jang. This work is supported by the Korea Institute of Science and Technology Information

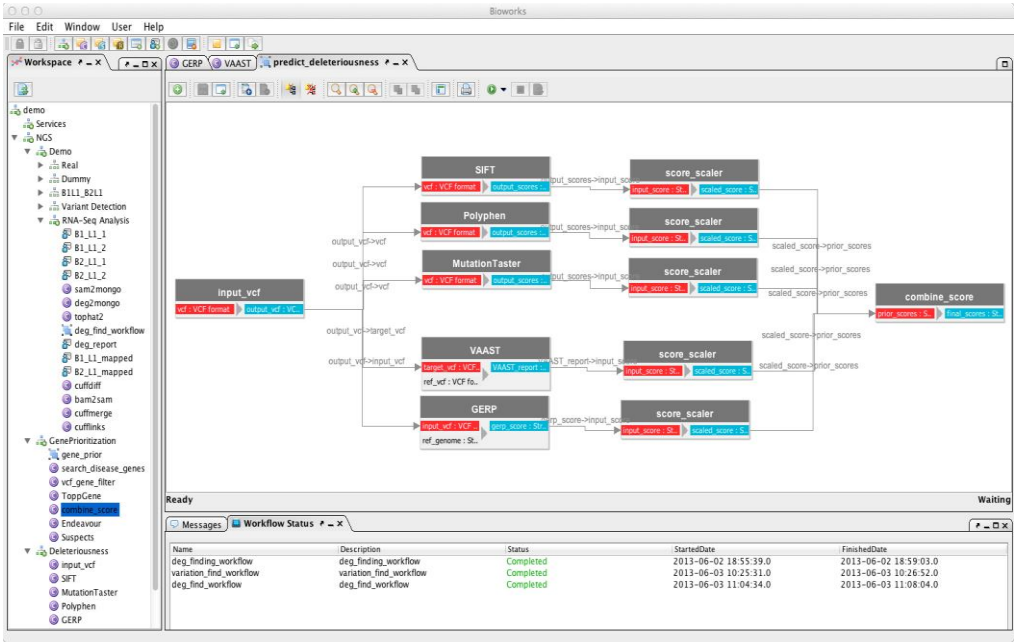

Figure 3. Implementation of Identifying Disease-causal Variants by using Bioworks

References

[1] A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini and V. A. McKusick, Nucleic Acids Res., vol. 33, no. 514, **(2005)**.
[2] P. D. Stenson, E. V. Ball, M. Mort, A. D. Phillips, J. A. Shiel, N. S. Thomas, S. Abeysinghe, M. Krawczak and D. N. Cooper, Hum. Mutat., vol. 21, no. 577, **(2003)**.
[3] P. C. Ng and S. Henikoff, Genome Res., vol. 12, no. 436, **(2001)**.
[4] J. K. Pritchard, Am. J. Hum. Genet, vol. 69, no. 124, **(2001)**.
[5] C. P. Ponting and L. Goodstadt, Eur. J. Hum. Genet, vol. 13, no. 269, **(2005)**.
[6] P. D. Thomas and A. Kejariwal, Proc. Natl. Acad. Sci. USA, vol. 101, no. 15398, **(2004)**.

[7] J. C. Cohen, R. S. Kiss, A. Pertsemlidis, Y. L. Marcel, R. McPherson and H. H. Hobbs, Science, vol. 305, no. 869, **(2004)**.
[8] R. Smigrodzki, J. Parks and W. D. Parker, Neurobiol. Aging, vol. 25, no. 1273, **(2004)**.
[9] G. M. Cooper, D. L. Goode, S. B. Ng, A. Sidow, M. J. Bamshad, J. Shendure and D. A. Nickerson, Nat. Methods, vol. 7, no. 250, **(2010)**.
[10] G. M. Cooper and Shendure, J., Nat. Rev. Genet., vol. 12, no. 628, **(2011)**.
[11] K. Musunuru, A. Strong, M. Frank-Kamenetsky, N. E. Lee, T. Ahfeldt, K. V. Sachs, X. Li, H. Li, N. Kuperwasser, V. M. Ruda, J. P. Pirruccello, B. Muchmore, L. Prokunina-Olsson, J. L. Hall, E. E. Schadt, C. R. Morales, S. Lund-Katz, M. C. Phillips, J. Wong, W. Cantley, T. Racie, K. G. Ejebe, M. Orho-Melander, O. Melander, V. Koteliansky, K. Fitzgerald, R. M. Krauss, C. A. Cowan, S. Kathiresan and D. J. Rader, Nature, vol. 466, no. 714, **(2010)**.
[12] L. D. Ward and M. Kellis, Nat. Biotechnol., vol. 30, no. 1095, **(2012)**.
[13] P. C. Ng and S. Henikoff, Genome Res., vol. 11, no. 863, **(2001)**.
[14] E. V. Koonin, Annu. Rev. Genet., vol. 39, no. 309, **(2005)**.
[15] G. M. Cooper and C. D. Brown, Genome Res., vol. 18, no. 201, **(2008)**.
[16] G. M. Cooper and A. Sidow, Curr. Opin. Genet. Dev., vol. 13, no. 604, **(2003)**.
[17] E. A. Stone, G. M. Cooper and A. Sidow, Annu. Rev. Genomics Hum. Genet., vol. 6, **(2005)**, pp. 143–164.
[18] R. J. Dobson1*, P. B. Munroe1, M. J. Caulfield1 and M. A. S. Saqi, BMC Bioinformatics, vol. 7, no. 217, **(2006)**.
[19] P. Yue, Z. Li and Moult, J., J. Mol. Biol., vol. 353, no. 459, **(2005)**.
[20] L. Bao and Y. Cui, Bioinformatics, vol. 21, no. 2185, **(2005)**.
[21] Y. Li, Z. Wen, J. Xiao, H. Yin, L. Yu, L. Yang and M. Li, BMC Bioinformatics, vol. 12, no. 14, **(2011)**.
[22] D. M. Jordan, V. E. Ramensky and S. R. Sunyaev, Curr Opin Struct Biol., vol. 20, no. 3, **(2010)**, pp. 342.
[23] E. A. Stone and A. Sidow, Genome Res., vol. 15, no. 978, **(2005)**.
[24] S. V. Tavtigian, A. M. Deffenbaugh, L. Yin1, T. Judkins, T. Scholl, P. B. Samollow, D. de Silva, A. Zharkikh and A. Thomas, J. Med. Genet., vol. 43, no. 295, **(2006)**.
[25] E. Mathe, M. Olivier, S. Kato, C. Ishioka, P. Hainaut and S. V. Tavtigian, Nucleic Acids Res., vol. 34, no. 1317, **(2006)**.
[26] P. C. Ng and S. Henikoff, Nucleic Acids Res., vol. 31, no. 13, **(2003)**, pp. 3812.
[27] C. Ferrer-Costa, J. L. Gelpí, L. Zamakola, I. Parraga, X. de la Cruz and M. Orozco, Bioinformatics, vol. 21, no. 3176, **(2005)**.
[28] S. Chun and J. C. Fay, Genome Research, vol. 19, no. 1553, **(2009)**.
[29] I. A. Adzhubei, S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork, A. S. Kondrashov and S. R. Sunyaev, Nat Methods, vol. 7, no. 248, **(2010)**.
[30] Y. Bromberg and B. Rost, Nucleic Acids Res., vol. 35, no. 3823, **(2007)**.
[31] C. L. Worth, S. Gong and T. L. Blundell, Nat. Rev. Mol. Cell Biol., vol. 10, no. 709, **(2009)**.
[32] J. Thusberg and M. Vihinen, Human Mut., vol. 30, no. 703, **(2009)**.
[33] D. Schell, J. Tsai, J. M. Scholtz and C. N. Pace, Proteins, vol. 63, no. 278, **(2006)**.
[34] C. Pál, B. Papp and M. J. Lercher, Nature Reviews Genetics, vol. 7, no. 337, **(2006)**.
[35] S. Gong and T. L. Blundell, PLoS Comput. Biol., vol. 4, no. e1000179, **(2008)**.
[36] A. Tiwari and A. K. T. Sekhar, Computational Biology and Chemistry, vol. 31, no. 305, **(2007)**.
[37] T. Oinn, M. Addis, J. Ferris, D. Marvin, M. Senger, M. Greenwood, T. Carver, K. Glover, M. R. Pocock, A. Wipat and P. Li, Bioinformatics, vol. 20, **(2004)**, pp. 3045–54.
[38] J. Goecks, A. Nekrutenko, J. Taylor1 and The Galaxy Team, Genome Biol, vol. 11, no. 8, **(2010)**, pp. R86.
[39] B. Liu, R. K. Madduri, B. Sotomayor, K. Chard, L. Lacinski, U. J. Dave, J. Li, C. Liu and I. T. Foster, Journal of Biomedical Informatics, vol. 49, no. 119, **(2014)**.
[40] Y. Han, Inter. J. Bio-Science and Bio-Technology, vol. 3, no. 59, **(2011)**.
[41] Y. Han, T. Q. Tung and I. Ahn, "Current Research Trend of Bioscience and Welfare III", Proceedings of International Workshop on Bioscience and Medical Research, **(2015)**, August 19-22, Jeju Island, Korea.
[42] J. M. Schwarz, C. Rödelsperger, M. Schuelke and D. Seelow, Nature Methods, vol. 7, no. 575, **(2010)**.
[43] M. Yandell, C. Huff, H. Hu, M. Singleton1, B. Moore, J. Xing, L. B. Jorde1 and M. G. Reese, Genome Res., vol. 23, **(2011)**, pp. 1529-1542.
[44] M. Kircher, D. M. Witten, P. Jain, B. J. O'Roak, G. M. Cooper and J. Shendure, Nat. Genetics, vol. 46, no. 310, **(2014)**.

