IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

STUDY ON NOVEL APPROACHES IN INVESTIGATING HIGH PERFORMANCE BIOINFORMATICS TOOL FOR DETECTING GENETIC MUTATIONS INVOLVING SNPS: A DETAILED REVIEW

[1]Upal D Joshi

[1]Student, [1]B.Tech IV[th] year, Department of Genetic Engineering.

[1]Bharath Institute of Higher Education and Research, Chennai.India

Abstract: This approach being regarded a vital protocol employed for determination of mutation, as there are certain genomic variations determined between individuals. Besides SNPs existed all through the genome which can be subdivided into several groups. Among the various types of SNPs, a non-synonymous SNP in the coding region of a gene is crucial because it modifies the amino acid composition. Consequently, such alterations can have an impact on protein structure, function and cell division. Although, the pin pointing the effects of the many non-synonymous SNPs using biochemical analysis is challenging, computational analysis tools predicting their impact on protein activity and stability have recently developed. Several research articles have stated is effectiveness in identifying the deleterious and disease associated mutations, thus predicting the pathogenic non-synonymous in connection to their functional and structural damaging properties.

Index Terms - SNPs, Mutations, Computational analysis tool, Bioinformatics

I. INTRODUCTION

Significance of Bioinformatics tools and their impact on SNPs detection

In recent times, the complexity concerning with molecular wet lab techniques has become substantially reduced and also being replaced or streamlined with the aid of major computational biology/ bioinformatics tools (Lesk, 2019). The very foundation about bioinformatics based on various factors likely the computational techniques, artificial intelligence, algorithms, database management and software engineering etc. These factors in turn leads to development of community of biological/ molecular data resources and from which, the very notion of their applications and its development of the bioinformatics for analysis of genetic data takes place (Binitha and Sathiya, 2012).

In simpler terms, the definition of Bioinformatics could be addressed as a very broad field, which encompasses primarily over the issues such as mapping, sequence comparison, sequencing, gene identification, network databases, protein modelling, visualization tools and ethics. It is an interdisciplinary subject that on one hand requires biological information- infrastructure building and on the other requires computation based biological research. All this depends on the large stores of experimental and derived data. The prime reason for selection of computational approaches for determination of SNP with regards to a single gene. In this study, we employ computational tools on the basis of reliability and standardised approach than that of labour-intensive molecular studies of interest (Wasserman et al., 2004). The following process, provides a generalized protocols for screening mutations involving SNPs as follows. With regards to determination of SNPs involving the essential requirement of computational and Bioinformatics tool become an utmost necessary factor. The significance and usage of Insilco tools for SNP could be understood and could be attributed chronologically as follows:

Discovery of SNPs:

The key prospect and area of investigation by CGAP-GAI serves as the discovery for determining novel SNP transcripts in the first place. The search for SNPs from publicly available form of mRNA and from EST sequences which remains to be the basic unit for this analysis (Campus and Hinxton, 2003). This is achieved through UniGene cluster. However ESTs remain to be of high-throughput low-quality sequences with false positives that arose as resultant sequencing artifacts posing to be one serious concern. For validation of new candidates of SNPs that are identified from *in silico* could be determined using MALDI-TOF mass spectrography.

SNP annotation

The prediction method involved in SNP could be determined and are rooted from UniGene clusters. With time, the Unigenes are further then refined. the new sequences in particular could be added alongside to a cluster (Pontius et al., 2003). In certain cases the cluster tend to merge with another UnigGene or may be split into sets of clusters. In such cases gene names, clusters and descriptions could change over the due course of time.

Analysis for Coding Region with regards to SNPs

Updating mapping for SNPs for genes as well as UniGene annotation or using annotation tools, determining the SNPs location with respect with mRNAs, motifs and 3-D protein structure. SNPs which led to substitutions of amino acid that are of particular interest since they express obvious potential for altering protein stability and activity. The initial stage involving NCBI RefSeqs curated from "Virtual as structures that are derived from public length transcripts that are high-quality sequences obtained via IMAGE consortium clones.

Display Tools

Following the adjunct to our SNP discovery efforts which provide Java-based viewer that display SNP under context of sequence assemblies from which they were predicted. SNP Launcher applet provides two views from assembly/ sequence alignment. Also from BLAST client and software for designing PCR primers for determining SNP assays could be accessed via tools menu. The assembly overview showed graphical representation of sequence assembly with clone coverage along with length of the consensus sequence. This approach being regarded vital and was employed for SNP determination, as there are certain genomic variations determined between individuals.

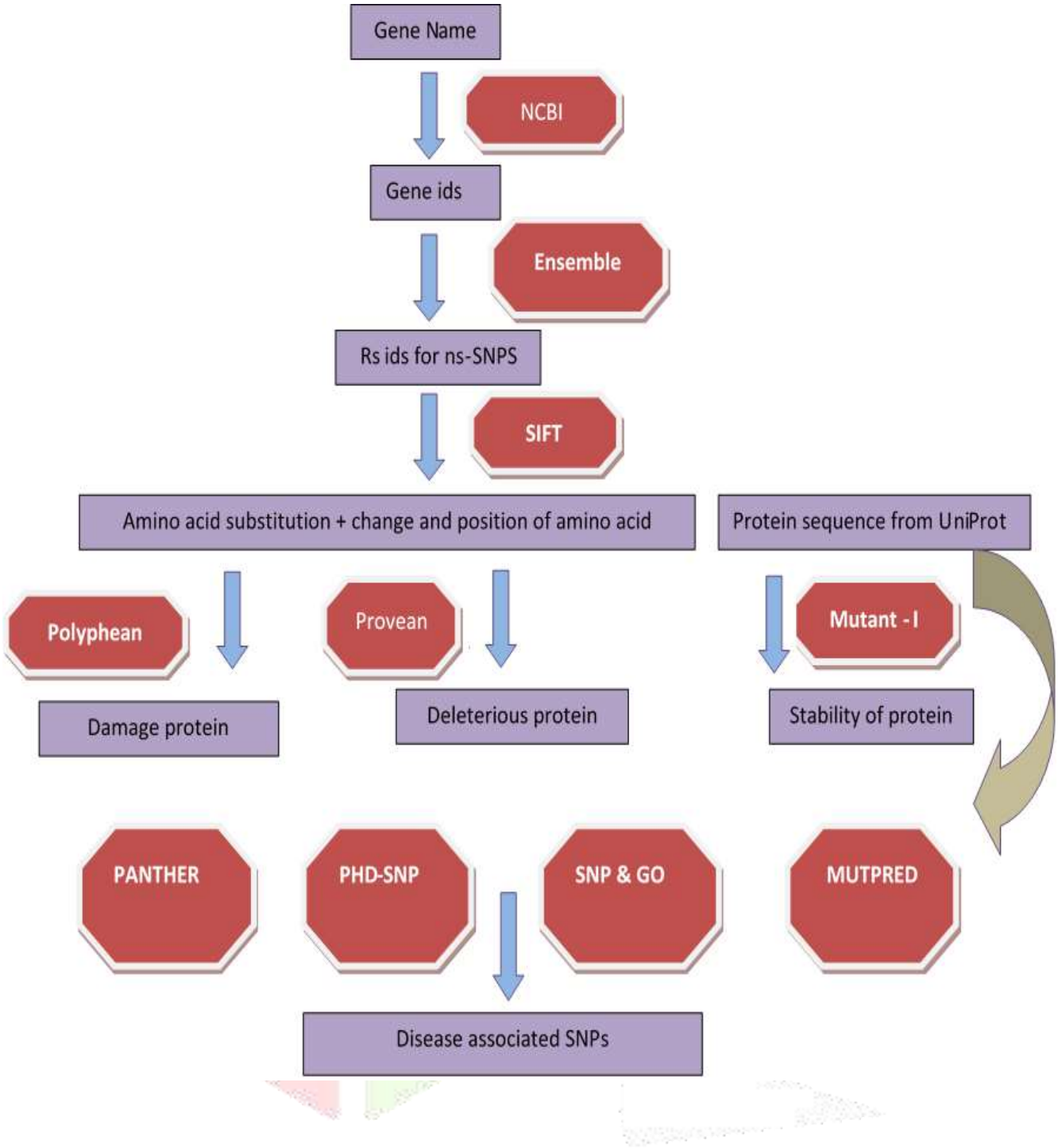

Figure 1. Flowchart illustration of heirarchial Insilico tools for SNPs detection

Besides SNPs existed all through the genome which can be subdivided into several groups. The research motive in employing insilico as a prominent technique is due to the method's regulatory purposes are well-defined, and promotes a transparent framework for researchers in choosing, handling and applying valid in silico methods. Besides the above-mentioned attributes, insilico technique is cost-effective and are not laborious in its application.

Public Accessibility over SNP Discovery Tools

The set of SNP discovery tools appeared freely available for research community. Researchers are given their liberty over submitting the own sets of DNA sequences for analyzed for presence of SNPs. The below section provides a detailed overview regarding a list of computational tools that are utilized for the present study for determination of SNP expression with emphasis on studying MTHFR gene and their association with RPL (Savage et al., 2005).

List of Computational Tools Employed in present study

Bioinformatic tools in the case of SNPs prediction offers a wide array of different softwares that functions unique.

BIOINFORMATICS ANALYSIS OF *MTHFR* GENE VARIATIONS

From the inferred Bioinformatics tool, initial the flow of the work as determined. Screening of Deleterious SNPs was achieved via NCBI, ENSEMBL and UNIPROT.

A. Screening of deleterious SNPs

The most commonly employed computational methods such as SIFT, PolyPhen-2, Provean, I-Mutant, PANTHER, Mut-pred, PMUT, Phd-SNP, SNP & GO were employed for prediction of the impact of nsSNPs on proteins. These insilico tools are indicated in table 1.

Table 1. Bioinformatics Tools involved in screening of deleterious SNPs

Name of the tool	Link	Reference
Screening of Deleterious SNPs		
NCBI	https://www.ncbi.nlm.nih.gov/	Chang et al., (2006)
ENSEMBL	https://asia.ensembl.org/index.html	McLaren et al., (2010)
UNIPROT	https://www.uniprot.org/	Yip et al., (2008)

SIFT

The expansion for SIFT is 'Sorting Intolerant From Tolerant'. This tool is an essential tool for determination of SNP on the basis of determining the amino acid substitution. By identifying on how an amino acid substitution hinders or affects the protein function, indirectly attributes on the underlying changes that are caused due to SNPs in the first place.

General Features

Sorting Intolerant from Tolerant (SIFT) (Ng and Henikoff, 2003) is a computational tool that facilitates primarily in detection of deleterious coding nonsynonymous SNPs. This program functionalizes via presuming the pertinent/ major amino acids, which will remain conserved in protein family and undergoes certain sets of modifications in specific positions that tend to be predicted under deleterious category (Rajasekaran et al., 2007). Considering the mutagenesis studies for human samples, wherein SIFT could attribute via differentiation of functionally neutral and deleterious polymorphisms (Ng and Henikoff, 2000).

Functional Attributes

From the sequence homology based tool SIFT could determine conservation of particular position of any types of amino acids or sequence within the given protein sequence taken into account. SIFT aligns the paralogous as well as orthologous protein sequences for determination as well as with influencing amino acid substitution, thus via considering some key outcomes such as: functional significance and physical properties. SIFT is confirmed to be sufficiently much more precise/ accurate in detection of disease especially with related SNPs via predicting the known disease related SNPs from database which was found to be comprising of only a total of 20% false positive result. Despite, there is a large number of SNP data that are presently available in database lacking structural information that are related to SNP, however SIFT algorithm performs analysis using sequence data, such that the tool predicts significantly large volume of SNPs from the database. As a result SIFT in turn provides advantages with other forms of deleterious SNP prediction algorithms (Ng and Henikoff, 2003). SIFT basically takes rsID from SNPs as its input and a .txt file uploaded comprising rsIDs to SIFT server. SIFT on its part, calculates tolerance index (TI) of particular amino acid substitution. SIFT score can be categorized as tolerant (0.201–1.00) or intolerant (0.051–0.10) and borderline (0.101–0.20). Thus a Single Nucleotide Polymorphisms functional consequence was found to be inversely proportional to the tolerance index (TI).

Polyphene

The expansion of Polyphene is 'Polymorphism Phenotyping'. The tool employs 'developed prediction method'. This computational tool enabled for analysing human non-synonymous SNPs that are present publicly in dbSNP database. From the following database that provides data collection for nsSNPs having a predicted impact over the structural and functional attributes of the protein synthesized (Ferguson et al., 2006).

General Features

SNPs were usually being re-analyzed using Polyphen software. This software predicts the effect of mutations for both structural as well as from functional sides as well.

Functional Attributes

The major functional attribute for this software tool is predicting the possible impact for the substitution of amino acid substitution from both structure and function of human protein via analysing multiple sequence alignment as well as its protein 3D structure, in addition the tool calculates position specific independent count scores (PSIC) in each of its two variants, and further then calculates PSIC scores and their difference between both variants.

With PSIC score difference getting higher, there is a higher chances for functional impact over a particular amino acid substitution that is likely tend to have its impact on the resultant SNP. Prediction outcomes are classified as probably damaging, which is possibly damaging/ benign according to PSIC values since it ranges from (0-1); with values closer to zero is regarded as benign whereas those values closer to 1 is regarded probably damaging and also could be indicated via vertical black marker inside color gradient bar, with green being benign and red as damaging. nsSNPs which are being predicted as ntolerant by Sift usually being submitted to Polyphen as protein sequence in the form of FASTA format obtained from UniproktB/Expasy after prior submission of relevant ensemble protein (ESNP) there, and with entered position of mutation, native amino acid and the newer substituent under both structural as well as from functional predictions (Oberholzer et al., 2012).

I-Mutant

This insilico tool is of neural-network-based web server that attributes for automatic prediction in the overall stability changes among protein as a result of single-site mutations that are observed. This tool was basically trained under datasets derived from ProTherm. The tool has presently been serving as one of the comprehensive form of database in determining protein mutations.

General Features

The insilicotool is of support vector machine (SVM)-based tool for automated prediction of protein stability and their changes upon single point mutations. The underlying predictions are performed starting either from protein structure or, more significantly , from that of protein sequence of interest.

Functional Attributes

The prime functional attribute to be considered is that the tool serves both as a classifier for predicting sign of protein stability change as a result of mutation and also by acting as regression estimator for predicting related Delta DeltaG values. Their Web interface facilitates in selection of predictive mode which is dependent on availability of protein structure as well as their sequence.

This tool acts as a unique and valuable helper for protein design, even when the protein structure is not yet known with atomic resolution. Their predictions are performed starting either from its protein structure or, from the protein sequence. This latter task, has been recently been exploited. The method was trained as well as tested on data set that are rendered from ProTherm, which presently acts as most comprehensive available database in thermodynamically experimental data in determining free energy changes in the protein stability with mutational changes and under varied conditions. In terms of the tool acting as a classifier, I-Mutant correctly predicts (with a crossvalidation procedure) 80% or 77% of data set, depending on usage of structural/ sequence information, respectively (Capriotti et al., 2005).

Panther

The expansion is Protein Analysis Through Evolutionary Relationships. This tool has widely been utilized as an online resource for comprehensive protein evolutionary as well as functional classification, comprising of tools over large-scale biological data analysis. Recent development that are intended to focus primarily under three main areas namely:

- Functional information (‘annotation’) coverage
- Genome coverage, and finally
- Accuracy, as well as improved genomic data analysis tools.

General Features

The goal of this particular tool is classifying proteins on the basis of their function. Any attempts in terms of their classification. Ontologies have been used for some time in computer science for precisely these kinds of applications. The latest version comprised almost 5000 new protein families (for a total of over 12 000 families), each with a reference phylogenetic tree including protein-coding genes from 104 fully sequenced genomes spanning all kingdoms of life.

Functional Attributes

The prime attribute for PANTHER is the Sequence alignments and phylogenetic trees were obtained from their database (Mi *et al.*, 2013). For each family that are residing in PANTHER, the researcher could easily reconstruct ancestral sequences for common ancestors (internal nodes) in tree using PAML, with PANTHER tree as well as their alignment serves as an input.

The input data which is of protein sequence is searched against that of PANTHER sequences using BLASTP (Camacho *et al.*, 2009) for identifying the best-matching PANTHER sequence (tree leaf node). The targeted amino acid could be traced from leaf back via increasingly older ancestral proteins from the tree. Such a trace stops target amino acid which is different from corresponding amino acid from ancestral sequence, or if amino acid reconstruction probability appeared less than predefined threshold, and age (in millions of years) of lastly preserved ancestor has also been reported (Tang and Thomas, 2016).

Phd-snp

The expansion for the insilico tool is Predictor of human deleterious Single Nucleotide Polymorphism. predicting the impact of human SNVs, both in coding and non-coding regions (Capriotti & Fariselli, 2017).

General Features

PhD SNP is available both as web server, and software as well for processing larger datasets with local variants. PhD-SNP, is designed for remaining as simple and lightweight. This SVM s acts as starting point for the prediction of protein sequence if they are of new phenotype that are derived from that of nsSNP which could be of newer genetic disease amongst humans.

Functional Attributes

The tool has greatly advanced as their functional attributes defined in the case of the new version with a predictor based on single SVM that are trained and tested for protein sequence and their profile information as well (Song et al., 2006). The key functional steps are to be regarded when utilizing this tool that are mentioned below:

- In case of the given set of mutation, the substitution in case of wild-type residue for mutant has been as encoded under 20 elements vector having -1 in the position relative to that of wild-type residue, 1 in the case of the position that is relative to mutant residues and with 0 in case of the remaining 18 positions.
- A second 20 elements vector encoding sequence environment for build reporting occurrence of residues using windows comprising 19 residue around mutated residue.
- In case of a given targeted protein, the sequence profile which is built in accordance to the procedure mentioned. From this the researcher could easily determine both frequency of the wild type (Fi(WT)) as well as mutated (Fi(MUT)) residues at position i. NAL is the numeber indicates number of sequences within the alignment given and position and the Conservation Index (CI).

SNAP

The expansion of the tool is Screening for non-acceptable polymorphisms, The insilico tool could potentially be able to classify nsSNPs throughout from all proteins into non-neutral (effect on function) as well as neutral (no effect) using sequence-based computationally acquired information alone (Bromberg et al., 2008).

General Features and Functional Attributs

Under each instances, SNAP provides reliability index, which serves as well-calibrated measure for reflecting the level of confidence over a particular prediction. SNAP serves as neural network-derived tool which in particular functions via accurately predicting the functional effects of nsSNPs from the newly compiled data set via incorporating evolutionary information (residue conservation within sequence families), predicted aspects from the protein structure (secondary structure, solvent accessibility), and also from other relevant information. All information that are required in the form of input was further obtained from sequence alone (Paul and Sundaridoss, 2016).

SNAP aids by a refined and extended previous machine-learning tools over many ways, by their extensive data set usage for assessment, and via particular approach for data handling, and also for its ubiquitous applicability (to sequences from all organisms, proteins with and without known structures, and entirely novel SNPs in scarcely characterized and un-annotated families) (Nailwal et al., 2017).

I tasser

The I-TASSER server acts as an on-line platform which in turn implements I-TASSER based algorithms in structure and functional predictions of protein sequences. Also it allows the researcher in automatic generation of rendering a high-quality model predictions with 3D structure as well as biological function for protein molecules from amino acid sequences.

General Features and Functional Attributes

- The prime features is that the researcher after identifying and isolating a sequence of amino acid then submits. The server tends to initially retrieve the template protein from that of similar folds that are rendered from the PDB library via LOMETS (Expansion: locally installed meta-threading approach).

- In case of the second step, wherein the continuous fragments that are excised from PDB templates are further then reassembled under a full-length models via replica-exchange Monte Carlo simulations from threading unaligned regions (mainly loops) that are built by ab initio modeling. In certain condition wherein there is no appropriate template is identified by LOMETS, I-TASSER will build the whole structures by ab initio modeling. The low free-energy states are identified via clustering simulation decoys.
- Thirdly, the fragment assembly simulation has been performed again starting from SPICKER cluster centroids, where the spatial restrains collected from both the LOMETS templates and the PDB structures via TM-align which is used for guiding simulations. The purpose for second iteration is primarily for removal of. the steric clash as well as to refine the global topology of the cluster centroids.

Pymol

The PyMOL tool was used for locating nsSNPs on protein structure and in analyzing RMS deviation via superimposing both native as well as mutant structures. Amino acids within the position of SNPs were further checked for polar interactions over other amino acids in protein via using PyMOL.

General Features and Functional Attributes

Pymol serves as potent and interactive molecular visualization software tool used for protein, ligand visualization (Itoh N and Ohta, 2013). The 3-D structure of protein and ligand were generally visualized using Pymol and are saved using PDB file format in subjecting them for further analysis.

PYMOL was used for generation of mutant models of SNPs with each of the selected PDB entries from corresponding amino acid substitutions. PYMOL allows the browsing via rotamer library for changing amino acids. A “Mutagenesis Wizard” was used for replacement of native amino acid with that of the new one. The mutation tool facilitates replacement of native amino acid using “best” rotamer for new amino acid. The “.pdb” files were further then saved for all the models.

REFERENCES

[1] Itoh, N., & Ohta, H. (2013). Roles of FGF20 in dopaminergic neurons and Parkinson's disease. *Frontiers in molecular neuroscience*, *6*, 15.

[2] Bromberg, Y., Yachdav, G., & Rost, B. (2008). SNAP predicts effect of mutations on protein function. *Bioinformatics*, *24*(20), 2397-2398.

[3] Nailwal, M., & Chauhan, J. B. (2017). Analysis of consequences of non-synonymous SNPs of USP9Y gene in human using bioinformatics tools. *Meta Gene*, *12*, 13-17.

[4] Capriotti, E., & Fariselli, P. (2017). PhD-SNPg: a webserver and lightweight tool for scoring single nucleotide variants. *Nucleic acids research*, *45*(W1), W247-W252.

[5] Tang, H., & Thomas, P. D. (2016). PANTHER-PSEP: predicting disease-causing genetic variants using position-specific evolutionary preservation. *Bioinformatics*, *32*(14), 2230-2232.

[6] Camacho, C., et al. (2009) BLAST+: architecture and applications, BMC Bioinformatics, 10, 421.

[7] Mi, H., Muruganujan, A., Casagrande, J. T., & Thomas, P. D. (2013). Large-scale gene function analysis with the PANTHER classification system. *Nature protocols*, *8*(8), 1551.

[8] E. Capriotti, R. Calabrese, and R. Casadio, "Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information," *Bioinformatics*, vol. 22, no. 22, pp. 2729–2734, 2006.

[9] Oberholzer, P. A., Kee, D., Dziunycz, P., Sucker, A., Kamsukom, N., Jones, R., ... & Piris, A. (2012). RAS mutations are associated with the development of cutaneous squamous cell tumors in patients treated with RAF inhibitors. *Journal of clinical oncology*, *30*(3), 316.

[10] Ferguson, P. J., Bing, X., Vasef, M. A., Ochoa, L. A., Mahgoub, A., Waldschmidt, T. J., ... & El-Shanti, H. (2006). A missense mutation in pstpip2 is associated with the murine autoinflammatory disorder chronic multifocal osteomyelitis. *Bone*, *38*(1), 41-47.

[11] Ng, P. C., & Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. *Nucleic acids research*, *31*(13), 3812-3814.

[12] Rajasekaran, R., Sudandiradoss, C., Doss, C. G. P., & Sethumadhavan, R. (2007). Identification and in silico analysis of functional SNPs of the BRCA1 gene. *Genomics*, *90*(4), 447-452.

[13] Binitha, S., & Sathya, S. S. (2012). A survey of bio inspired optimization algorithms. *International journal of soft computing and engineering*, *2*(2), 137-151

[14] Campus, W. T. G., & Hinxton, C (2003).. GENOME SEQUENCE AND VARIATION.

[15] Pontius, J. U., Wagner, L., & Schuler, G. D. (2003). 21. UniGene: A unified view of the transcriptome. *The NCBI Handbook. Bethesda, MD: National Library of Medicine (US), NCBI*.

[16] Chang, R. W., Javid, P. J., Oh, J. T., Andreoli, S., Kim, H. B., Fauza, D., & Jaksic, T. (2006). Serial transverse enteroplasty enhances intestinal function in a model of short bowel syndrome. *Annals of surgery*, *243*(2), 223.

[17] McLaren, W., Pritchard, B., Rios, D., Chen, Y., Flicek, P., & Cunningham, F. (2010). Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics*, *26*(16), 2069-2070.

[18] Yip, Y. L., Famiglietti, M., Gos, A., Duek, P. D., David, F. P., Gateau, A., & Bairoch, A. (2008). Annotating single amino acid polymorphisms in the UniProt/Swiss-Prot knowledgebase. *Human mutation*, *29*(3), 361-366.