Gene 680 (2019) 20–33

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Review

A review study: Computational techniques for expecting the impact of non-synonymous single nucleotide variants in human diseases

Marwa S. Hassan[a,b,*], A.A. Shaalan[c], M.I. Dessouky[d], Abdelaziz E. Abdelnaiem[c], Mahmoud ElHefnawi[a,e]

[a] Systems and Information Department and Biomedical Informatics Group, Engineering Research Division, National Research Center, Giza, Egypt
[b] Patent Office of Scientific Research Academy, Egypt
[c] Electronics and Communication Department, Faculty of Engineering, Zagazig University, Zagazig, Egypt
[d] Electronics and Electrical Communications Department, Faculty of Electronic Engineering, Menoufia University, Menouf 32952, Egypt
[e] Center for Informatics, Nile University, Giza, Egypt

ARTICLE INFO

Keywords:
Non-synonymous single nucleotide variants
Genotype
Phenotype
Machine learning techniques (MLTs)
Predictive power
Meta-tool
Pathogenic
Coding and noncoding variants
Protein stability

ABSTRACT

Non-Synonymous Single-Nucleotide Variants (nsSNVs) and mutations can create a diversity effect on proteins as changing genotype and phenotype, which interrupts its stability. The alterations in the protein stability may cause diseases like cancer. Discovering of nsSNVs and mutations can be a useful tool for diagnosing the disease at a beginning stage. Many studies introduced the various predicting singular and consensus tools that based on different Machine Learning Techniques (MLTs) using diverse datasets. Therefore, we introduce the current comprehensive review of the most popular and recent unique tools that predict pathogenic variations and Meta-tool that merge some of them for enhancing their predictive power. Also, we scanned the several types computational techniques in the state-of-the-art and methods for predicting the effect both of coding and noncoding variants. We then displayed, the protein stability predictors. We offer the details of the most common benchmark database for variations including the main predictive features used by the different methods. Finally, we address the most common fundamental criteria for performance assessment of predictive tools.

This review is targeted at bioinformaticians attentive in the characterization of regulatory variants, geneticists, molecular biologists attentive in understanding more about the nature and effective role of such variants from a functional point of views, and clinicians who may hope to learn about variants in human associated with a specific disease and find out what to do next to uncover how they impact on the underlying mechanisms.

1. Introduction

1.1. Background

The modern advances in bioinformatics technologies produce a huge amount of genetic data. Since gene or protein variants are irregular and most variations have no dangerous consequences, recognition of genetic deviations is usually an essential (Zhou et al., 2016). Variations in the protein have influence not only in the protein structure but also its stability and function. Non-synonymous Single-Nucleotide Variants (nsSNVs) can create adverse effects on proteins as changing genotype and phenotype of any protein which may be a source of disease as cancer (Kulshreshtha et al., 2016).

Bioinformatics analysis is necessary to forecast participating Amino Acid Substitutions (AAS) to human diseases for each genome (Zeng et al., 2014). Expecting the functional effect of amino acid substitution

Abbreviations: Abbreviation, Annotation; nsSNVs, Non-Synonymous Single-Nucleotide Variants; MLTs, Machine Learning Techniques; SVM, Support Vector Machine; RF, Random Forest; ANN, Artificial Neural Network; SDM, Site-Directed Mutate; APOGEE, The pathogenicity Prediction through the Logistic Model tree; LMT, the Logistic Model Tree; CHASM, Cancer-Specific High-throughput Annotation of Somatic Mutations; HOPE, Homotopy optimization method; SIFT, The Sorting Intolerant from Tolerant; PROVEAN, Protein Variation Effect Analyzer; PANTHER, Protein Analysis Through Evolutionary Relationships; FATHMM, Functional Analysis through Hidden Markov Models; iFish, Integrated Functional inference of SNVs in human; SNAP, Screening for Non-Acceptable Polymorphisms; Polyphen-2, Polymorphism Phenotyping v2; PaPI, Pseudo Amino Acid Composition; GO, Gene Ontology; PSIC, Specific Independent Count; Condel, Combined Annotation Dependent Depletion; REVEL, Rare Exome Variant Ensemble Learner; COVEC, REVEL consensus Variant Effect Classification; RBF, radial based function; ACC, accuracy; AUC, Area under the curve; MCC, Math well Correlation Coefficient

* Corresponding author at: 12311, Egypt.
E-mail address: ms.hassan-elsayed@nrc.sci.eg (M.S. Hassan).

https://doi.org/10.1016/j.gene.2018.09.028
Received 17 August 2018; Accepted 14 September 2018
Available online 18 September 2018
0378-1119/

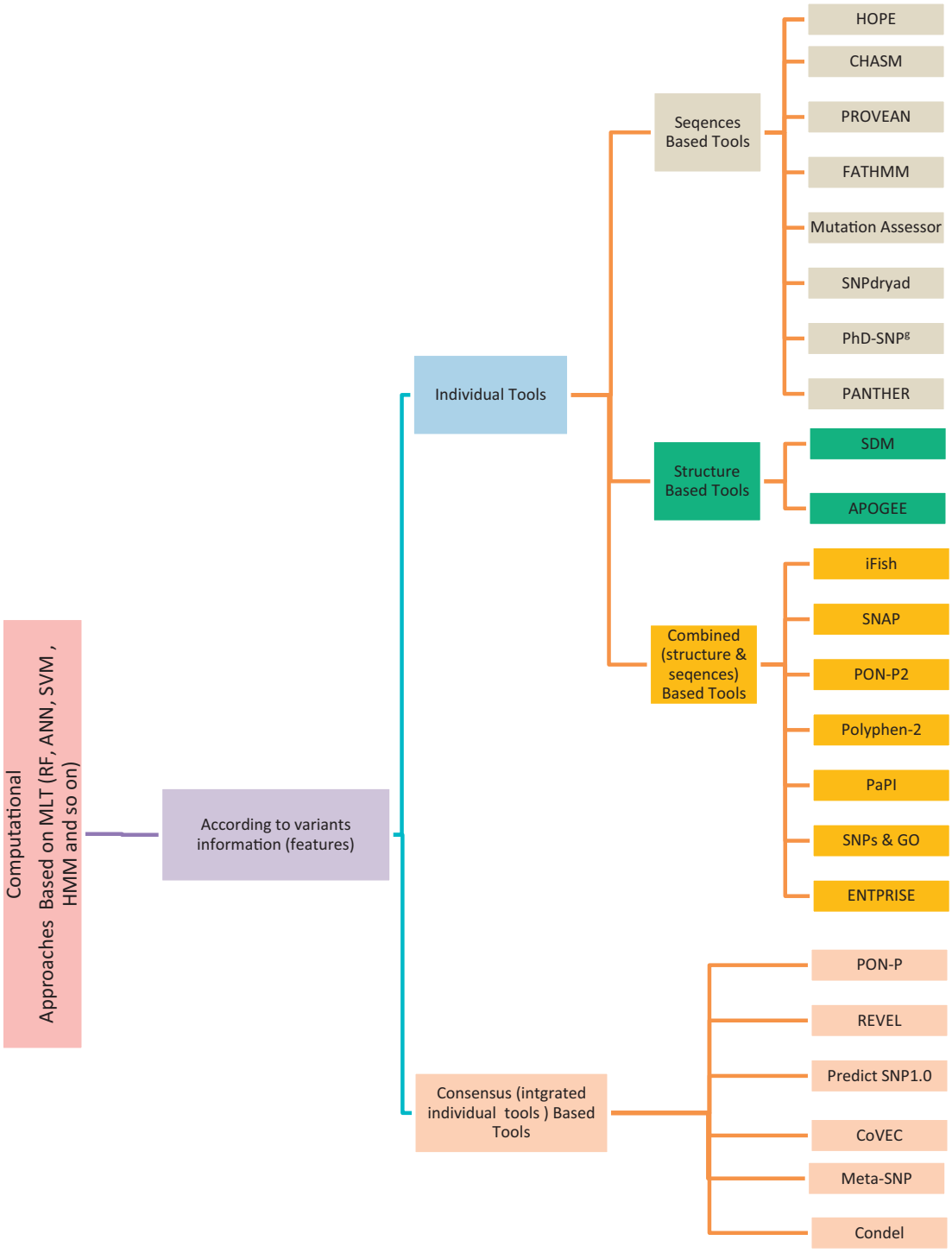

Fig. 1. Overall framework of the categories of computational approaches.

caused by nsSNPs is becoming increasingly significant as more and more novel variants are detected to distinguish between harmful and neutral mutations (Kulshreshtha et al., 2016; Zeng et al., 2014). Besides, early detection of nsSNVs and mutations can be contributory in prediction and diagnosing the disease at the primary stage (Kulshreshtha et al., 2016). One of the main challenges in the human genome is - to recognize the functional effects of SNVs. Although some of the variants found in genes assessed by the experimenter, many others have not estimated for their likely deleterious effects on protein function and structure (Hepp et al., 2015).

Many studies evaluated several tools on the same datasets. Thusberg et al. (2011) assessed the performance of nine pathogenicity prediction tools; MutPred, PolyPhen2, SIFT, nsSNPAnalyzer, Panther, PhD-SNP, PolyPhen, SNAP, and SNPs & GO. The methods were evaluated with a set of over 40,000 pathogenic and neutral variants. They also determine whether the type of original or substituting amino acid residue, the structural environment of the amino acid substitution, the structural class of the protein had an effect on the prediction performance (Thusberg et al., 2011). Grimm et al. (2015) evaluated and ranked twelve tools on five datasets (Grimm et al., 2015). We recently presented an evaluation study of eight individual tools on a VariBenchSelectedPure dataset and the integration model of the best of them. This