SAVDerma: An integrated tool and database for Single Amino Acid Variations associated with Dermatological Disorders

A DISSERTATION

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE AWARD OF THE DEGREE

OF

MASTER OF TECHNOLOGY

IN

BIOINFORMATICS

SUBMITTED BY

JAISHREE MEENA

(2K16/BIO/01)

UNDER THE SUPERVISION

OF

DR. YASHA HASIJA

DEPARTMENT OF BIOTECHNOLOGY
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road
Delhi-110042, India

JUNE 2018

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi-110042

CANDIDATE'S DECLARATION

I, **Jaishree Meena**, Roll no. 2K16/BIO/01, student of **M.Tech (Bioinformatics)**, hereby declare that the project Dissertation titled "**SAVDerma: An integrated tool and database for Single Amino Acid Variations associated with Dermatological Disorders**" which is submitted by me to the Department of Biotechnology, Delhi Technological University, Delhi in partial fulfilment of the requirement for the award of the degree of Master of Technology, is original and not copied from any source without proper citation. This work has not previously formed the basis for the award of any Degree, Diploma Associateship, Fellowship or other similar title or recognition.

Place: Delhi
Date:

Jaishree Meena
M.Tech. (Bioinformatics)
2K16/BIO/01

DEPARTMENT OF BIOTECHNOLOGY

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

CERTIFICATE

I, hereby certify that the Project Dissertation titled "**SAVDerma: An integrated tool and database for Single Amino Acid Variations associated with Dermatological Disorders**" which is submitted by **Jaishree Meena**, Roll no. 2K16/BIO/01, Department of Biotechnology, Delhi Technological University, Delhi in partial fulfilment of the requirement for the award of the degree of Master of Technology, is a record of the project work carried out by the student under my supervision. To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Place: Delhi
Date:

(**Dr. Yasha Hasija)**
Supervisor
Assistant Professor
Department of Biotechnology
Delhi Technological University

Title: “SAVDerma: An integrated tool and database for Single Amino Acid Variations associated with Dermatological Disorders

Jaishree Meena
2K16/BIO/01

ABSTRACT

Single Amino Acid Variations (SAVs) also called non-synonymous single nucleotide polymorphisms SNPs (nsSNPs), are the most inexhaustible type of single nucleotide polymorphisms (SNPs) that lead to amino acid substitutions in the protein products. Among different SAVs, some may cause pernicious infections and diseases while other amino acid substitutions are neutral which won't influence the capacity of the protein. Various researches on protein structures and functions have proposed that some SAVs are accountable for certain diseases, and it is responsible for about 60% of Mendelian diseases caused by amino acid substitutions. Studies on SAVs also showed that a mutation in an amino acid sequence leads to different skin disease. SAVs are of utmost relevance in checking their association with disease. SAVDerma is an integrated tool and database which contains information about SAVs and their association with Dermatological disorders. Data related to SAVs were retrieved through various databases and used to prepare training and testing data for machine learning. Various classifiers used to train data, among them Random forest classifier is used for testing data because of its highest accuracy of 87.29 percent and highest precision of 87.40 percent in predicting the association of SAVs in dermatological disorders. After compiling information on SAVs through various databases and utilising prediction data which gives insight about an SAV and its association with dermatological disorders, SAVDerma was developed. SAVDerma compiles highly curated physicochemical and sequence-based data of SAVs. Users can retrieve information regarding a particular SAV through three different parameters (HumsaVar FT id, Uniprot id and Swissport id). Results will show amino acid position at which variation is present along with their association with dermatological disease.

ACKNOWLEDGEMENTS

This dissertation is not only a part of my MTech degree but indeed also a very important part of my life which I'll remember and cherish forever. There were many ups and downs during this time, and I, from the bottom of my heart, want to thank and acknowledge everyone who has helped, supported and guided me in this journey of mine. First and foremost, my deepest gratitude goes to my mentor cum counsellor, Dr. Yasha Hasija, for her endless assistance both professionally and personally. Madam, this acknowledgement is not enough to describe your support and tell you how much I owe you. Thank you for all the inspiring discussions and for making me learn from my mistakes by "not holding my hand and walking with me but rather by not letting me fall". I enjoyed being taught by you immensely and you are an excellent teacher and have inspired me to continue learning with an open and positive mind. I feel extremely privileged to be a part of this department and a major part of this is due to the presence of excellent faculty members here. I am extremely thankful to all my other teachers, Prof. Jaigopal Sharma, Prof. B. D. Malhotra, Prof. Pravir Kumar, Dr. Asmita Das, Dr. Sourabh Saxena, Dr. Prakash Chandra for enriching me with all the basic details during the interactive classes which really helped me during my dissertation.

Thank you, Isha Shrivastava mam, for being so supportive and always encouraging me by appreciating my efforts and guiding and helping me with not just during my dissertation but personally too. Rajkumar Chakraborty, thank you for all the help you provided me during my dissertation and for your constant support. I am very thankful to you for being a true friend and a patient listener. Your views and opinions about certain things are truly exceptional and I have indeed learnt a lot from you. A very special thanks to Neha and Sunil for cheering me up and making me realize that I am not the only one who is facing problems. Thanks to all my classmates, Ruchi Sharma, Vikrant Khokhar, Rohan Ajit, Rohan Gupta and Varsha Singh for being there by my side and for constantly boosting me when it was required the most. You all will always remain to be a special part of my life.

Above all I would like to thank my Parents and God whom I consider equivalent for their love, patience and persistent support throughout my life. A very special thank you to my brother for patiently listening me and even providing suggestions, even though you didn't understand anything. Thank you to all my sisters for maintaining a cheerful environment at home and making me forget all my worries. Thank you, Roshan, for all the wonderful ways you make

me happy, even when you don’t realize it. Thanks for being so supportive during my entire duration of MTech.

Thanking you all for the support and help you gave me. I could have never done this without you all. Thank you once again!

Jaishree Meena
Department of Biotechnology
Delhi Technological University

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

SAVs	Single Amino Acid Variations
nsSNPs	Non-synonymous Single Nucleotide Polymorphisms
PSSM	Position-Specific Scoring Matrix
WEKA	Waikato Environment for Knowledge Analysis
ESD	Erythemato Squamous Skin Disease
DNA	Deoxyribonucleic Acid
SVM	Support Vector Machines
NN	Neutral Network
MLP	Multilayer Perceptron
CMS	Content Management System
SEO	Search Engine Optimization
GPL	General Public License
PROFEAT	Protein Feature Server
ML	Machine Learning
ROC	Receiver Operator Curve
MCC	Mathews Correlation Coefficient
SAVDerma	Single Amino Acid Variations related to Dermatological disorders

Chapter 1 Introduction

Dermatological disorders are one of the common illnesses that affects human and their lifestyle. It penetrates through all societies, happens at all ages, and influences between thirty to seventy percent of people, with much higher rates in at-risk subpopulations. The International Classification of Disease records more than thousand skin or skin-related ailments, a pattern overwhelmed by a couple of conditions representing the greater part of the skin illness load. However, despite this significant effect, skin illness keeps on getting generally little consideration in the national or worldwide health debate [1]. This investigation is aimed at examining the role of single amino acid variations (SAVs) in skin disorders by utilizing machine learning techniques. The data of SAVs can be utilized to follow the migration patterns of bygone humans and the ancestor line of present day humans. In many case, its most critical application might be to decipher the useful impact and effect of genomic variation, relating complex connections with phenotypes and making an interpretation of these disclosures into therapeutic practices [2].

In the post-genomic era, discrimination between disease-associated variants and neutral variants is of utmost significance, helping in comprehension of the genotype/phenotype correlations and creating treatment systems for diseases. From the viewpoint of diagnosis of disease and infection, it is likewise critical to distinguish between a neutral SAV and a non-neutral SAV. An assortment of computational strategies has been produced to foresee the functional effect of SAVs in a protein in the previous couple of decades employing methodologies, for example, statistical principles or machine learning algorithms. These methods utilise input features like amino acid sequence and the physicochemical properties associated with them, three-dimensional structure, evolutionary information, complex residue-contact network features etc. The clear majority of these techniques are being used by research community which can be applied as independent software or webservers to furnish free prediction of the functional effect of SAVs for educational and non-commercial purpose [3]. In this study, data related to dermatological disorders have been retrieved through various web resources and literature available online and all the SAVs, neutral as well as disease-associated were extracted from Humsavar.txt. Machine learning is carried out on the training data, model is prepared using different classifiers, and applied on test data to predict their association with dermatological disorders. On basis of prediction data, “SAVDerma: Single Amino Acid

Variations related to Dermatological disorders” is developed, it is an integrated tool and database, which compiles all the information of SAVs and their association with deramtological disorders. SAVDerma targets research community which provides service and solution in healthcare sector. Using data of SAVDerma, researchers can make tools for easy and fast diagnosis of skin diseases, moreover SAVDerma predicted novel SAVs whose role in skin disease have not been recognised, but they may have role in dermatological disorders. SAVs predicted to be associated with skin disorders can be targeted to make efficient drugs. Thus, this study is really helpful in providing machine learning based diagnostic systems and clinical advancements.

Chapter 2 Review of Literature

2.1 Overview

Skin is the major organ of human body having a prominent role in thermoregulation, sensation, protection from external environment, metabolism specially Vitamin D synthesis, helps in restricting fluid and water loss etc. Skin provides a complex barrier structure to the body. Socio-economic scarcity is the overall everyday reason of a higher risk of skin infections. People living within socio-economically deprived areas are further apt on the way to exhibit a skin infection than people beginning the slightest deprived areas. In India, due to environmental or personal factors, skin diseases are increasing among humans affecting their life. Extensive proliferation of skin infections and diseases and their associated clinical risks arise the need for elucidation of these skin diseases. Conventional diagnosis of skin diseases requires higher level of expertise and knowledge to recognise the disease from many similar diseases. The prevalence of skin related infections and disease in the all-inclusive community has shifted from 7.86% to 11.16% in different studies. Social and cultural stigma attached to skin diseases pose a lot of challenges to the sufferers. Diagnosis of skin diseases which employs the use of computer aided system is normally taken into consideration to accomplish the primary objective of accurate diagnosis. With the betterment of computing knowledge, computerisation of medical fields and data science implementations, diagnosis through computer aided systems has become an actuality. To develop such computer aided diagnosis, various data science techniques are being used, Machine learning is one such method which is now used for diagnosis purpose**.** In this study, single amino acid variations in protein related to skin disorders are taken into consideration and the data related to them have been retrieved through Profeat, Tango, Waltz, Limbo and Grantham like tools. Various Machine Learning classifiers are used in this study, but Random forest classifier was chosen because of its high accuracy and precision in predicting skin disease related single amino acid variations (SAVs). After predicting the association of SAVs in the skin disorders, an integrated tool and database has been developed, "SAVDerma: Single Amino Acid Variations related to Dermatological disorders. SAVDerma contains all the data related to SAVs and its association with skin disease. This tool will help in finding the association of an unknown SAV with skin disease. Its information can be used by Dermatologist for diagnosing skin disease in very less time.

2.2 Skin: The vital organ of human body

Skin being the largest and major organ have a prominent role in providing the interface between environment and the human. Weighing about an average of 4 kg and covering an area of 2 m^2, skin have many functions like thermoregulation, sensation, metabolism, protection from harsh external conditions, injury, chemicals and pathogens etc. It serves as a barrier to prevent the loss of vital body components like water from body. It mirrors the strength of the body and destruction of skin may lead to a person’s death reminding us of its major role of protection. Shockingly, at some time, almost everybody has some sort of dermatological condition. Skin issue contrast conspicuously in seriousness and side effects; it can be either changeless, transitory, painless or excruciating. Different external and internal factors which play important role in skin diseases. Some have situational causes, while others might be hereditary. Some skin conditions are minor, and others can be dangerous. The reason for the turmoil are not known. Skin is divided into 2 layers- outer one is called epidermis which is strongly and firmly attached and supported by connective tissues in the underlying inner layer called dermis. There is another layer beneath the dermis, called subcutaneous layer or hypodermis which is composed of loose connective tissues rich in fat which act as a shock absorber [4].

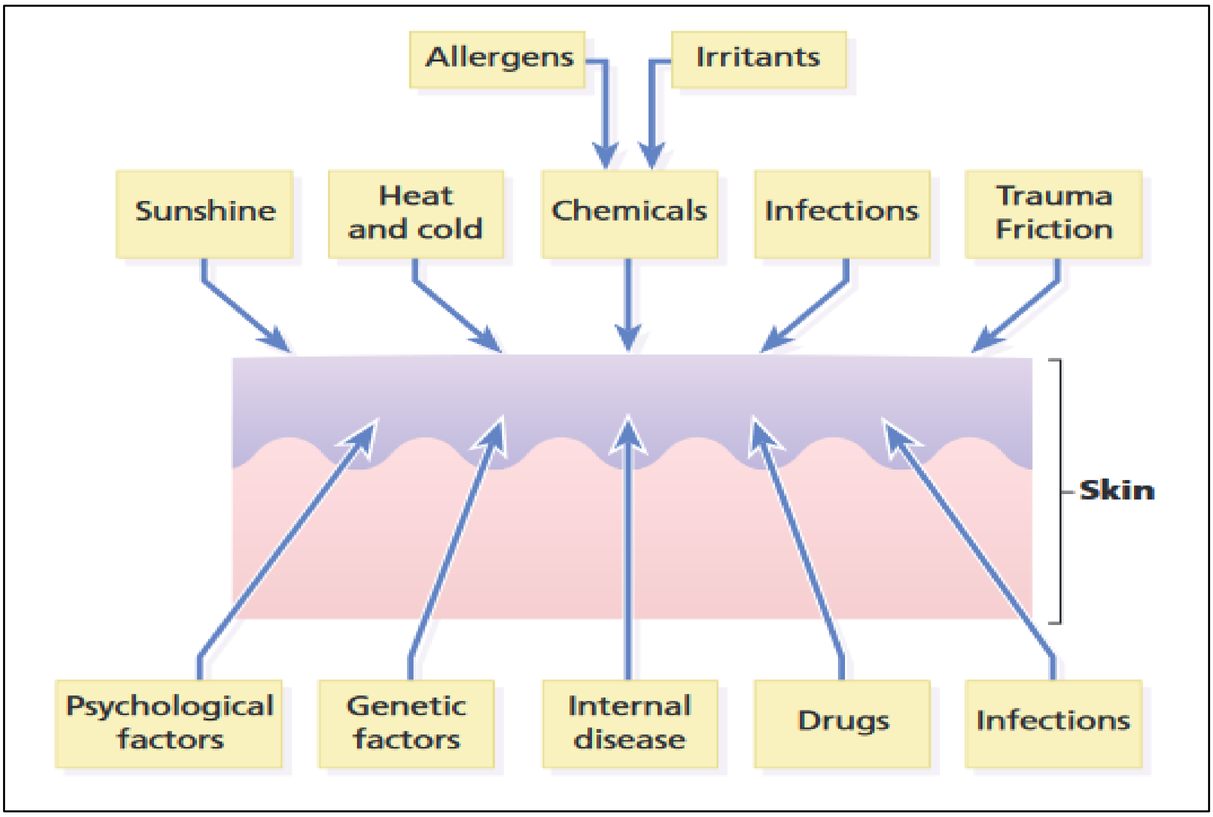


Figure 1: Different external and internal factors which play important role in skin diseases.

2.2.1 Epidermis

Epidermis is the outermost tough protective layer and is avascular in nature, means it has no blood supply, but it gets nourishment through blood vessels in the dermis. It is made of stratified squamous epithelial tissue. Depending on the site, the thickness of epidermis greatly varies. The epidermis comprises of 3 different types of cells namely Keratinocytes, Melanocytes and Langerhans cells. First type of cells are keratinocytes cells, which are the building blocks for the protein named keratin. Next type of cells in epidermis are melanocytes, that synthesises melanin, the pigment that gives colour to our skin. Very palest and the very darkest human skins have about the same number of melanocytes. Skin colour is not dependent on the number of melanocytes but instead it depends on the amount of melanin that a melanocyte contained. Third type of star shaped cells are Langerhans cells, originating in bone marrow. Once the Langerhans cells relocated towards the epidermis, their long thin tendrils circled the keratinocytes and invest quite a bit of time in ingesting the undesirable trespassers that are endeavouring to sneak around the skin. Merkel cells are the fourth type of cells which arises deep down at the ambit between the epidermis and the dermis [4].

2.2.2 Dermis

Dermis contains collagen and elastin and help connects the epidermis to the layer of skin underneath it. Dermis consists of blood capillaries and vessels and shelters the nerve fibres that recognise sensations like temperature, pressure and pain, as well as parts of hair follicles, ducts of oil and sweat glands that points towards the surface of the skin. It has three layers where all these functions take place, upper layer called papillary layer is composed of a thin sheet of areolar connective tissue that’s riddled with little peg like projections called dermal papillae. Just beneath the papillary layer, there is another layer called reticular layer which is deeper and thicker and made up of dense irregular connective tissue, making up to eighty percent of dermis. Dermis has the most important role in connecting epidermis to other skin layers as well as to make skin strong and flexible and in making skin to sense temperature, pressure and any kind of harm or injury. All these features of dermis make skin a very important and functional organ [4].

2.2.3 Hypodermis

Hypodermis consists of mostly adipose connective tissue and provides insulation, energy storage, shock absorption and helps anchor the skin. It contains cells like fibrocytes, macrophages and adipocytes [4].

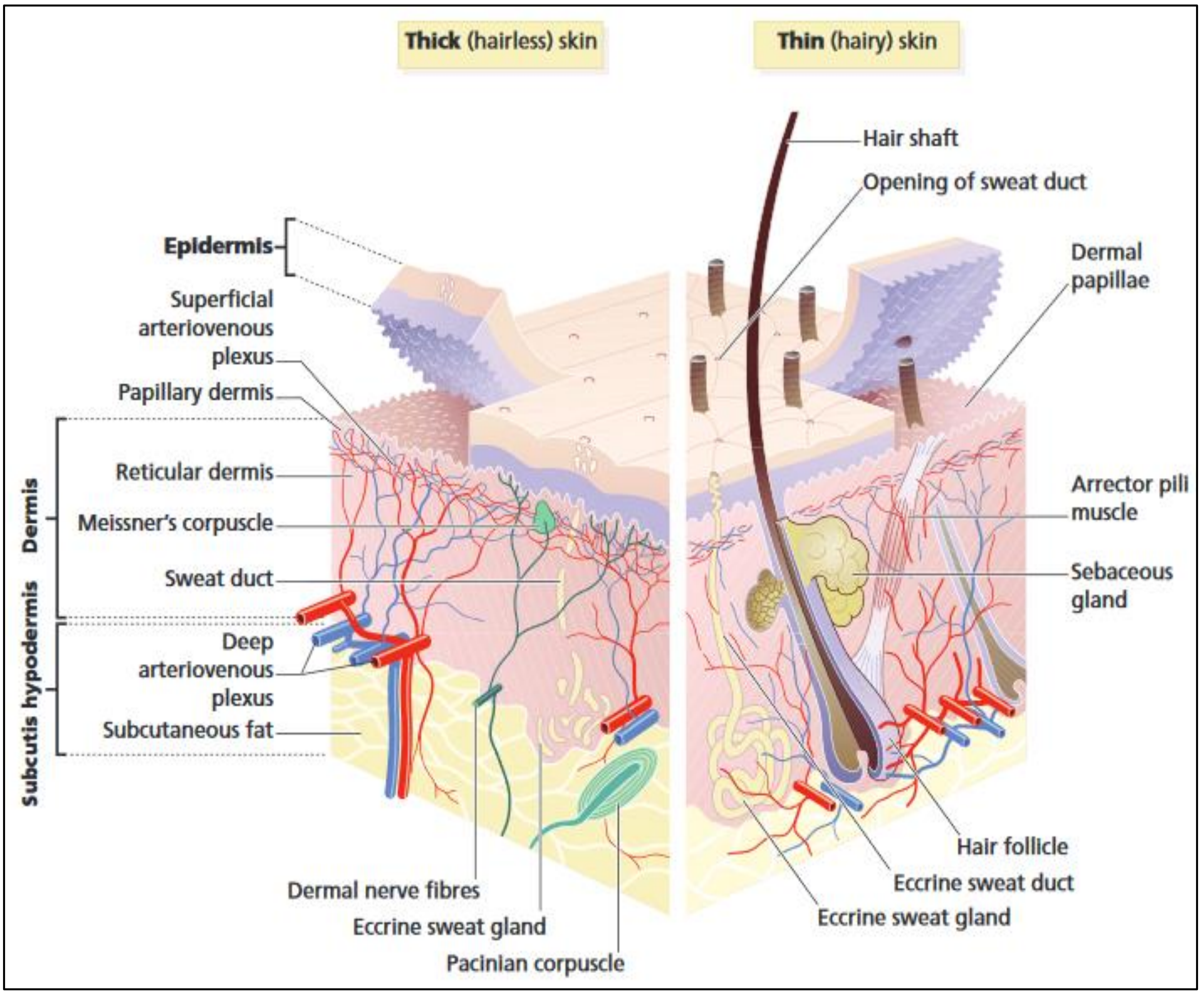


Figure 2: Cross-sectional area of Skin alongwith hair follicles.

2.3 Single Amino Acid Variations (SAVs) and their role in diseases

About 99.9 percent of DNA code is genetically alike in humans but the rest 0.1 percent leads to biological differences in individuals and may result in susceptibility to diseases among them. With the advancement in present whole genome sequencing, genetic testing is anticipated to be a routine procedure and its ability in discriminating between the disease-causing and neutral variants [5]. Though, anticipating the phenotype and association with disease, particularly for novel variations or variations happening in genes never ensnared in diseases, is certainly not a paltry issue [6].

Critical endeavours were endowed with creating methods to foresee disease related SAVs by employing techniques which are based on sequences or combination of sequence information with physico-chemical and structural characteristics [7][8]. In parallel, many methodologies were created to foresee the impacts of variations on thermodynamic properties, for example, folding and binding free energies. Some of the techniques focussed around determining the direction of the energy change, while others ascertain its extent or magnitude of change. These advancements show that research community has hunted down features and patterns to distinguish disease-causing variations from safe variations on a variety of levels like: sequencing, structure, physico-chemical and thermodynamics [9]. The execution among the best strategies using one or a few of the previously mentioned attributes, as far as distinguishing disease-causing variations or anticipating the free energy change, was observed to be quite, albeit direct correlation ought to be finished with care. Maybe this shows all pertinent data is accessible at each of these levels and essentially must be utilized properly [10].

Regardless of whether we accept that these distinctive levels contain a similar amount of data with respect to having the ability to distinguish disease-causing and innocuous variants, a researcher still needs to convey the predictions as well as to give point by point investigation of impacts initiated by the amino acid mutations [9]. Contingent upon the objectives of a specific study, this extra information may incorporate the perception that the variations happen at exceptionally conserved sequence position and thusly is anticipated to be disease-causing, including non-coding conserved position. For the reasons for another investigation, one might be occupied with making sense of whether the mutation includes modification of the physico-chemical properties of mutation site. Besides, specific studies could be concentrating on uncovering the alterations of the hydrogen bonds and salt extensions initiated by the amino acid change [11][12]. Along a similar line, much of the time, scientists will be intrigued to discover the impact of amino acid mutation on thermodynamic properties of the comparing protein. From various experimental methodologies employed to measure amino acid change, it is easier to estimate the last few quantities, changes of the folding and binding free energy, hydrogen bonds and salt extensions and subsequently giving extreme input for the *in-silico* modelling [10][13]. Here, we are especially keen on examining the contrast between disease-causing and innocuous variations showed at these four levels- sequencing, structure, physico-chemical properties and thermodynamics without conjuring predictions, using mainly experimental information taken from different databases.

2.4 Databases and Tools used for retrieving SAVs associated data

2.4.1 HumsaVar

HumsaVar is a compilation of all missense variants which are annotated in human UniprotKB/Swiss-Prot entries. It provides information of missense variants in the form of disease variants (missense variants having a role in disease), polymorphic or neutral variants (variants which are not reported in any kind of disease) and unclassified (variants whose role is still unknown and their role in causing disease is still doubtful). HumsaVar insightful data is very helpful for research communities in clinical and diagnostic applications.

2.4.2 PROFEAT

Sequence inferred structural and physicochemical highlights have been widely utilized for examining and anticipating structural, functional, articulation and interaction profiles of proteins and peptides. PROFEAT is a web server used for processing ordinarily employed features of proteins and peptides from amino acid sequence. The features retrieved through PROFEAT such as amino acid composition and physiochemical properties are exceedingly helpful for representing and recognizing proteins or peptides of various structural, functional and interaction profiles, which is fundamental for the effective utilization of statistical learning strategies in foreseeing the structural, functional and connection profiles of proteins and peptides regardless of sequence similarities [14].

2.4.3 TANGO

Protein aggregation is observed to be related to an expanding number of human ailments. In several cases, aggregation specifically adds to or modulates the pathology with which it is related. The method of activity of these protein aggregates in ailment is classified into loss-of-function and gain-of-function effects. Loss-of-function comes about because of the sequestration of misfolded proteins into dormant cell inclusions and can functionally be made even to a genetic deletion. Likewise, aggregated proteins can secure novel aggregation specific purpose that further add to the sickness. The impact of the mutations on protein aggregation is assessed with TANGO figuring the inherent aggregation propensity of the unfolded protein chain. TANGO depends on the physico-chemical principles of beta-sheet generation, stretched out by the belief that the core regions of an aggregate are completely buried [15][16].

2.4.4 LIMBO

For distinguishing sites of chaperone binding in proteins, position specific algorithm, Limbo is used. It depends on a position-specific scoring matrix (PSSM) trained from in vitro peptide binding information and structure modelling. Limbo based applications include: Predicting the mutational changes which influences binding of chaperone; Certain disease transformations which are identified as the consequence of modified chaperone binding can be recognized; Protein-fusions designs that can enhance solubility of proteins [17].

2.4.5 WALTZ

Various studies have tended to the impact of SAVs on protein steadiness, protein-protein connections and protein capacities. SAVs which are caused by nsSNPs frequently disrupts the function of protein by making alteration into protein structure as well as protein stability. SAVs can be very harmful when they play an important role in destroying functional binding sites in proteins. There are so many softwares and applications which provide data about the mutational effects of SAVs on protein structure and stability, for this study, we have used one such tool i.e., WALTZ. It is a position-specific prediction algorithm which is used to check amyloid propensity.

2.4.6 Grantham Matrix

The Grantham score endeavors to foresee the distance between two amino acids, in an evolutionary sense. A lower Grantham score indicates less evolutionary distance. A higher Grantham score indicates a more prominent evolutionary distance. Higher Grantham scores are viewed as more injurious: the more inaccessible two amino acids are, the more outlandish the amino acids are to be substituted with each other and the more distant two amino acids are, the more harming is their substitution. The distance scores go from 5 to 215 as published by Grantham [18].

2.5 Machine learning on the data retrieved for SAVs

Machine learning trains computers to do what easily falls into place for people and creatures: gain for a fact. Machine learning algorithms utilize computational techniques to "learn" data specifically from information without depending on a foreordained condition as a model. The algorithms adaptively enhance their execution as the quantity of tests accessible for learning increases. Machine learning algorithms discover common and natural patterns in information that create understanding and help in settling on better choices and forecasts [19]. They are utilized each day to make decisions for diagnosis of a medical condition, stock exchanging, energy load anticipation, and that's just the beginning. Media sites depend on machine learning to figure out how to filter through many alternatives to give a song or movie proposals to a particular user of that site. Retailers utilize it to pick up knowledge into their clients' purchasing behaviour. Machine learning utilizes two kinds of procedures: supervised learning, which prepares a model on known information and output information with the goal that it can foresee future yields, and unsupervised learning, which finds shrouded patterns or natural structures in input information.

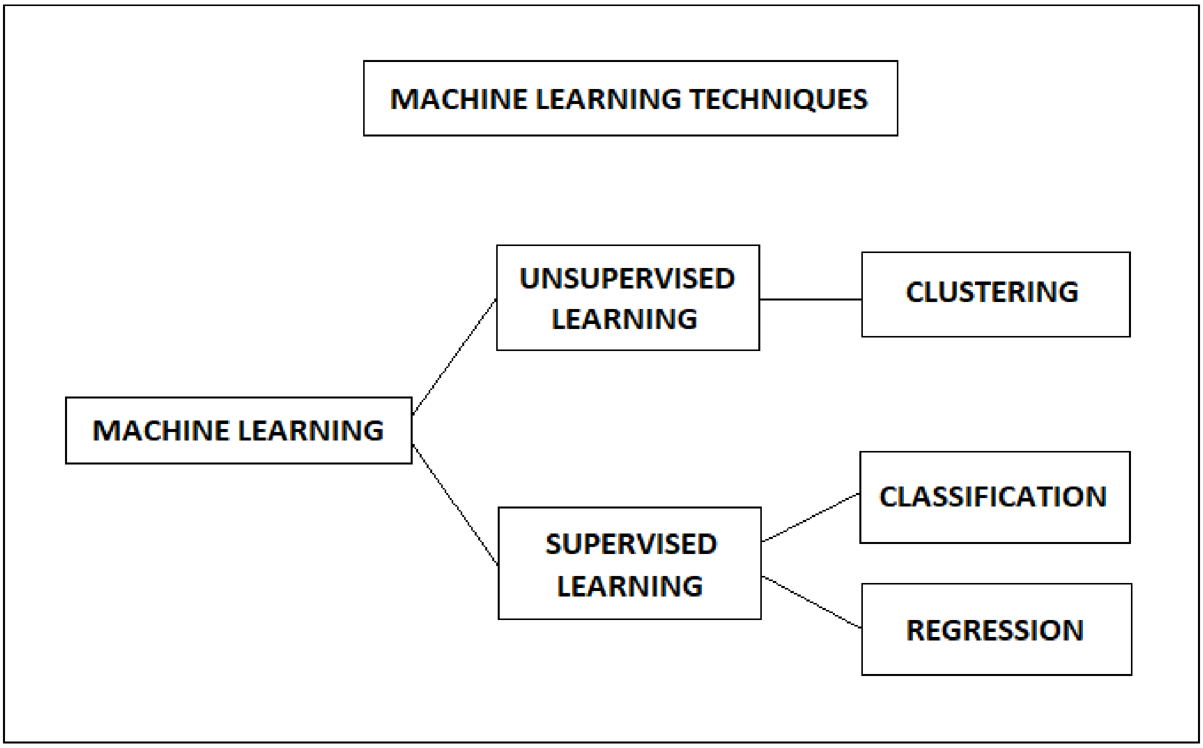


Figure 3: Types of Machine learning

The point of supervised machine learning is to manufacture a model that makes predictions based on facts within the sight of uncertainity. An algorithm based on supervised learning, takes a known set of input information and known reactions to the yielded information and prepares a model to create sensible predictions for the reaction to new information.

Supervised learning utilizes classification and regression methods to create prescient models.
• Classification procedures anticipate discrete reactions—for instance, regardless of whether an email is bona fide or spam, or whether a tumor is benign or harmful. Classification models input information into categories. Some applications of machine learning incorporate medical imaging, recognition of speech, and credit scoring.

• Regression procedures anticipate continuous reactions—for instance, changes in temperature or variations in power demand. Some applications of this procedure incorporate electricity load forecasting and algorithmic exchanging.

Unsupervised learning finds concealed or hidden patterns or inborn structures in information. It is utilized to draw inductions from datasets containing input information without marked responses.

Clustering is the most widely recognized unsupervised learning procedure. It is utilized for exploratory information investigation to discover shrouded patterns or groupings in information. Clustering based applications incorporate gene sequence examination, statistical surveying, and recognition of objects [19].

2.5.1 WEKA (Waikato Environment for Knowledge Analysis)

In this study, Waikato Environment for Knowledge Analysis (Weka) is used for machine learning tasks, it is an easy to use, user-friendly machine learning platform which is developed by Waikato University in New Zealand [20]. It is open source programming written in Java (GNU Public License) and utilized for research, training and projects. It comprises of group of machine learning algorithms for executing information mining errands which are really helpful in machine learning. Numerous classification strategies have been produced with the guide of learning algorithms, for example, Bayesian, DecisionTree, k-NN, Support Vector Machine (SVM) and boosting which show high accuracy in prediction of unknown data. Every one of these classifiers are fundamentally learning techniques and embrace sets of principles. In Classification, training instances are utilized to make a model that can group and classify the data tests into known classes.

Detection of Skin Disease is a topic of research for many scholars in the field of Machine Learning and Artificial Intelligence. Skin Disease Detection is fundamentally a classification assignment. Scientists have utilized distinctive data mining, statistics, machine learning algorithms in past. Machine learning based Neural networks and Support Vector Machine (SVM) are generally utilized in literature which gives great outline of various information mining procedures utilized for diagnosis of skin disease. In various studies conducted on the erythemato squamous skin disease (ESD), leucoderma, psoriasis and many other skin related diseases, researchers used various machine learning algorithms like Multilayer Perceptron (MLP), Naive Bayes, k-NN, random forest, decision tree etc to make predictions and showed the accuracy and precision of various classifiers in accurately predicting the disease class, showing the importance and utility of machine learning for diagnosis purpose.

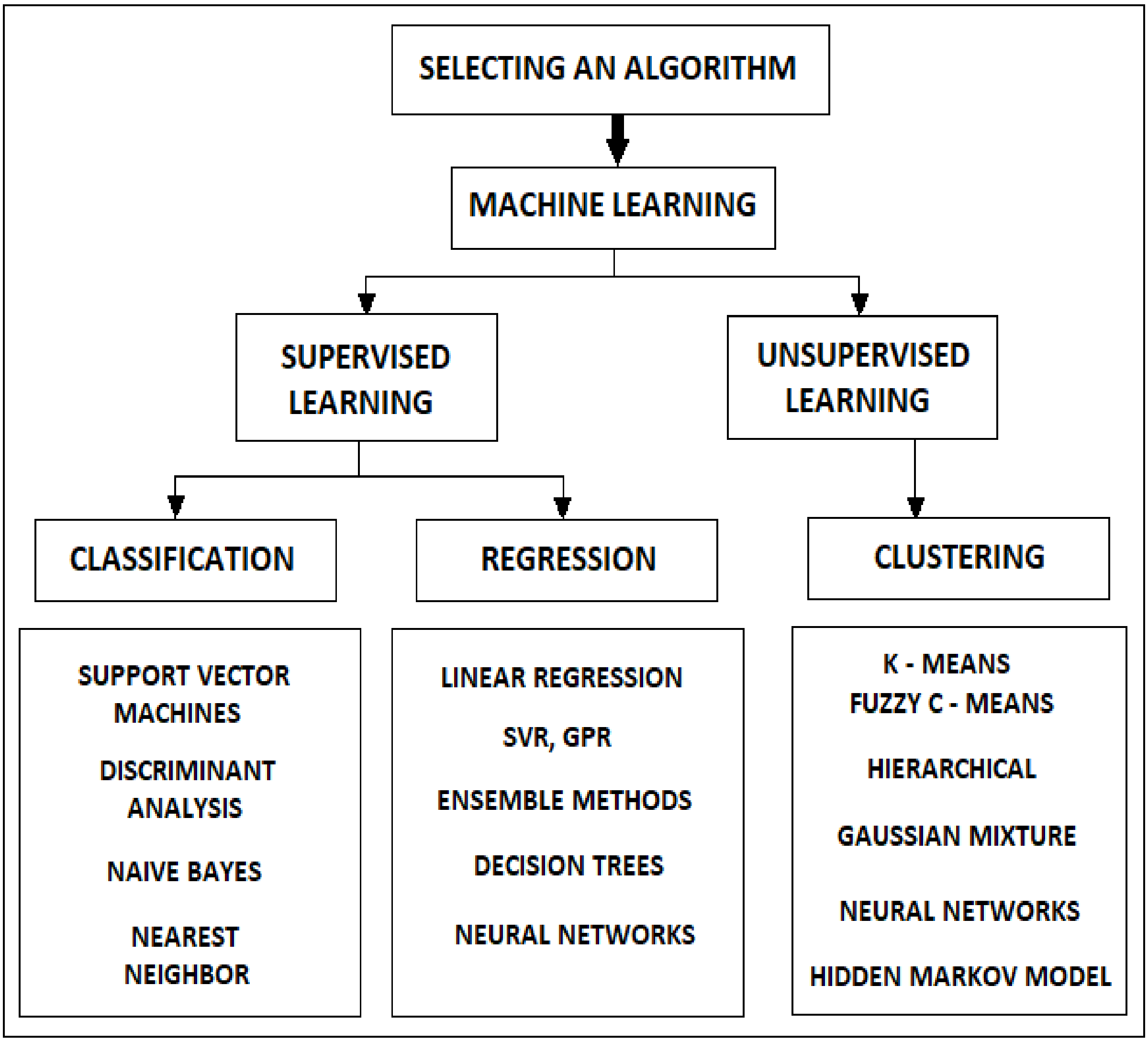


Figure 4: Selection of Machine learning algorithms

2.6 WordPress and Caspio for developing SAVDerma

WordPress is a Content Management System (CMS), and is an open source which was made to oversee online sites and blogs. WordPress enables its users to effortlessly make and deal with online blogs and sites content without coding which can be utilized to make a completely operational site. CMS is an application programming that gives a simple domain to deal with digital content information, for example, content, pictures, music, reports and so forth [21].

Properties of WordPress

WordPress is thought to be the most prominent content management framework because of its qualities.

- The most critical highlights of WordPress are that user can make a dynamic website with no programming and design information.

- Being theme-based WordPress gives you alternatives for many open source and premium theme subjects, which can be coordinated effortlessly with no designing information.
- Plugins broaden the usefulness of WordPress, which can be utilized to include new required modules.
- Websites created through WordPress are search engine optimization (SEO) friendly, it implies that sites worked in WordPress can be effortlessly optimised for internet searcher postings.
- Being Multilingual, WordPress enables its clients to decipher content in their dialect.
- WordPress has inbuilt Media Management System which is utilized to oversee pictures, music, archives and so forth and can be utilized with text content.

Advantages of WordPress

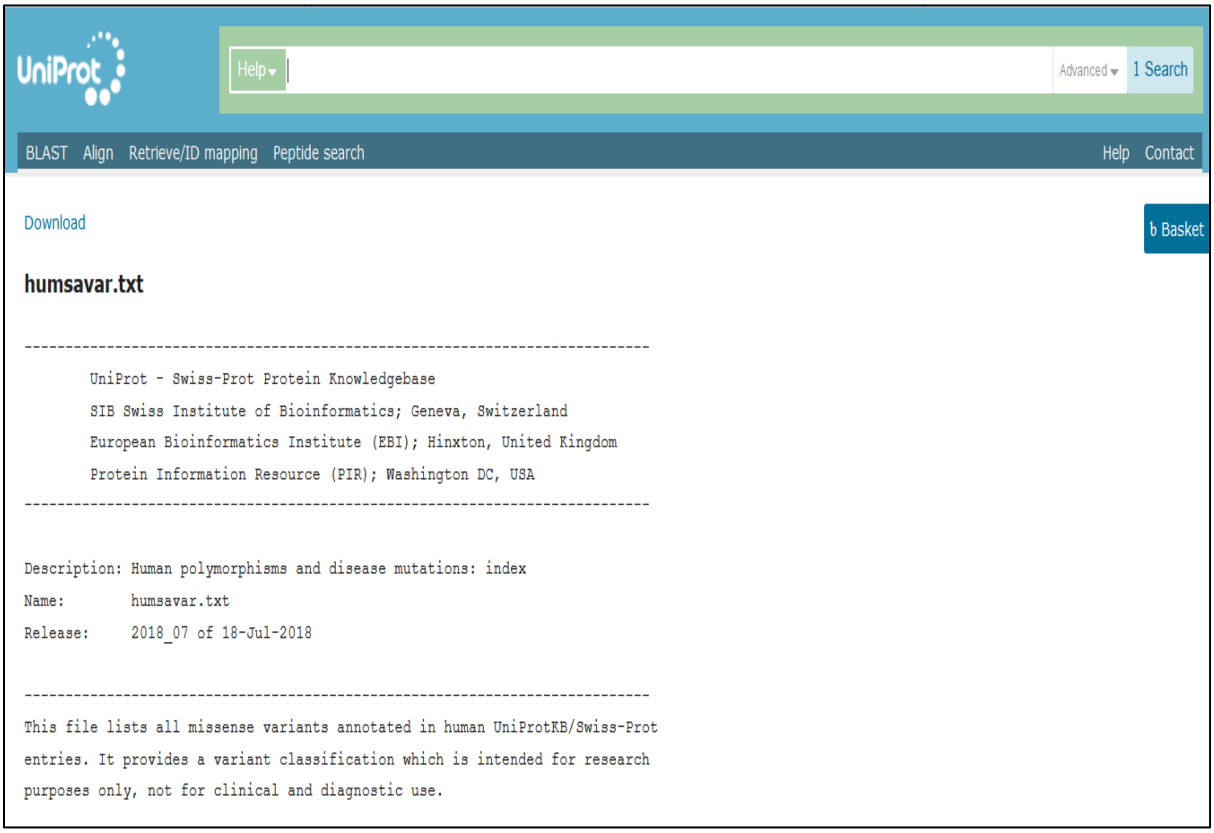
- WordPress is free, easy to use open source network under the GNU General Public License (GPL) [21].
- Customization of WordPress themes according to users need is very simple.
- It enables users to oversee clients with various roles and authorizations.
- WordPress media administration is snappy and quick and simple to utilize.
- WordPress gives WYSIWYG editor to deal with text content which is exceptionally helpful for controlling the format of the record.

Caspio Bridge gives an easy to understand development interface for creating database applications spreading over from a straightforward site to an intricate online business arrangement. Caspio Bridge contains development wizards which are helpful in creating applications, built-in forms and database, enabling developers to make a predefined arrangement utilizing input and inclinations taken from the client. The Caspio Bridge development system is coordinated with database development programming that is facilitated and executed completely by Caspio's cloud servers and is effortlessly conveyed inside a current or new site or application. The Caspio Bridge also provides platform for development of interactive mobile applications with careful customization attributes custom-made by the versatile working stage.

Chapter 3 Materials and Methods

3.1 Retrieval of SAVs

It is appealing to consolidate distinctive normal data sources and it has been found that both sequence-based, and topological information are basic for evaluating whether amino acid substitutions are infection or disease-related or not. Single amino acid substitutions related with comparative diseases are commonly depicted with high comparability of sequence-based features, functional properties and physical relationship between their products. In this way, for our model building reason, we included sequence-based and physico-chemical features. Genes identified to have a role in skin disease were curated from literature and after that they were mapped with Uniport ids. In this examination, we have focussed on single amino acid variations (SAVs) and recognized two classes of variations for creating positive and negative instances: skin disease causing and neutral polymorphisms (no role in skin sicknesses) [22]. By utilizing **HumsaVar**, Uniport ids were mapped with disease variations and neutral polymorphic variations. Protein sequences of all the human proteins were retrieved from **UniProt** and sequences for particular polymorphisms have been retrieved by changing the particular amino acid residues at particular positions as indicated by **HumsaVar** information.



humsavar.txt

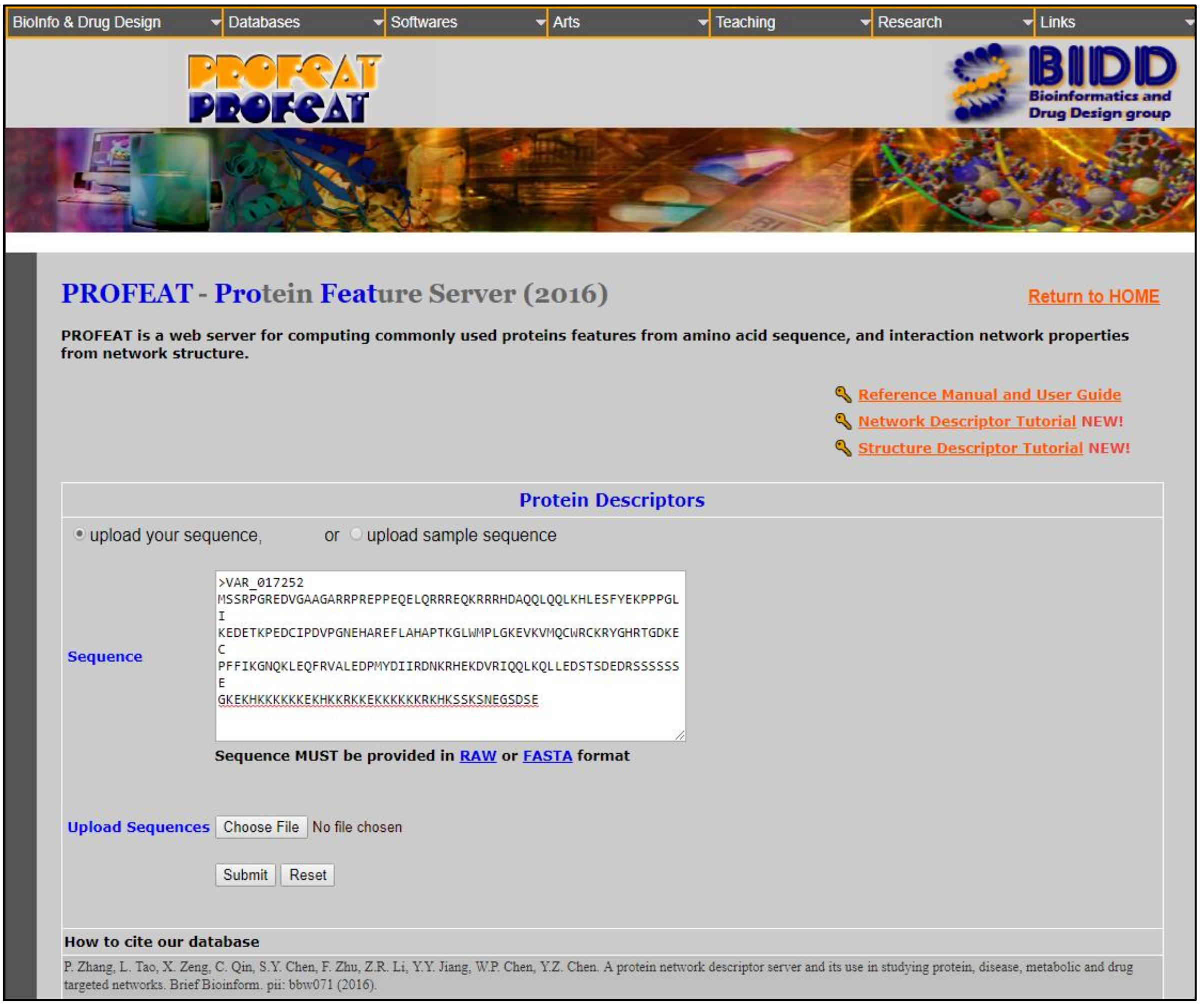
```
----------------------------------------------------------------------------
        UniProt - Swiss-Prot Protein Knowledgebase
        SIB Swiss Institute of Bioinformatics; Geneva, Switzerland
        European Bioinformatics Institute (EBI); Hinxton, United Kingdom
        Protein Information Resource (PIR); Washington DC, USA
----------------------------------------------------------------------------

Description: Human polymorphisms and disease mutations: index
Name:        humsavar.txt
Release:     2018_07 of 18-Jul-2018

----------------------------------------------------------------------------
This file lists all missense variants annotated in human UniProtKB/Swiss-Prot
entries. It provides a variant classification which is intended for research
purposes only, not for clinical and diagnostic use.
```

Figure 5: Welcome page HumsaVar.txt

3.2 Analysis of SAVs

3.2.1 Retrieving physico-chemical properties of SAVs

Sequence inferred structural and physicochemical highlights have been widely utilized for examining and anticipating structural, functional, articulation and interaction profiles of proteins and peptides. **PROFEAT** is a web server used for processing ordinarily employed features of proteins and peptides from amino acid sequence. The features retrieved through **PROFEAT** such as amino acid composition and physiochemical properties are exceedingly helpful [14] for representing and recognizing proteins or peptides of various structural, functional and interaction profiles, which is fundamental for the effective utilization of statistical learning strategies in foreseeing the structural, functional and connection profiles of proteins and peptides regardless of sequence similarities.



PROFEAT - Protein Feature Server (2016)

PROFEAT is a web server for computing commonly used proteins features from amino acid sequence, and interaction network properties from network structure.

Reference Manual and User Guide
Network Descriptor Tutorial NEW!
Structure Descriptor Tutorial NEW!

Protein Descriptors

● upload your sequence, or ○ upload sample sequence

Sequence

```
>VAR_017252
MSSRPGREDVGAAGARRPREPPEQELQRRREQKRRRHDAQQLQQLKHLESFYEKPPPGL
I
KEDETKPEDCIPDVPGNEHAREFLAHAPTKGLWMPLGKEVKVMQCWRCKRYGHRTGDKE
C
PFFIKGNQKLEQFRVALEDPMYDIIRDNKRHEKDVRIQQLKQLLEDSTSDEDRSSSSSS
E
GKEKHKKKKKKEKHKKRKKEKKKKKKRKHKSSKSNEGSDSE
```

Sequence MUST be provided in RAW or FASTA format

Upload Sequences Choose File No file chosen

Submit Reset

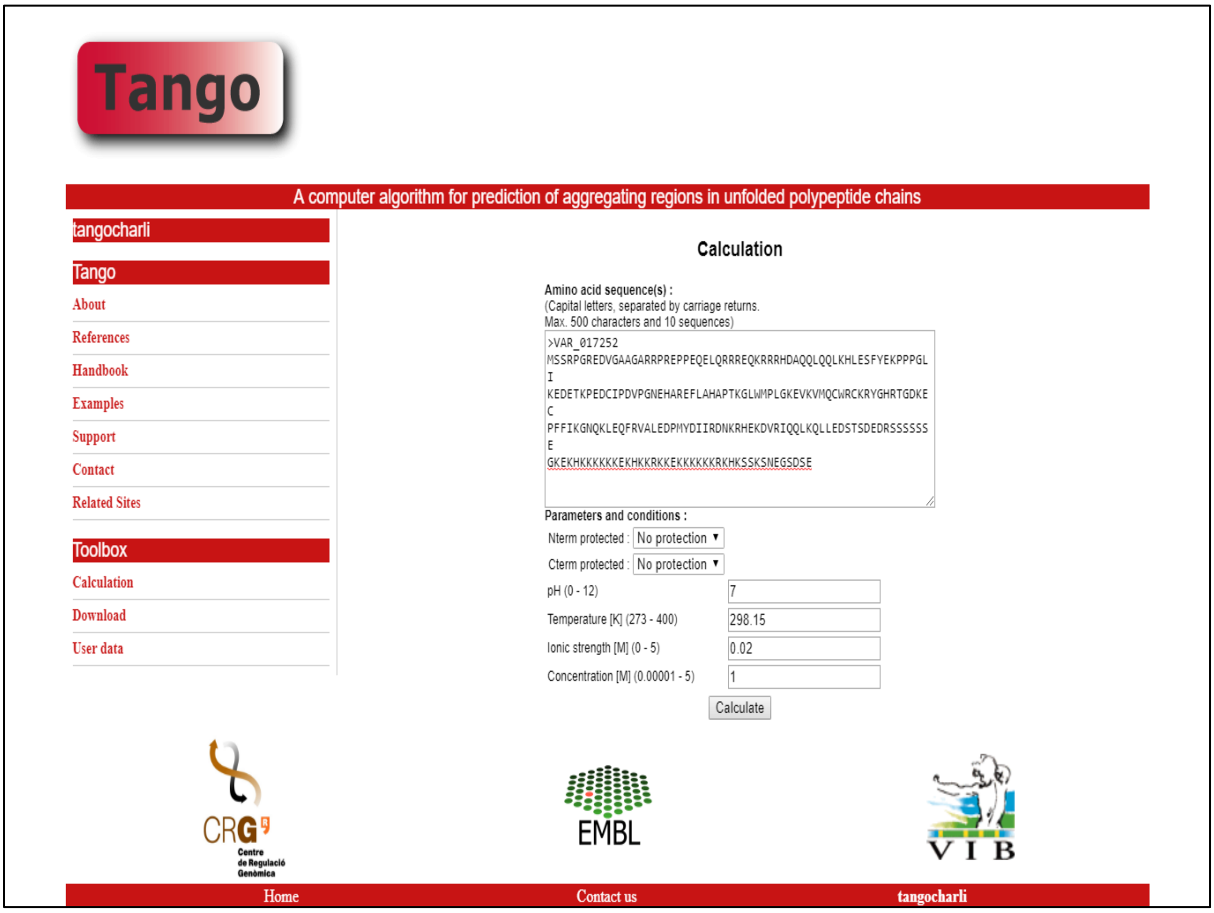
How to cite our database

P. Zhang, L. Tao, X. Zeng, C. Qin, S.Y. Chen, F. Zhu, Z.R. Li, Y.Y. Jiang, W.P. Chen, Y.Z. Chen. A protein network descriptor server and its use in studying protein, disease, metabolic and drug targeted networks. Brief Bioinform. pii: bbw071 (2016).

Figure 6: Query submission page of PROFEAT

3.2.2 Retrieving properties related to mutational effect of SAVs

Protein aggregation is observed to be related to an expanding number of human ailments. In several cases, aggregation specifically adds to or modulates the pathology with which it is related. The method of activity of these protein aggregates in ailment is classified into loss-of-function and gain-of-function effects. Loss-of-function comes about because of the sequestration of misfolded proteins into dormant cell inclusions and can functionally be made even to a genetic deletion. Likewise, aggregated proteins can secure novel aggregation specific purpose that further add to the sickness. The impact of the mutations on protein aggregation is assessed with **TANGO** figuring the inherent aggregation propensity of the unfolded protein chain. **TANGO** depends on the physico-chemical principles of beta-sheet generation, stretched out by the belief that the core regions of an aggregate are completely buried [15][16].



Calculation

Amino acid sequence(s) :
(Capital letters, separated by carriage returns.
Max. 500 characters and 10 sequences)

```
>VAR_017252
MSSRPGREDVGAAGARRPREPPEQELQRRREQKRRRHDAQQLQQLKHLESFYEKPPPGL
I
KEDETKPEDCIPDVPGNEHAREFLAHAPTKGLWMPLGKEVKVMQCWRCKRYGHRTGDKE
C
PFFIKGNQKLEQFRVALEDPMYDIIRDNKRHEKDVRIQQLKQLLEDSTSDEDRSSSSSS
E
GKEKHKKKKKKEKHKKRKKEKKKKKKRKHKSSKSNEGSDSE
```

Parameters and conditions :

Nterm protected : No protection ▾

Cterm protected : No protection ▾

pH (0 - 12) 7

Temperature [K] (273 - 400) 298.15

Ionic strength [M] (0 - 5) 0.02

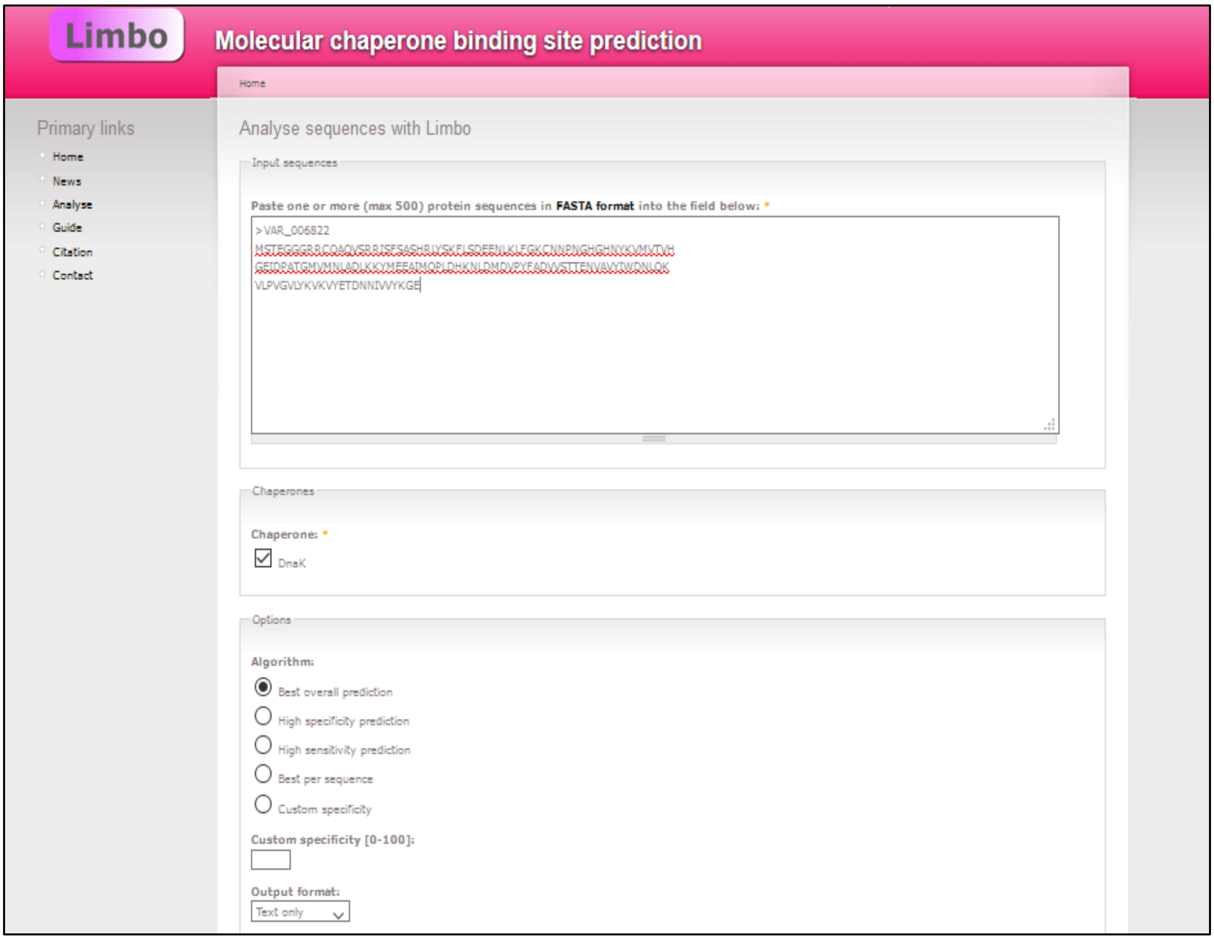
Concentration [M] (0.00001 - 5) 1

Calculate

Figure 7: Query submission page of TANGO

For distinguishing sites of chaperone binding in proteins, position specific algorithm, **LIMBO** is used. It depends on a position-specific scoring matrix (PSSM) trained from in vitro peptide binding information and structure modelling [17]. **LIMBO** based applications include:

Predicting the mutational changes which influences binding of chaperone; Certain disease transformations which are identified as the consequence of modified chaperone binding can be recognized; Protein-fusions designs that can enhance solubility of proteins.



Analyse sequences with Limbo

Input sequences

Paste one or more (max 500) protein sequences in **FASTA format** into the field below: *

```
>VAR_006822
MSTEGGGRRCQAQVSRRISFSASHRLYSKFLSDEENLKLFGKCNNPNGHGHNYKVMVTVH
GEIDPATGMVMNLADLKKYMEEAIMQPLDHKNLDMDVPYFADVVSTTENVAVYIWDNLQK
VLPVGVLYKVKVYETDNNIVVYKGE
```

Chaperones

Chaperone: *

☑ DnaK

Options

Algorithm:

● Best overall prediction

○ High specificity prediction

○ High sensitivity prediction

○ Best per sequence

○ Custom specificity

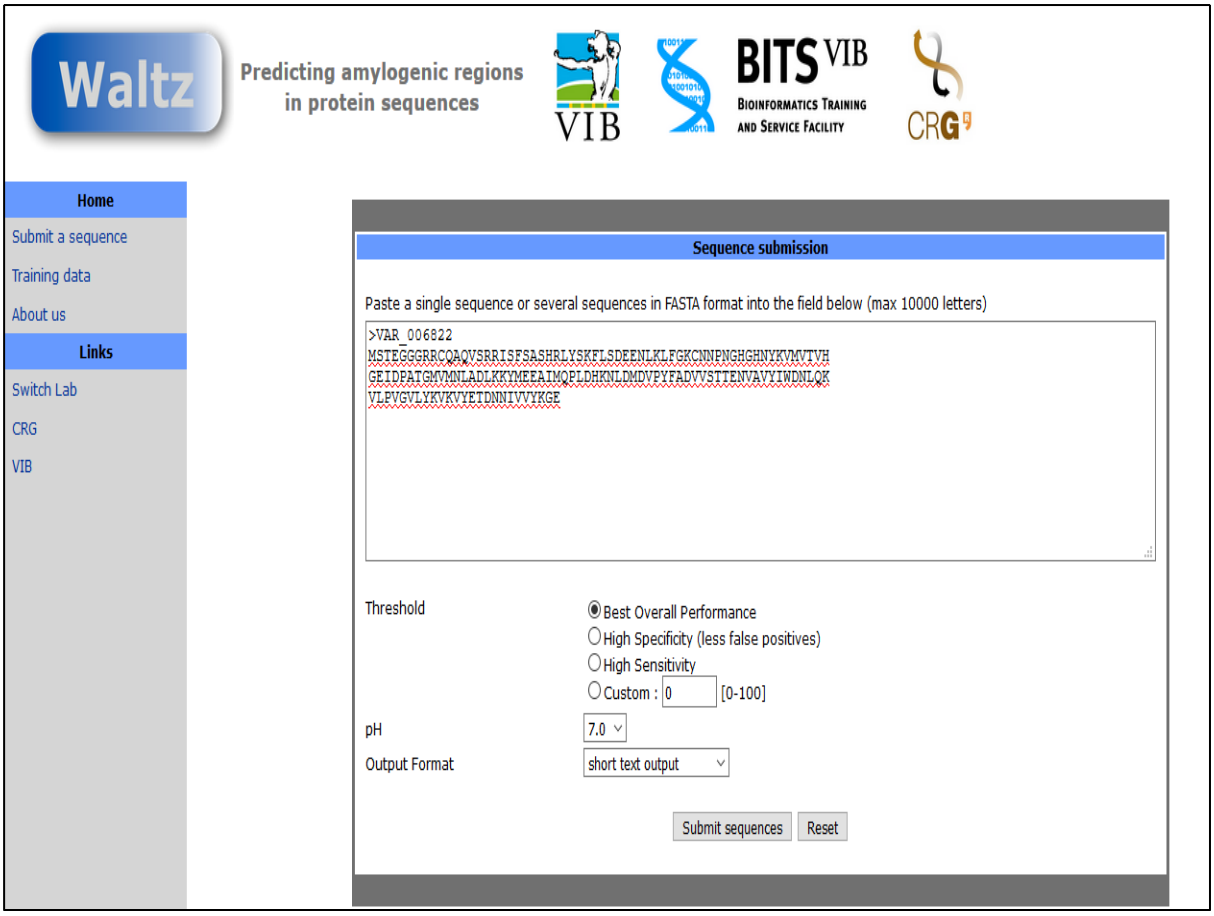
Custom specificity [0-100]:

Output format:

Text only

Figure 8: Query submission page of LIMBO

Various studies have tended to the impact of SAVs on protein steadiness, protein-protein connections and protein capacities. SAVs which are caused by nsSNPs frequently disrupts the function of protein by making alteration into protein structure as well as protein stability. SAVs can be very harmful when they play an important role in destroying functional binding sites in proteins. There are so many softwares and applications which provide data about the mutational effects of SAVs on protein structure and stability, for this study, we have used one such tool i.e., **WALTZ**.



Waltz
Predicting amylogenic regions in protein sequences
VIB
BITS VIB
Bioinformatics Training and Service Facility
CRG
Home
Submit a sequence
Training data
About us
Links
Switch Lab
CRG
VIB

Sequence submission

Paste a single sequence or several sequences in FASTA format into the field below (max 10000 letters)

```
>VAR_006822
MSTEGGGRRCQAQVSRRISFSASHRLYSKFLSDEENLKLFGKCNNPNGHGHNYKVMVTVH
GEIDPATGMVMNLADLKKYMEEAIMQPLDHKNLDMDVPYFADVVSTTENVAVYIWDNLQK
VLPVGVLYKVKVYETDNNIVVYKGE
```

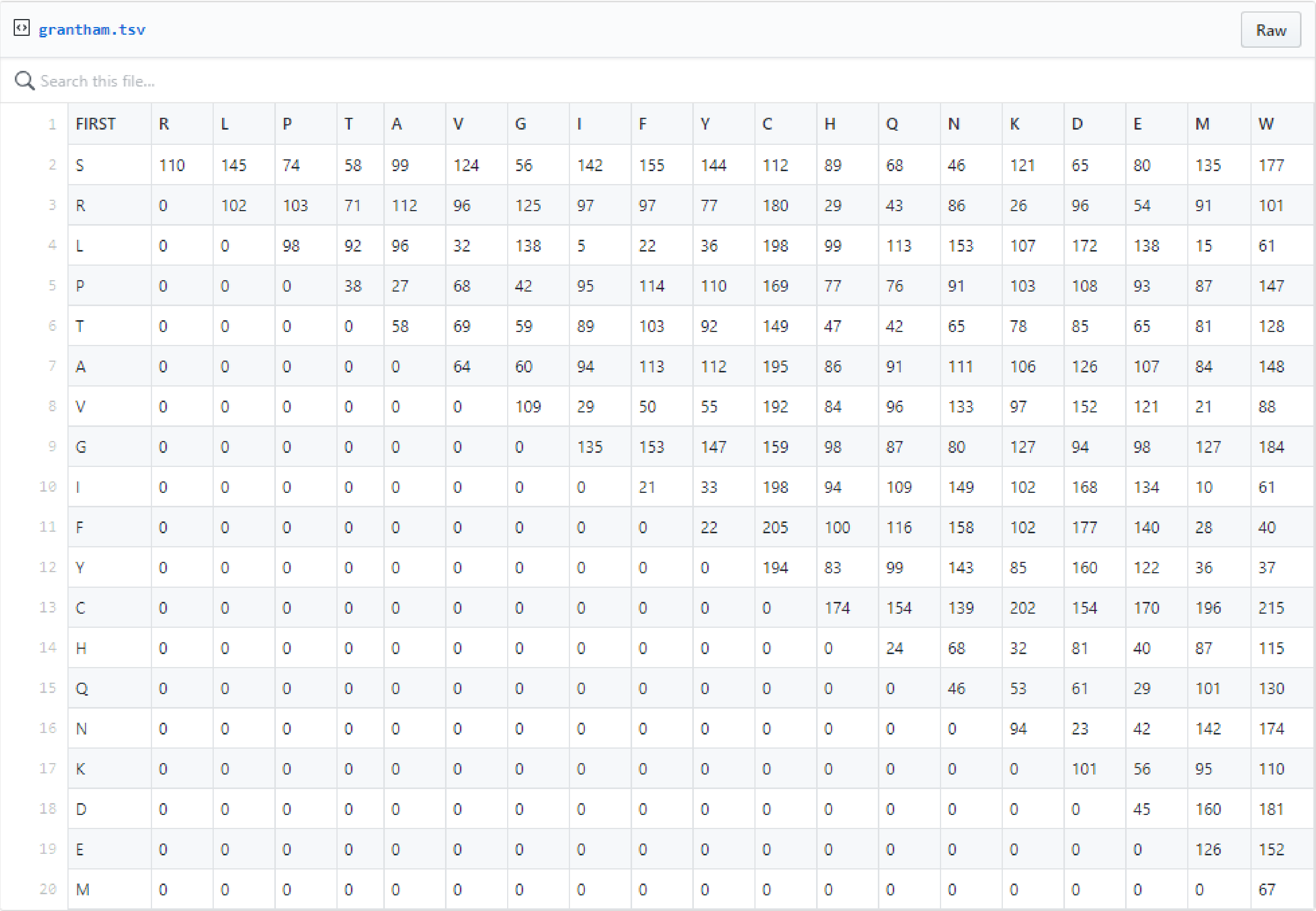
Threshold ● Best Overall Performance
○ High Specificity (less false positives)
○ High Sensitivity
○ Custom : 0 [0-100]

pH 7.0

Output Format short text output

Figure 9: Query submission page of WALTZ

The **Grantham score** endeavours to foresee the distance between two amino acids, in an evolutionary sense. A lower **Grantham score** indicates less evolutionary distance. A higher **Grantham score** indicates a more prominent evolutionary distance. Higher **Grantham scores** are viewed as more injurious: the more inaccessible two amino acids are, the more outlandish the amino acids are to be substituted with each other and the more distant two amino acids are, the more harming is their substitution. The distance scores go from 5 to 215 as published by Grantham [18].



FIRST	R	L	P	T	A	V	G	I	F	Y	C	H	Q	N	K	D	E	M	W
S	110	145	74	58	99	124	56	142	155	144	112	89	68	46	121	65	80	135	177
R	0	102	103	71	112	96	125	97	97	77	180	29	43	86	26	96	54	91	101
L	0	0	98	92	96	32	138	5	22	36	198	99	113	153	107	172	138	15	61
P	0	0	0	38	27	68	42	95	114	110	169	77	76	91	103	108	93	87	147
T	0	0	0	0	58	69	59	89	103	92	149	47	42	65	78	85	65	81	128
A	0	0	0	0	0	64	60	94	113	112	195	86	91	111	106	126	107	84	148
V	0	0	0	0	0	0	109	29	50	55	192	84	96	133	97	152	121	21	88
G	0	0	0	0	0	0	0	135	153	147	159	98	87	80	127	94	98	127	184
I	0	0	0	0	0	0	0	0	21	33	198	94	109	149	102	168	134	10	61
F	0	0	0	0	0	0	0	0	0	22	205	100	116	158	102	177	140	28	40
Y	0	0	0	0	0	0	0	0	0	0	194	83	99	143	85	160	122	36	37
C	0	0	0	0	0	0	0	0	0	0	0	174	154	139	202	154	170	196	215
H	0	0	0	0	0	0	0	0	0	0	0	0	24	68	32	81	40	87	115
Q	0	0	0	0	0	0	0	0	0	0	0	0	0	46	53	61	29	101	130
N	0	0	0	0	0	0	0	0	0	0	0	0	0	0	94	23	42	142	174
K	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	101	56	95	110
D	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	45	160	181
E	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	126	152
M	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	67

Figure 10: Grantham score matrix providing information about the codon replacements in terms of conservation score.

3.3 Machine learning approaches used in the study to predict association of SAVs with Dermatological disorders

In this study, Waikato Environment for Knowledge Analysis (Weka) is used for machine learning tasks, it is an easy to use, user-friendly machine learning platform which is developed by Waikato University in New Zealand. It is open source programming written in Java (GNU Public License) and utilized for research, training and projects. It comprises of group of machine learning algorithms for executing information mining errands which are really helpful in machine learning. Numerous classification strategies have been produced with the guide of learning algorithms, for example, Bayesian, DecisionTree, k-NN, Support Vector Machine (SVM) and boosting which show high accuracy in prediction of unknown data. Every one of these classifiers are fundamentally learning techniques and embrace sets of principles. Bayesian Decision Theory is the main source for developing bayesisan classifiers [23][24]. In Classification, training instances are utilized to make a model that can group and classify the data tests into known classes. The Classification procedure includes following steps:

a. Creating a set of training data.

b. Identifying the class attribute and classes.

c. Identifying useful and relevant attributes for classification.

d. Using training instances in a set of training data to make learn a model

e. Using the model for classification of the unknown data samples.

Various classification methodologies have been utilised and reported in the literature for different types of machine learning applications [23]. To check the performance of various machine learning classifiers, features like accuracy, precision [22], f-measure, recall, MCC, ROC curve are measured[22], which are defined as follows:

- Accuracy-

$$\text{ACCURACY} = \frac{\text{TRUE POSITIVE} + \text{TRUE NEGATIVE}}{\text{TRUE POSITIVE} + \text{TRUE NEGATIVE} + \text{FALSE POSITIVE} + \text{FALSE NEGATIVE}}$$

- Precision -.

$$\text{PRECISION} = \frac{\text{TRUE POSITIVE}}{\text{TRUE POSITIVE} + \text{FALSE POSITIVE}}$$

- Recall –

$$\text{RECALL} = \frac{\text{TRUE POSITIVE}}{\text{TRUE POSITIVE + FALSE POSITIVE}}$$

- F-measure –

$$\text{F - MEASURE} = \frac{2 * (\text{PRECISION} * \text{RECALL})}{(\text{PRECISION} + \text{RECALL})}$$

- MCC (Mathews Correlation Coefficient) – It is the quality measure of binary classifications which is a two-class based classification in machine learning. It is based on true positives and negatives and false positives and negatives as shown in the formula below.

$$\text{MCC} = \frac{(\text{TP} * \text{TN}) - (\text{FP} * \text{FN})}{\sqrt{(\text{TP} + \text{FP})\ (\text{TP} + \text{FN})\ (\text{TN} + \text{FP})\ (\text{TN} + \text{FN})}}$$

here, MCC = MATHEWS CORRELATION COEFFICIENT,
TP = TRUE POSITIVE,
TN = TRUE NEGATIVE,
FP = FALSE POSITIVE,
FN = FALSE NEGATIVE

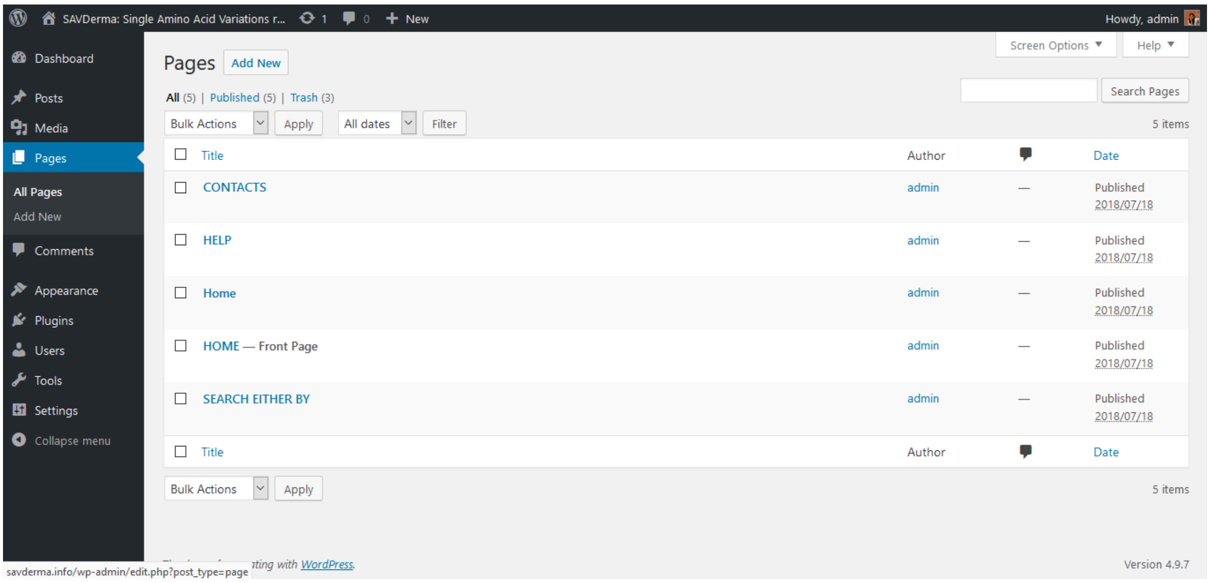
In this study, seven machine learning classifiers were used for prediction purpose, namely random forest, J48 tree, function logistic, JRip classifier, filtered classifier, classification via regression and multilayer perceptron classifier for training data. Random forest was chosen to make predictions on test data due to its high accuracy and precision in prediction tasks.

Machine learning features like Accuracy, Precision, Recall, F-measure, Mathews correlation coefficient, and ROC curve were measured for data on SAVs to select optimum classifier and after selecting the classifier, it was applied on the unclassified test datasets of SAVs to predict their association with skin diseases.

3.4 Development of SAVDerma

SAVDerma was developed using WordPress and Caspio. SAVDerma contains all the information of these databases and hence showed highest accuracy in predicting the test data.

WordPress is a Content Management System (CMS), and is an open source which was made to oversee online sites and blogs. WordPress enables its users to effortlessly make and deal with online blogs and sites content without coding which can be utilized to make a completely operational site. Content Management System (CMS) is an application programming that gives a simple domain to deal with digital content information, for example, content, pictures, music, reports and so forth [21].



☐	Title	Author	💬	Date
☐	CONTACTS	admin	—	Published 2018/07/18
☐	HELP	admin	—	Published 2018/07/18
☐	Home	admin	—	Published 2018/07/18
☐	HOME — Front Page	admin	—	Published 2018/07/18
☐	SEARCH EITHER BY	admin	—	Published 2018/07/18
☐	Title	Author	💬	Date

Figure 11: User Interface of WordPress

Chapter 4 Results

4.1 Collection of SAVs data and Machine learning applications

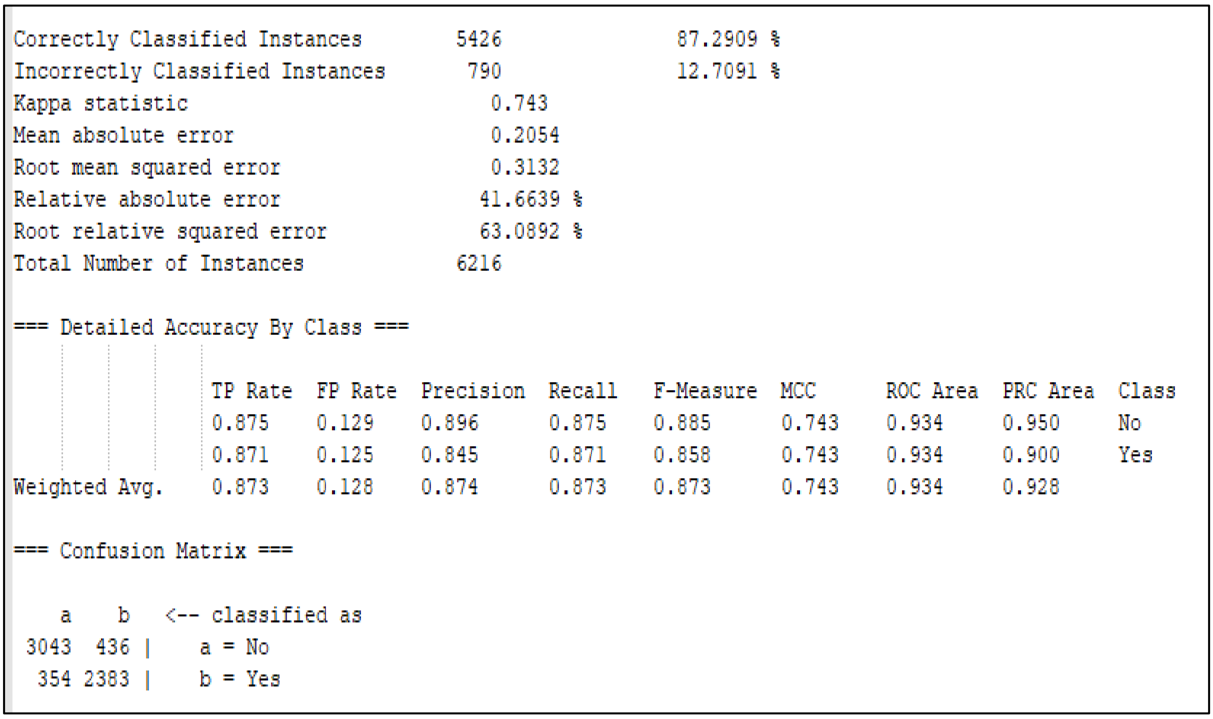
Genes identified to have a role in skin disease were curated from literature and after that they were mapped with Uniport ids. In this study, we have focussed on SAVs and recognized two classes of variations for creating positive and negative instances: skin disease causing and neutral polymorphisms. By utilizing HumsaVar.txt, Uniport ids were mapped with 4279 instances of disease variations and 4895 instances of neutral polymorphic variations. Protein sequences of all the human proteins were retrieved from UniProt and sequences for particular polymorphisms have been retrieved by changing the particular amino acid residues at particular positions as indicated by HumsaVar information [refer to Annexure 1].

PROFEAT program was then used to collect physico-chemical information of disease-related and neutral SAVs. Data about mutational effect of SAVs on aggregation propensity was collected through TANGO program. To collect data about mutational effect of SAVs on amyloid propensity and chaperone binding, WALTZ and LIMBO algorithms were used. Grantham score was calculated for all the instances which gives insights about the codon replacements in terms of conservation score. This data collected to make training datasets for machine learning. Data on unclassified SAVs, from HumsaVar.txt that are not recognised in any kind of skin disease was also collected, which then used as testing data to predict the association of SAVs with dermatological disorders.

4.2 Selection of Optimum Classifier

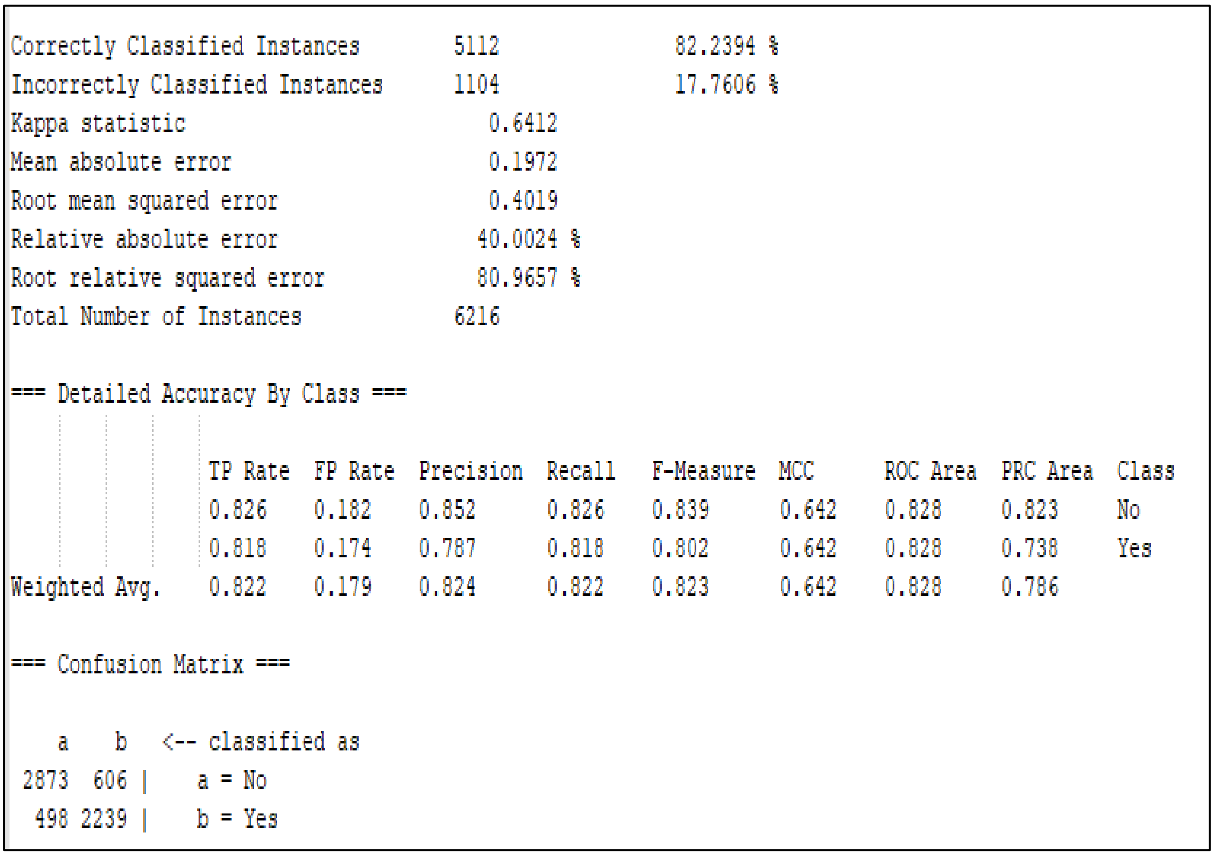
It is very important to choose a suitable machine learning classifier which can be applied on training data providing highest accuracy and precision for prediction purpose. So, for selecting optimum classifier, we used several machine learning algorithms and check their performance in terms of accuracy, precision, f-measure, recall, MCC value and ROC curve. Seven machine learning classifiers namely Random forest, J48 tree, Function logistic, JRip classifier, Filtered classifier, Classification via Regression and Multilayer perceptron were used on training data to make prediction on association of SAVs with skin diseases. Following results regarding selection of optimum classifier, were obtained for various classifiers used in the study. Based on the results we choose Random forest classifier for further analysis of single amino acid variations.

Random Forest Classifier



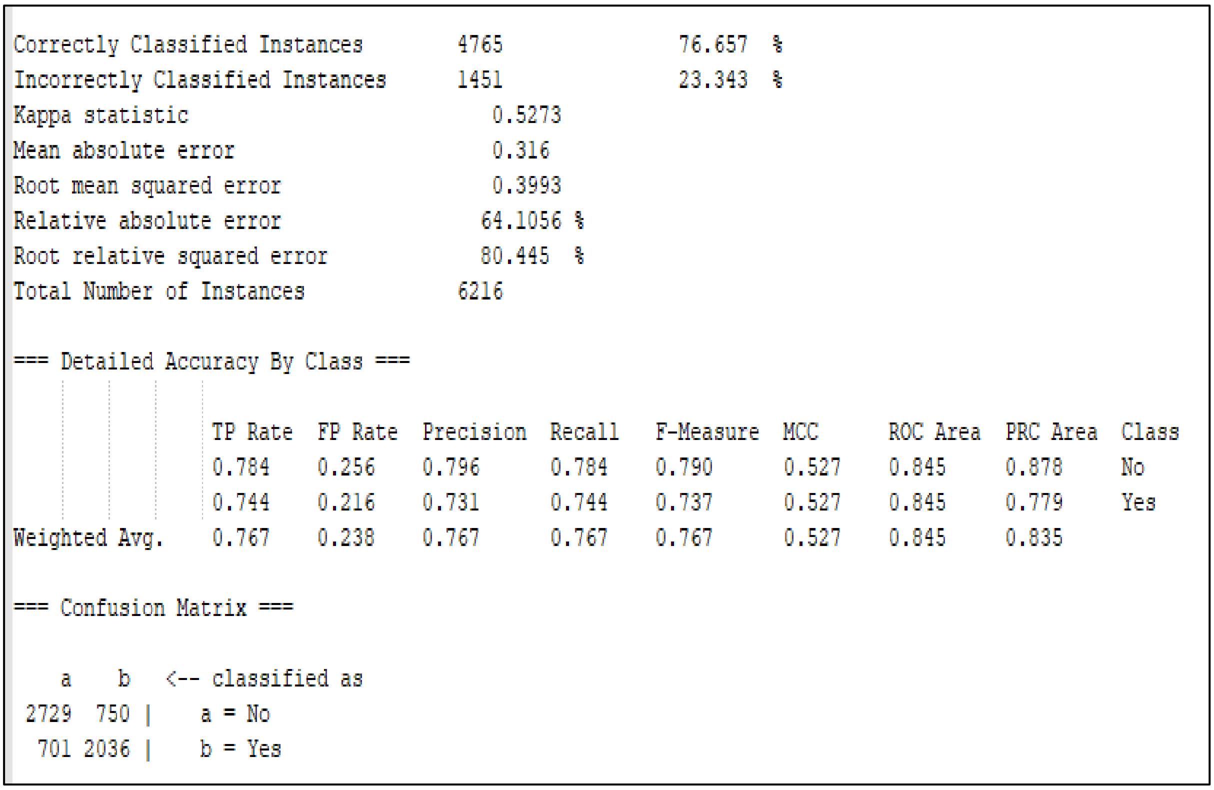
```
Correctly Classified Instances        5426               87.2909 %
Incorrectly Classified Instances       790               12.7091 %
Kappa statistic                          0.743
Mean absolute error                      0.2054
Root mean squared error                  0.3132
Relative absolute error                 41.6639 %
Root relative squared error             63.0892 %
Total Number of Instances             6216

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.875    0.129    0.896      0.875    0.885      0.743    0.934     0.950     No
                 0.871    0.125    0.845      0.871    0.858      0.743    0.934     0.900     Yes
Weighted Avg.    0.873    0.128    0.874      0.873    0.873      0.743    0.934     0.928

=== Confusion Matrix ===

    a    b   <-- classified as
 3043  436 |    a = No
  354 2383 |    b = Yes
```

J48 Tree



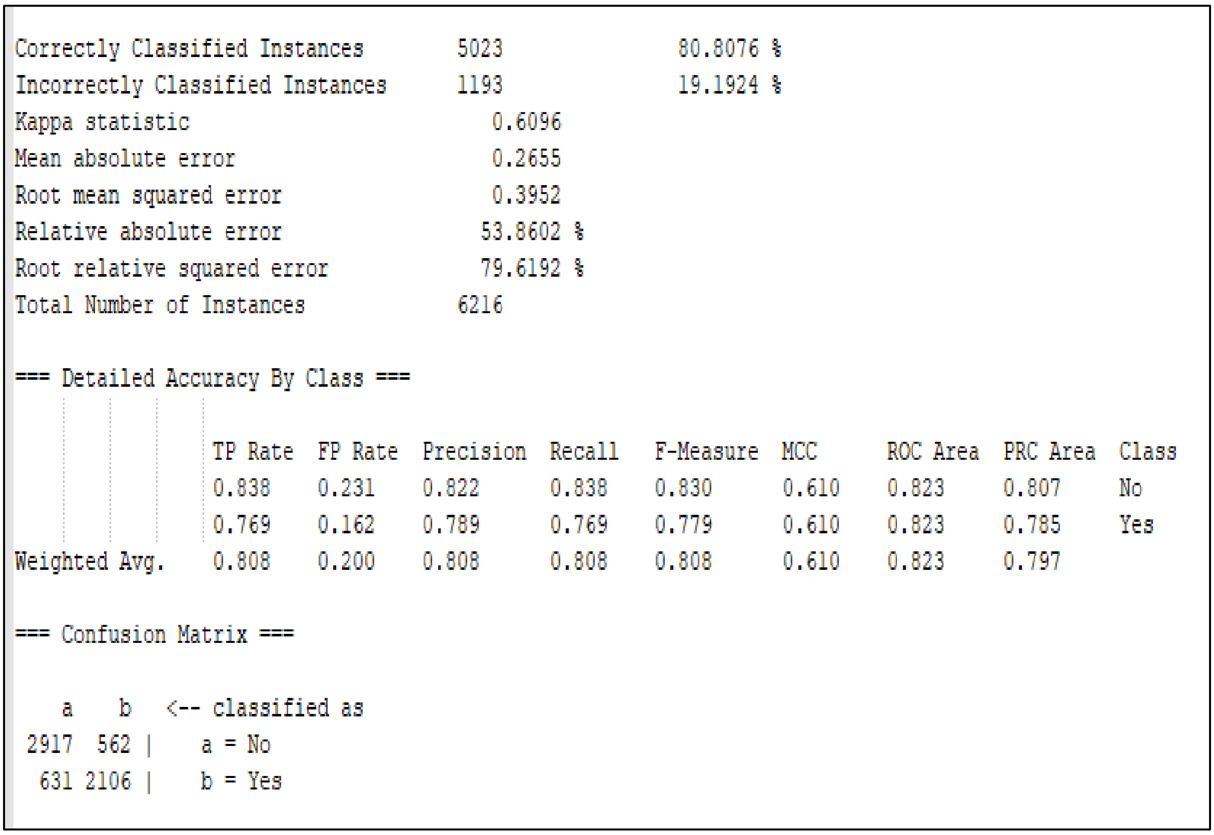
```
Correctly Classified Instances        5112               82.2394 %
Incorrectly Classified Instances      1104               17.7606 %
Kappa statistic                          0.6412
Mean absolute error                      0.1972
Root mean squared error                  0.4019
Relative absolute error                 40.0024 %
Root relative squared error             80.9657 %
Total Number of Instances             6216

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.826    0.182    0.852      0.826    0.839      0.642    0.828     0.823     No
                 0.818    0.174    0.787      0.818    0.802      0.642    0.828     0.738     Yes
Weighted Avg.    0.822    0.179    0.824      0.822    0.823      0.642    0.828     0.786

=== Confusion Matrix ===

    a    b   <-- classified as
 2873  606 |    a = No
  498 2239 |    b = Yes
```

Function Logistic Classifier



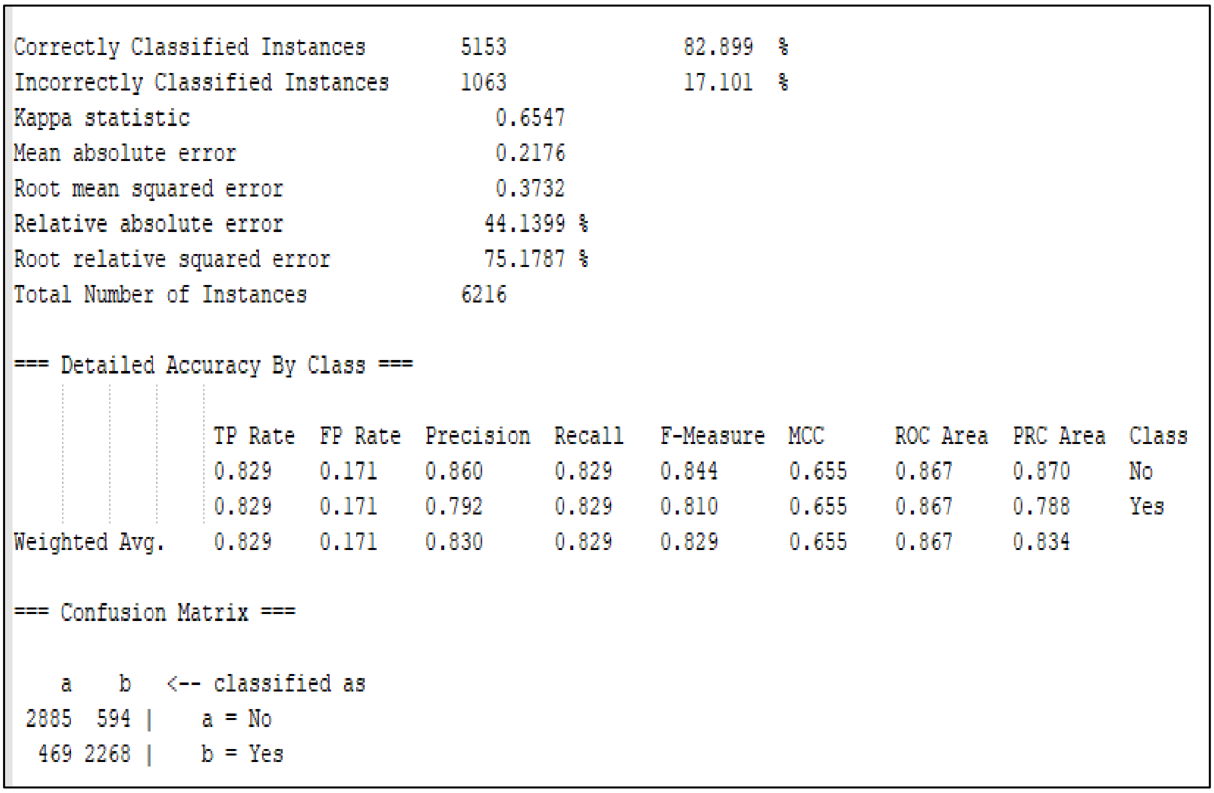
```
Correctly Classified Instances        4765               76.657  %
Incorrectly Classified Instances      1451               23.343  %
Kappa statistic                          0.5273
Mean absolute error                      0.316
Root mean squared error                  0.3993
Relative absolute error                 64.1056 %
Root relative squared error             80.445  %
Total Number of Instances             6216

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.784    0.256    0.796      0.784    0.790      0.527    0.845     0.878     No
                 0.744    0.216    0.731      0.744    0.737      0.527    0.845     0.779     Yes
Weighted Avg.    0.767    0.238    0.767      0.767    0.767      0.527    0.845     0.835

=== Confusion Matrix ===

    a    b   <-- classified as
 2729  750 |    a = No
  701 2036 |    b = Yes
```

JRip Classifier



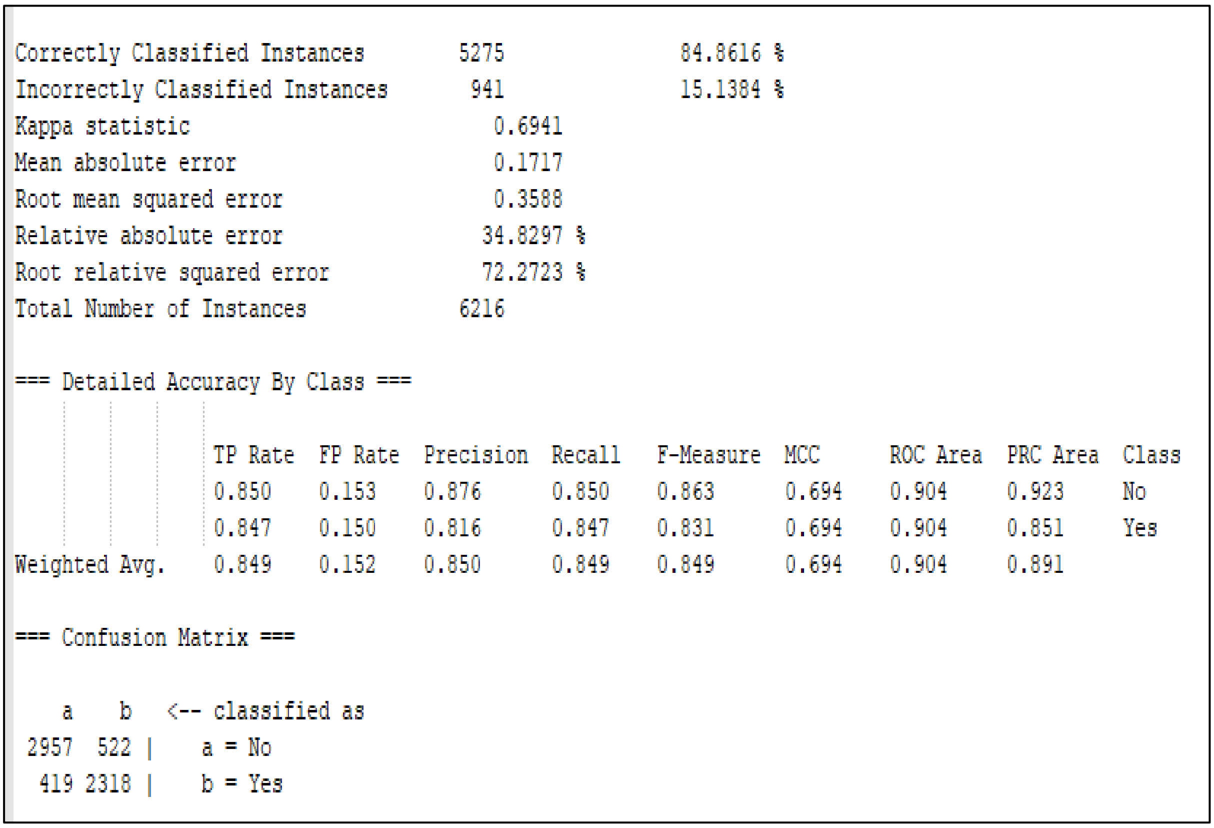
```
Correctly Classified Instances        5023               80.8076 %
Incorrectly Classified Instances      1193               19.1924 %
Kappa statistic                          0.6096
Mean absolute error                      0.2655
Root mean squared error                  0.3952
Relative absolute error                 53.8602 %
Root relative squared error             79.6192 %
Total Number of Instances             6216

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.838    0.231    0.822      0.838    0.830      0.610    0.823     0.807     No
                 0.769    0.162    0.789      0.769    0.779      0.610    0.823     0.785     Yes
Weighted Avg.    0.808    0.200    0.808      0.808    0.808      0.610    0.823     0.797

=== Confusion Matrix ===

    a    b   <-- classified as
 2917  562 |    a = No
  631 2106 |    b = Yes
```

Filtered Classifier



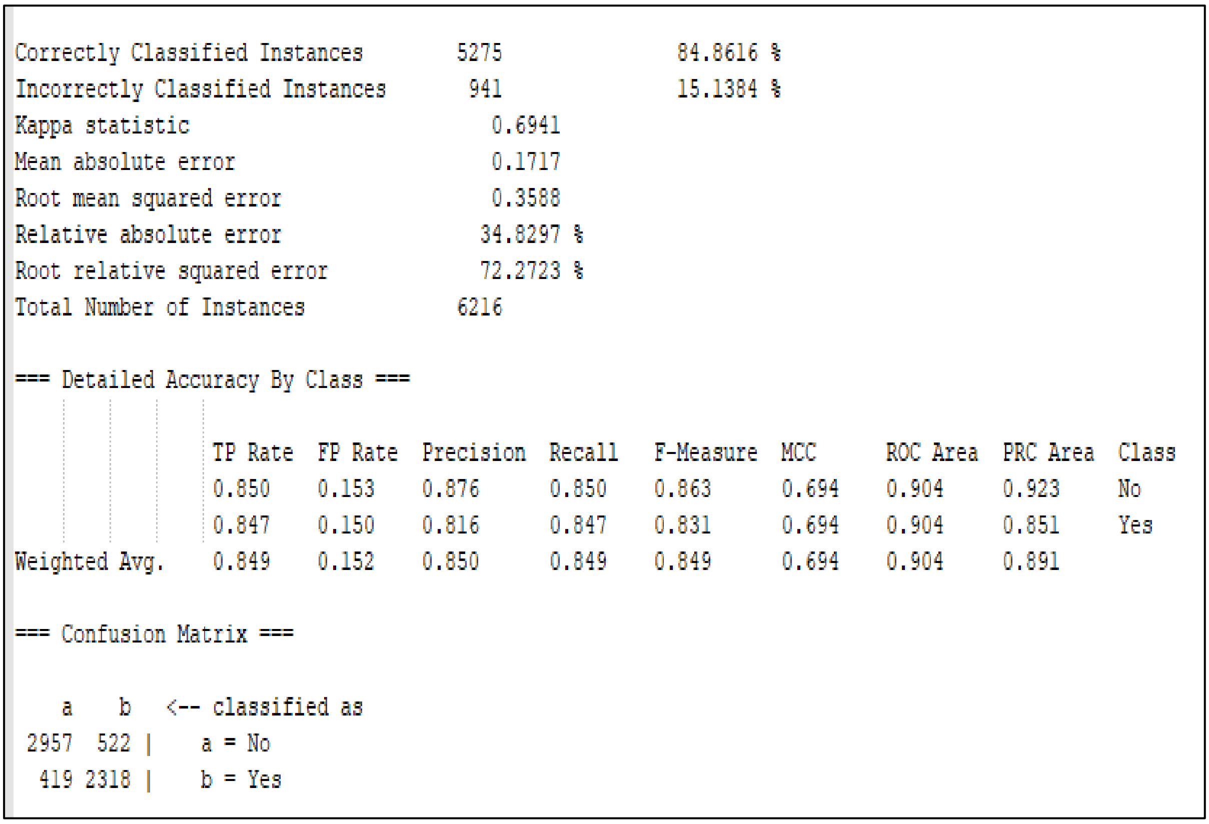
```
Correctly Classified Instances        5153               82.899  %
Incorrectly Classified Instances      1063               17.101  %
Kappa statistic                          0.6547
Mean absolute error                      0.2176
Root mean squared error                  0.3732
Relative absolute error                 44.1399 %
Root relative squared error             75.1787 %
Total Number of Instances             6216

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.829    0.171    0.860      0.829    0.844      0.655    0.867     0.870     No
                 0.829    0.171    0.792      0.829    0.810      0.655    0.867     0.788     Yes
Weighted Avg.    0.829    0.171    0.830      0.829    0.829      0.655    0.867     0.834

=== Confusion Matrix ===

    a    b   <-- classified as
 2885  594 |    a = No
  469 2268 |    b = Yes
```

Classification via Regression



```
Correctly Classified Instances        5275               84.8616 %
Incorrectly Classified Instances       941               15.1384 %
Kappa statistic                          0.6941
Mean absolute error                      0.1717
Root mean squared error                  0.3588
Relative absolute error                 34.8297 %
Root relative squared error             72.2723 %
Total Number of Instances             6216

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.850    0.153    0.876      0.850    0.863      0.694    0.904     0.923     No
                 0.847    0.150    0.816      0.847    0.831      0.694    0.904     0.851     Yes
Weighted Avg.    0.849    0.152    0.850      0.849    0.849      0.694    0.904     0.891

=== Confusion Matrix ===

    a    b   <-- classified as
 2957  522 |    a = No
  419 2318 |    b = Yes
```

Multilayer Perceptron



```
Correctly Classified Instances        5275               84.8616 %
Incorrectly Classified Instances       941               15.1384 %
Kappa statistic                          0.6941
Mean absolute error                      0.1717
Root mean squared error                  0.3588
Relative absolute error                 34.8297 %
Root relative squared error             72.2723 %
Total Number of Instances             6216

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.850    0.153    0.876      0.850    0.863      0.694    0.904     0.923     No
                 0.847    0.150    0.816      0.847    0.831      0.694    0.904     0.851     Yes
Weighted Avg.    0.849    0.152    0.850      0.849    0.849      0.694    0.904     0.891

=== Confusion Matrix ===

    a    b   <-- classified as
 2957  522 |    a = No
  419 2318 |    b = Yes
```

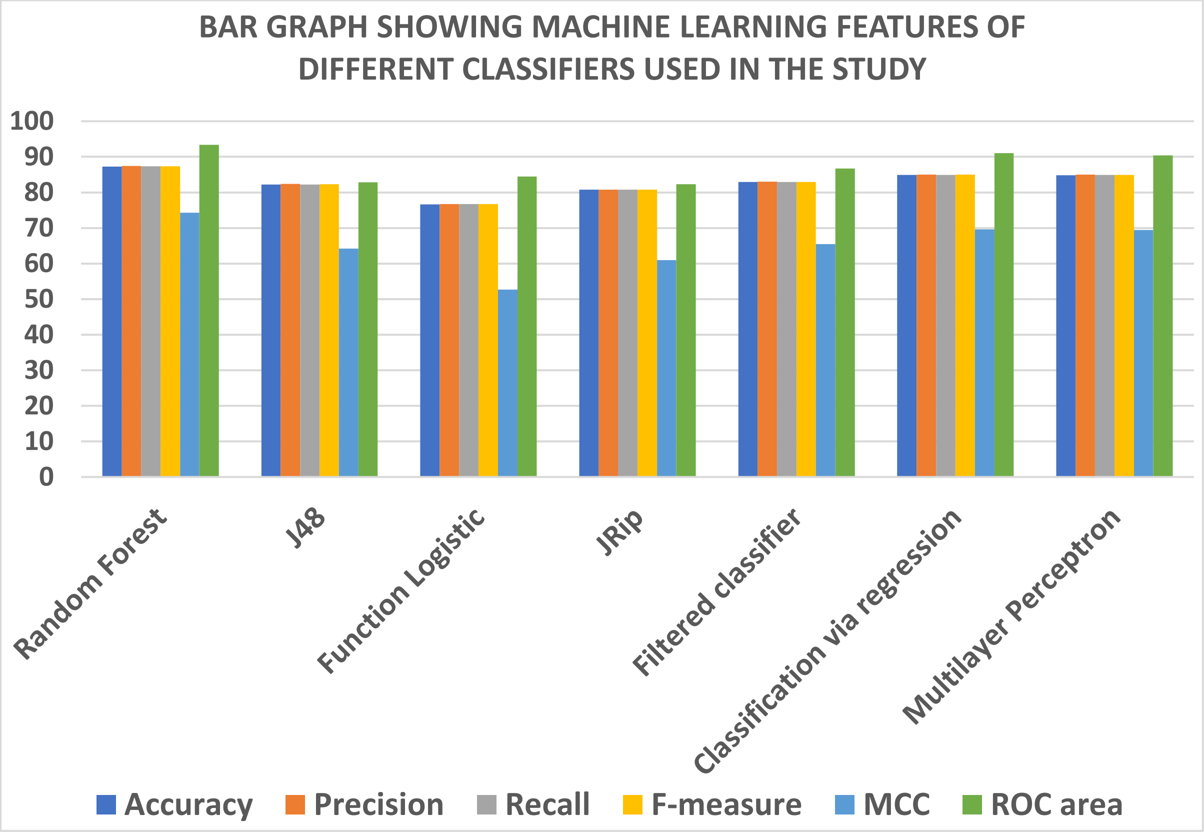


Figure 12: Bar graph showing performance of different classifiers used in the study

Table 1: Results obtained for various Machine Learning Algorithms used in this study.

Classifier	***Accuracy***	***Precision***	***Recall***	***F-measure***	***MCC***	***ROC area***
Random Forest	87.29	87.40	87.30	87.30	74.30	93.40
J48	82.23	82.40	82.20	82.30	64.20	82.80
Function Logistic	76.65	76.70	76.70	76.70	52.70	84.50
JRip	80.80	80.80	80.80	80.80	61.00	82.30
Filtered classifier	82.89	83.00	82.90	82.90	65.50	86.70
Classification via regression	84.92	85.00	84.90	85.00	69.60	91.00
Multilayer Perceptron	84.86	85.00	84.90	84.90	69.40	90.40

Classification via Regression and Multilayer perceptron have also showed high accuracy and precision in predicting association of SAVs with skin disease and performed well in terms of area under the ROC curve. But Random forest classifier of machine learning was chosen for testing the unclassified data because of its high accuracy (87.29%) and precision (87.40%), moreover its gives better performance among other classifiers in terms of area under the ROC curve (93.40%) in predicting the disease association of SAVs with skin disease.

4.3 Training of Random Forest classifier on different types of datasets

Random forest classifier was used to predict association of unclassified SAVs with skin diseases. Data retrieved from PROFEAT, TANGO, WALTZ, LIMBO and Grantham were individually tested to check their individual performance for comparison purpose with SAVDerma which compiles the following attributes related to SAVs.

Table 2: Attributes of SAVs that are compiled in SAVDerma

Property	Class 1	Class 2	Class 3
Hydrophobicity	Polar RKEDQN	Neutral GASTPHY	Hydrophobicity CLVIMFW
Normalized van der Waals volume	0-2.78 GASTPD	2.95-4.0 NVEQIL	4.03-8.08 MHKFRYW
Polarity	4.9-6.2 LIFWCMVY	8.0-9.2 PATGS	10.4-13.0 HQRKNED
Polarizability	0-1.08 GASDT	0.128-0.186 CPNVEQIL	0.219-0.409 KMHFRYW
Charge	Positive KR	Neutral ANCQGHILMFPSTWYV	Negative DE
Secondary structure	Helix EALMQKRH	Strand VIYCWFT	Coil GNPSD
Solvent accessibility	Buried ALFCGIVW	Exposed PKQEND	Intermediate MPSTHY
Surface tension	-0.20~0.16 GQDNAHR	-0.3~ -0.52 KTSEC	-0.98~-2.46 ILMFPWYV
Molecular Weight	Low (75-105) AGS	Medium (115-155) CDEHIKLMNQPTV	High (165-204) FRWY

cLogP	-4.2 - -3.3 RKDNEQH	-3.07 – 2.26 PYSTGACV	-1.78 - -1.05 WMFLI
Solubility in water	High (9-65 g/100g) ACGKRT	Medium (1.14-7.44 g/100g) EFHILMNPQSVW	Low (0.048-0.82 g/100g) DY
Amino acid flexibility index	Very flexible EGKNQS	Moderately flexible ADHIPRTV	Less flexible CFLMWY
TANGO score	Mutation increase aggregation propensity (>50)	Neutral (between -50 and 50)	Mutation decrease aggregation propensity (<-50)
WALTZ score	mutations can increase amyloid propensity (>50)	Neutral (between -50 and 50)	mutations can decrease amyloid propensity (<-50)
LIMBO score	mutations can increase chaperone binding (>50)	Neutral (between -50 and 50)	Mutation decrease chaperone binding (<-50)
Grantham score	Conservative (0-50)	Moderately conservative (51-100), Moderately radical (101-150)	Radical (≥151)

Individual performance of various databases and SAVDerma used in the study was tested using Random forest classifier and performance results were compiled as shown in the Table 3. Tool based on sequence properties, i.e., PROFEAT gives better performance in terms of accuracy and precision in comparison with other mutational effects-based tools i.e., TANGO, WALTZ, LIMBO and Grantham score. But SAVDerma performs very well and comes out to be the best performer in overall performance measures. Performance table shows the highest accuracy of SAVDerma i.e., 87.29 percent and precision of 87.40 percent, ROC curve which depicts the quality of binary measure comes out to be 93.40 percent, which proves that SAVDerma compiles data of high quality.

Table 3: Performance of various databases and SAVDerma

Tools	Accuracy	Precision	Recall	F-Measure	MCC	ROC curve
Profeat	81.59	81.60	81.60	81.60	62.70	81.40
Tango	57.72	56.70	57.70	53.80	10.50	55.70
Waltz	57.23	56.40	57.20	50.20	08.40	52.00
Limbo	56.74	55.30	56.70	51.10	07.40	51.90
Granthum	62.00	61.60	62.00	60.60	21.20	64.70
SAVDerma	87.29	87.40	87.30	87.30	74.30	93.40

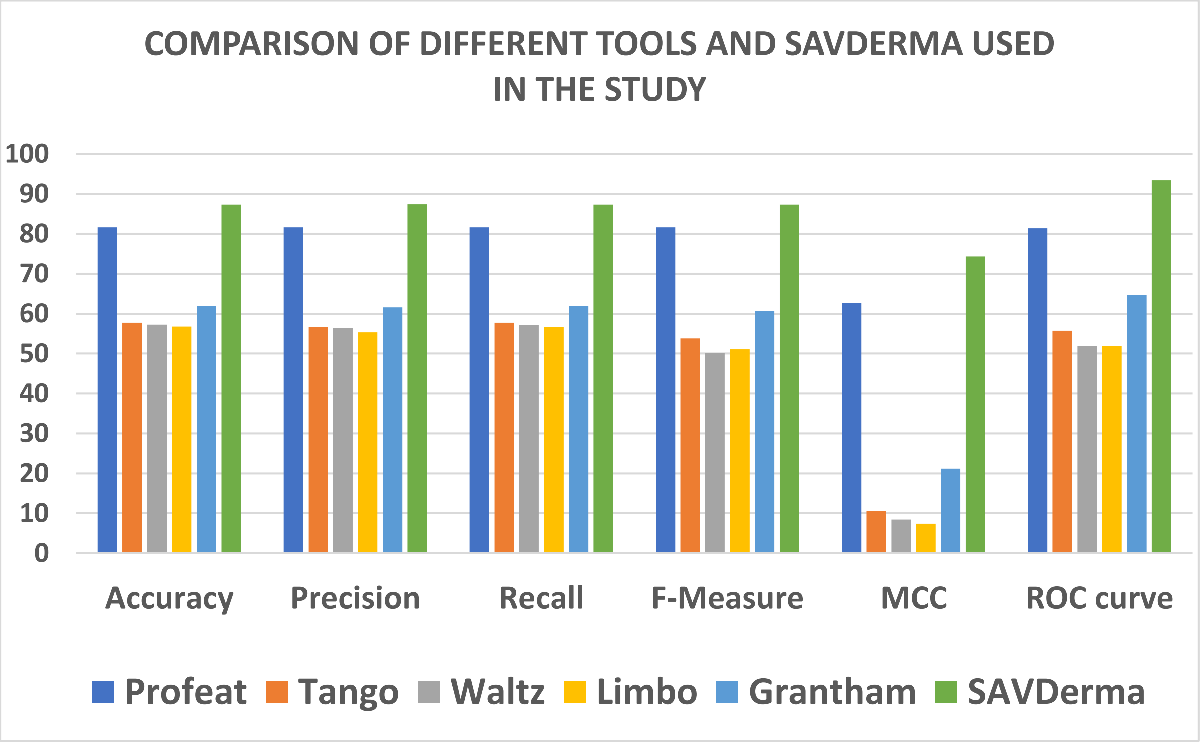


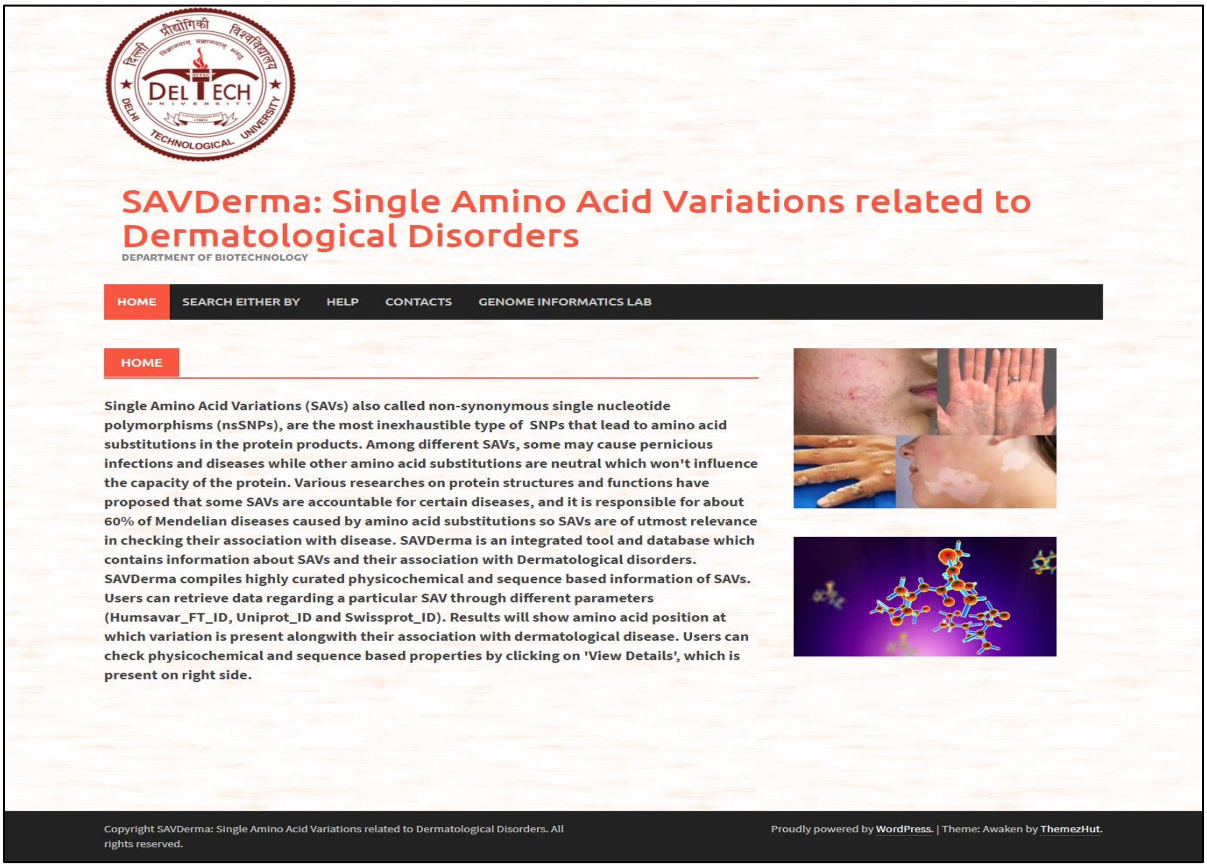
Figure 13: Bar graph showing performance of different tools and SAVderma

On the basis of prediction result and information retrieved from various databases, SAVDerma was developed using WordPress and Caspio program which provides easy to use, no coding platform for application development. WordPress is an open source, Content Management System (CMS), which was made to oversee online sites and blogs.

4.4 User interface development for SAVDerma

URL: http://savderma.info/

SAVDerma is an integrated tool and database which contains information about SAVs and their association with Dermatological disorders. SAVDerma compiles highly curated physicochemical and sequence-based information of SAVs. Users can retrieve data regarding a particular SAV through different parameters (Humsavar_FT_ID, Uniprot_ID and Swissprot_ID). Results will show amino acid position at which variation is present alongwith their association with dermatological disease.



SAVDerma: Single Amino Acid Variations related to Dermatological Disorders

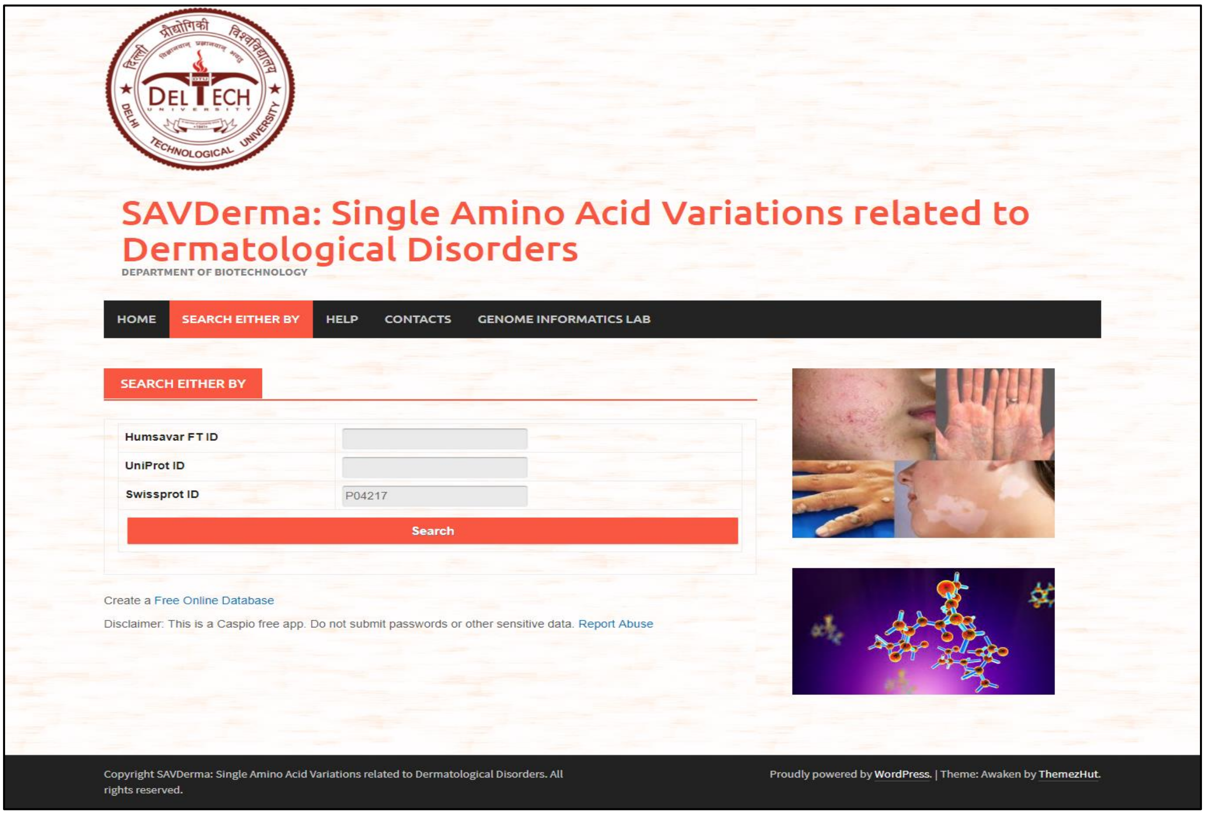
DEPARTMENT OF BIOTECHNOLOGY

Single Amino Acid Variations (SAVs) also called non-synonymous single nucleotide polymorphisms (nsSNPs), are the most inexhaustible type of SNPs that lead to amino acid substitutions in the protein products. Among different SAVs, some may cause pernicious infections and diseases while other amino acid substitutions are neutral which won't influence the capacity of the protein. Various researches on protein structures and functions have proposed that some SAVs are accountable for certain diseases, and it is responsible for about 60% of Mendelian diseases caused by amino acid substitutions so SAVs are of utmost relevance in checking their association with disease. SAVDerma is an integrated tool and database which contains information about SAVs and their association with Dermatological disorders. SAVDerma compiles highly curated physicochemical and sequence based information of SAVs. Users can retrieve data regarding a particular SAV through different parameters (Humsavar_FT_ID, Uniprot_ID and Swissprot_ID). Results will show amino acid position at which variation is present alongwith their association with dermatological disease. Users can check physicochemical and sequence based properties by clicking on 'View Details', which is present on right side.

Copyright SAVDerma: Single Amino Acid Variations related to Dermatological Disorders. All rights reserved.

Proudly powered by WordPress. | Theme: Awaken by ThemezHut.

Figure 14: Homepage of SAVDerma



SAVDerma: Single Amino Acid Variations related to Dermatological Disorders

DEPARTMENT OF BIOTECHNOLOGY

Humsavar FT ID	
UniProt ID	
Swissprot ID	P04217

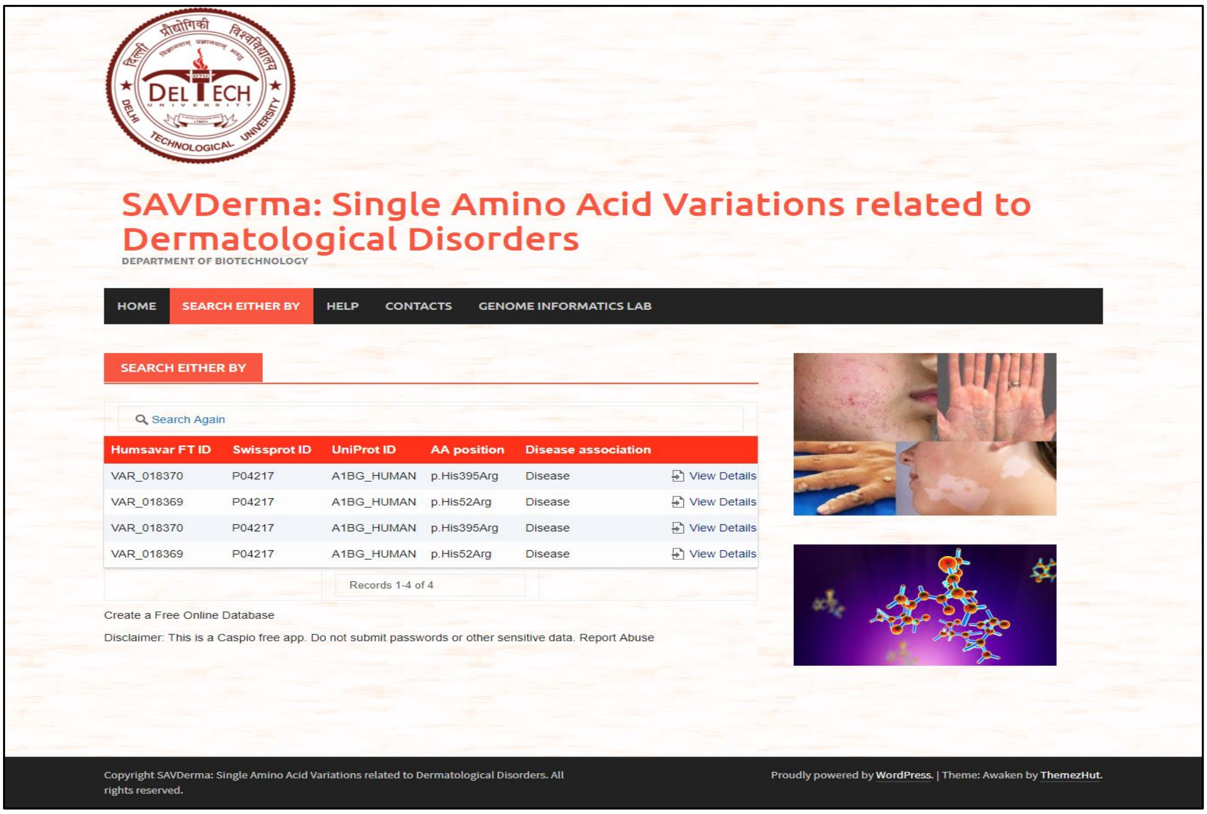
Create a Free Online Database

Disclaimer: This is a Caspio free app. Do not submit passwords or other sensitive data. Report Abuse

Copyright SAVDerma: Single Amino Acid Variations related to Dermatological Disorders. All rights reserved.

Proudly powered by WordPress. | Theme: Awaken by ThemezHut.

Figure 15: Search page of SAVDerma



SAVDerma: Single Amino Acid Variations related to Dermatological Disorders

DEPARTMENT OF BIOTECHNOLOGY

Humsavar FT ID	Swissprot ID	UniProt ID	AA position	Disease association	
VAR_018370	P04217	A1BG_HUMAN	p.His395Arg	Disease	View Details
VAR_018369	P04217	A1BG_HUMAN	p.His52Arg	Disease	View Details
VAR_018370	P04217	A1BG_HUMAN	p.His395Arg	Disease	View Details
VAR_018369	P04217	A1BG_HUMAN	p.His52Arg	Disease	View Details

Records 1-4 of 4

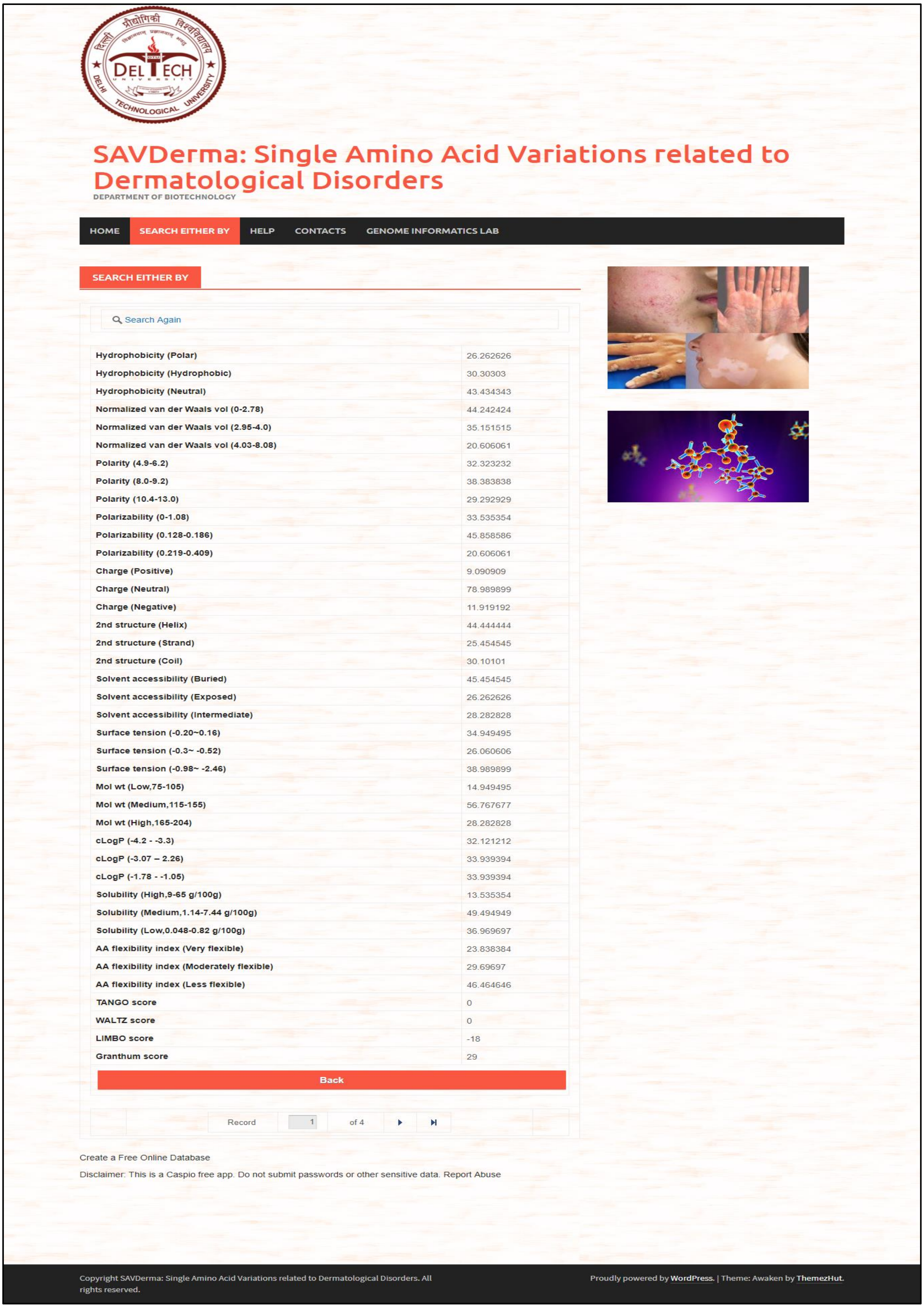
Create a Free Online Database

Disclaimer: This is a Caspio free app. Do not submit passwords or other sensitive data. Report Abuse

Copyright SAVDerma: Single Amino Acid Variations related to Dermatological Disorders. All rights reserved.

Proudly powered by WordPress. | Theme: Awaken by ThemezHut.

Figure 16: Result page of SAVDerma showing position of SAV alongwith its association with Dermatological disorders



SAVDerma: Single Amino Acid Variations related to Dermatological Disorders
DEPARTMENT OF BIOTECHNOLOGY
HOME
SEARCH EITHER BY
HELP
CONTACTS
GENOME INFORMATICS LAB
SEARCH EITHER BY
Search Again

Hydrophobicity (Polar)	26.262626
Hydrophobicity (Hydrophobic)	30.30303
Hydrophobicity (Neutral)	43.434343
Normalized van der Waals vol (0-2.78)	44.242424
Normalized van der Waals vol (2.95-4.0)	35.151515
Normalized van der Waals vol (4.03-8.08)	20.606061
Polarity (4.9-6.2)	32.323232
Polarity (8.0-9.2)	38.383838
Polarity (10.4-13.0)	29.292929
Polarizability (0-1.08)	33.535354
Polarizability (0.128-0.186)	45.858586
Polarizability (0.219-0.409)	20.606061
Charge (Positive)	9.090909
Charge (Neutral)	78.989899
Charge (Negative)	11.919192
2nd structure (Helix)	44.444444
2nd structure (Strand)	25.454545
2nd structure (Coil)	30.10101
Solvent accessibility (Buried)	45.454545
Solvent accessibility (Exposed)	26.262626
Solvent accessibility (Intermediate)	28.282828
Surface tension (-0.20~0.16)	34.949495
Surface tension (-0.3~ -0.52)	26.060606
Surface tension (-0.98~ -2.46)	38.989899
Mol wt (Low,75-105)	14.949495
Mol wt (Medium,115-155)	56.767677
Mol wt (High,165-204)	28.282828
cLogP (-4.2 - -3.3)	32.121212
cLogP (-3.07 – 2.26)	33.939394
cLogP (-1.78 - -1.05)	33.939394
Solubility (High,9-65 g/100g)	13.535354
Solubility (Medium,1.14-7.44 g/100g)	49.494949
Solubility (Low,0.048-0.82 g/100g)	36.969697
AA flexibility index (Very flexible)	23.838384
AA flexibility index (Moderately flexible)	29.69697
AA flexibility index (Less flexible)	46.464646
TANGO score	0
WALTZ score	0
LIMBO score	-18
Granthum score	29

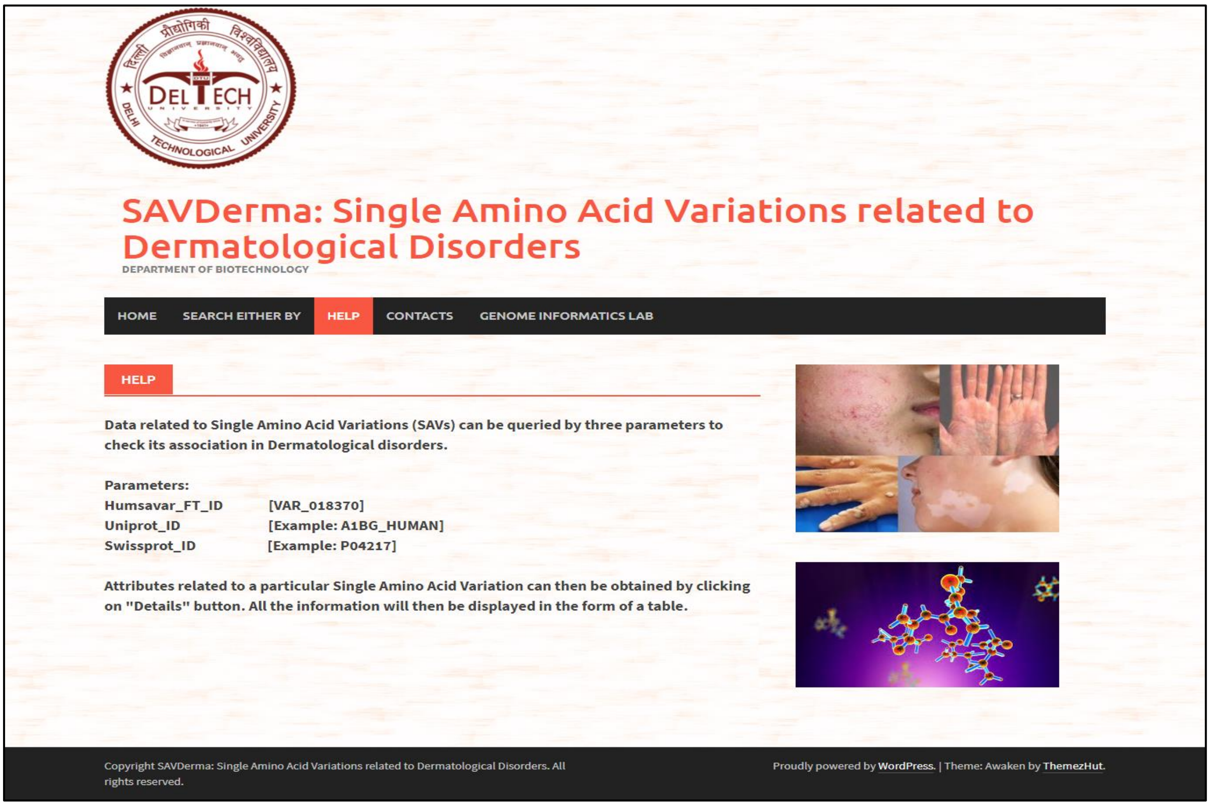
Create a Free Online Database

Disclaimer: This is a Caspio free app. Do not submit passwords or other sensitive data. Report Abuse

Copyright SAVDerma: Single Amino Acid Variations related to Dermatological Disorders. All rights reserved.

Proudly powered by WordPress. | Theme: Awaken by ThemezHut.

Figure 17: Result page containing information of a particular SAV which will be displayed upon clicking on 'View Details' option



SAVDerma: Single Amino Acid Variations related to Dermatological Disorders

DEPARTMENT OF BIOTECHNOLOGY

HELP

Data related to Single Amino Acid Variations (SAVs) can be queried by three parameters to check its association in Dermatological disorders.

Parameters:

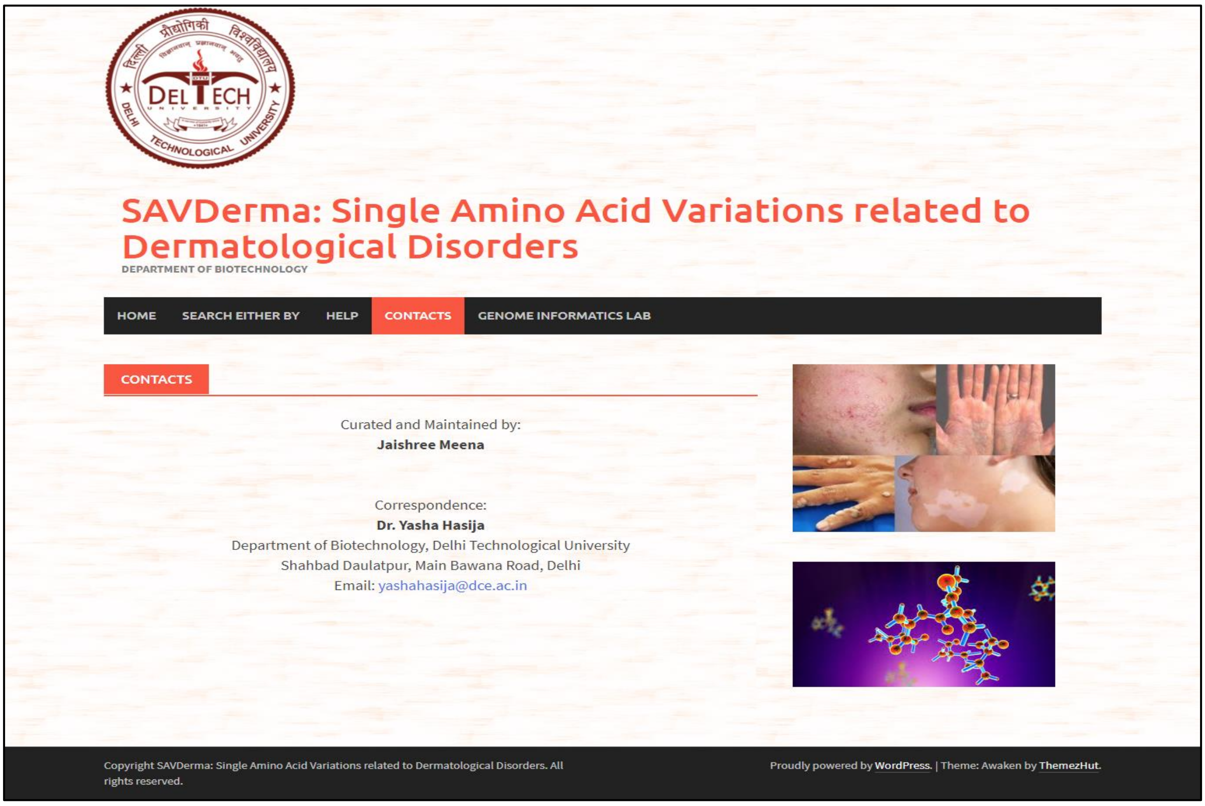
Humsavar_FT_ID	[VAR_018370]
Uniprot_ID	[Example: A1BG_HUMAN]
Swissprot_ID	[Example: P04217]

Attributes related to a particular Single Amino Acid Variation can then be obtained by clicking on "Details" button. All the information will then be displayed in the form of a table.

Copyright SAVDerma: Single Amino Acid Variations related to Dermatological Disorders. All rights reserved.

Proudly powered by WordPress. | Theme: Awaken by ThemezHut.

Figure 18: Help page of SAVDerma containing information about how to retrieve data from the database



SAVDerma: Single Amino Acid Variations related to Dermatological Disorders

DEPARTMENT OF BIOTECHNOLOGY

CONTACTS

Curated and Maintained by:
Jaishree Meena

Correspondence:
Dr. Yasha Hasija
Department of Biotechnology, Delhi Technological University
Shahbad Daulatpur, Main Bawana Road, Delhi
Email: yashahasija@dce.ac.in

Copyright SAVDerma: Single Amino Acid Variations related to Dermatological Disorders. All rights reserved.

Proudly powered by WordPress. | Theme: Awaken by ThemezHut.

Figure 19: Contacts page of SAVDerma

4.5 Validation of SAVDerma

To test the efficacy of SAVDerma, we took various experimentally validated dermatological disorders associated with SAVs which were not present in our training sets by reviewing previous literature. To accomplish the purpose of testing the association of SAVs with dermatological disorders, we used DisGeNET and thoroughly check the query for published literature by providing Swissprot ids. We found various dermatological disorders like Lupus erythematosus, Xeroderma pigmentosum, Melanoma, Hamartoma tumor syndromes, Acrodermatitis enteropathica etc that were caused due to mutations in protein’s amino acid sequence.

Xeroderma pigmentosum is an autosomal recessive disease, where the persons are deficient of highly specific UV-damaged DNA-binding (UV-DDB) activity which leads to UV sensitivity and make the person susceptible to skin cancers. There are two subunits of UV-DDB and mutation in any of these subunits i.e., p125 (DDB1) and p48 (DDB2), leads to Xeroderma pigmentosum progression [28]. Single amino acid substitutions at position 244, replacing lysine with glutamic acid (p.Lys244Glu) in UV-DDB protein sequence results in loss of activity of UV-DDB.

Similarly, we found evidences for Hamartoma tumor syndromes and its association with SAVs. Hamartoma tumor syndromes are a range of hamartomatous outgrowth syndromes which includes Cowden syndrome which is caused due to Single amino acid variation (p.Ile67Arg) at position 67 where isoleucine is replaced by arginine in PTEN (Phosphatase and tensin homolog) protein sequence [29][30]. Patients suffered from Cowden syndrome develop oral papillomas and lesions called trichilemmomas on face, around the mouth and ears which can be severe if untreated.

BRAF is a serine/threonine protein kinase activating the MAP kinase/ERK-signaling pathway. Around 50 % of melanomas harbors activating BRAF mutations [31]. Mutation in BRAF protein sequence at position 600 replacing valine with glutamic acid implicated in different mechanisms underlying melanomagenesis, most of which due to the deregulated activation of the downstream MEK/ERK effectors resulting in progression of melanomas and skin lesions [32].

Acrodermatitis enteropathica is a rare autosomal recessive disorder which is associated with zinc deficiency resulting in retarded growth, loss of hairs, immune dysfunctions and skin lesions [33][34]. This disorder is associated with single amino acid substitution in protein sequence of SLC39A4 at position 84 where proline is replaced with leucine (p.Pro84Leu), position 95 at which arginine is replaced by cysteine (p.Arg95Cys) or position 106 at which asparagine is replaced by lysine (p.Asn106Lys). All the above stated cases are properly predicted by our tools.

Chapter 5 Conclusion and Discussion

Skin is the major organ of human body having a prominent role in thermoregulation, sensation, protection from external environment, metabolism etc. Socio-economic scarcity is the overall everyday reason of a higher risk of skin infections and diseases. People living within socio-economically deprived areas are further apt on the way to exhibit a skin disease than people beginning the slightest deprived areas. In India, due to environmental or personal factors, skin diseases are increasing among humans affecting their lifestyle and social interaction. It penetrates through all cultures, happens at all ages, and influences between thirty to seventy percent of people, with much higher rates in at-risk subpopulations. Extensive proliferation of skin infections and diseases and their associated clinical risks arise the need for elucidation of these skin diseases. Conventional diagnosis of skin diseases requires higher level of expertise and knowledge to recognise the disease from many similar diseases, all these issues related to skin diseases pose a challenge to look for better treatment and diagnosis options and made them available to the patients.

An unparalleled amount of data about SAVs has been produced using genomic profiling technologies. It is evaluated that there are three to five million SAVs in a person as indicated by the sequencing of the entire human genome. SAVs, otherwise called nsSNPs, are the most inexhaustible type of SNPs that lead to amino acid substitutions in the protein products. Among different SAVs, some may cause pernicious infections and diseases while other amino acid substitutions are neutral which won't influence the capacity of the protein. This investigation was aimed at examining the role of SAVs in skin disorders by utilizing machine learning techniques.

By utilising the machine learning techniques, we have made predictions on association of SAVs with dermatological disorders. On basis of the findings, we have made SAVDerma, which is an integrated tool and database which contains information about SAVs and their association with Dermatological disorders. SAVDerma contains more than fifty-seven thousand SAVs and their associated physico-chemical as well sequence-based properties. Many SAVs predicted to be associated with dermatological disorders like Lupus erythematosus, Xeroderma pigmentosum, Psoriasis, Acrodermatitis enteropathica etc, which were not present in our training datasets. SAVDerma is first of its kind repository for the SAVs related to dermatological disorders. It is user friendly-interface providing information of amino acid substitutions. This integrated tool and database will help users to find biologically significant information which will probably going to give insights about the cause of the skin disorders. The data compiled in SAVDerma primarily focusses on giving information about dermatological disease association of amino acid substitutions and their related properties. We intended to update more and more data associated with single amino acid substitutions in future. The data about single amino acid substitutions that we obtained from machine learning which was not known to have any kind of relation with skin diseases can be further used to test their relationship with dermatological diseases and can be targeted to make drugs for better treatment of skin related illnesses. Also, this study will help researchers to find better diagnostic systems based on machine learning for fast and reliable diagnosis of skin disorders.

Chapter 6 References

[1] R. J. Hay *et al.*, “The global burden of skin disease in 2010: An analysis of the prevalence and impact of skin conditions,” *J. Invest. Dermatol.*, vol. 134, no. 6, pp. 1527–1534, 2014.

[2] G. H. Fernald, E. Capriotti, R. Daneshjou, K. J. Karczewski, and R. B. Altman, “Bioinformatics challenges for personalized medicine,” *Bioinformatics*. 2011.

[3] M. Wang *et al.*, “FunSAV: Predicting the Functional Effect of Single Amino Acid Variants Using a Two-Stage Random Forest Model,” *PLoS One*, vol. 7, no. 8, p. e43847, 2012.

[4] R. P. J. B. Weller, J. A. A. Hunter, J. A. Savin, and M. V. Dahl, *Clinical Dermatology, Fourth Edition*. 2009.

[5] S. E. Plon *et al.*, “Sequence variant classification and reporting: recommendations for improving the interpretation of cancer susceptibility genetic test results,” *Human Mutation*. 2008.

[6] Z. Zhang, M. A. Miteva, L. Wang, and E. Alexov, “Analyzing effects of naturally occurring missense mutations,” *Computational and Mathematical Methods in Medicine*. 2012.

[7] E. Alexov and M. Sternberg, “Understanding molecular effects of naturally occurring genetic differences,” *Journal of Molecular Biology*. 2013.

[8] Y. Yang and Y. Zhou, “Specific interactions for ab initio folding of protein terminal regions with secondary structures,” *Proteins Struct. Funct. Genet.*, 2008.

[9] T. G. Kucukkal, Y. Yang, S. C. Chapman, W. Cao, and E. Alexov, “Computational and experimental approaches to reveal the effects of single nucleotide polymorphisms with respect to disease diagnostics,” *International Journal of Molecular Sciences*. 2014.

[10] M. Petukh, T. G. Kucukkal, and E. Alexov, “On human disease-causing amino acid variants: Statistical study of sequence and structural patterns,” *Hum. Mutat.*, 2015.

[11] L. Boccuto *et al.*, “A mutation in a ganglioside biosynthetic enzyme, ST3GAL5, results in salt & pepper syndrome, a neurocutaneous disorder with altered glycolipid and glycoprotein glycosylation,” *Hum. Mol. Genet.*, 2014.

[12] N. Dolzhanskaya *et al.*, “A novel p.leu(381)phe mutation in presenilin 1 is associated with very early onset and unusually fast progressing dementia as well as lysosomal inclusions typically seen in kufs disease,” *J. Alzheimer’s Dis.*, 2014.

[13] K. Takano *et al.*, “An x-linked channelopathy with cardiomegaly due to a CLIC2 mutation enhancing ryanodine receptor channel activity,” *Hum. Mol. Genet.*, 2012.

[14] H. B. Rao, F. Zhu, G. B. Yang, Z. R. Li, and Y. Z. Chen, “Update of PROFEAT: A web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence,” *Nucleic Acids Res.*, vol. 39, no. SUPPL. 2, pp. 385–390, 2011.

[15] G. De Baets, L. Van Doorn, F. Rousseau, and J. Schymkowitz, “Increased Aggregation Is More Frequently Associated to Human Disease-Associated Mutations Than to Neutral Polymorphisms,” *PLoS Comput. Biol.*, 2015.

[16] A. M. Fernandez-Escamilla, F. Rousseau, J. Schymkowitz, and L. Serrano, “Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins,” *Nat. Biotechnol.*, 2004.

[17] J. Van Durme, S. Maurer-Stroh, R. Gallardo, H. Wilkinson, F. Rousseau, and J. Schymkowitz, “Accurate prediction of DnaK-peptide binding via homology modelling and experimental data,” *PLoS Comput Biol*, 2009.

[18] R. Grantham, “Amino acid difference formula to help explain protein evolution,” *Science (80-.).*, 1974.

[19] C. M. Bishop, *Pattern Recognition and Machine Learning*. 2006.

[20] B. T. Femina and E. M. Sudheep, “An efficient CRM-data mining framework for the prediction of customer behaviour,” in *Procedia Computer Science*, 2015.

[21] L. Sabin-Wilson, “WordPress for Dummies,” *Director*, 2008.

[22] I. Srivastava, L. K. Gahlot, P. Khurana, and Y. Hasija, “DbAARD & AGP: A computational pipeline for the prediction of genes associated with age related disorders,” *J. Biomed. Inform.*, vol. 60, pp. 153–161, 2016.

[23] S. Dreiseitl, L. Ohno-Machado, H. Kittler, S. Vinterbo, H. Billhardt, and M. Binder, “A Comparison of Machine Learning Methods for the Diagnosis of Pigmented Skin

Lesions," *J. Biomed. Inform.*, vol. 34, no. 1, pp. 28–36, 2001.

[24] T. C. Sharma and M. Jain, "WEKA Approach for Comparative Study of Classification Algorithm," *Int. J. Adv. Res. Comput. Commun. Eng.*, vol. 2, no. 4, pp. 1925–1931, 2013.

Chapter 7 Annexures

Annexure 1. Python code for retrieving sequence of proteins having single amino acid substitution. Protein sequences having SAVs were retrieved by changing the particular amino acid residue at respective position as indicated by HumsaVar information.

--------------------------------------*--*------------------------------------

```
import urllib.request as ur

def deleteContent(pfile):

    pfile.seek(0)

    pfile.truncate()




def seq_pos_replace(seq,y,aa_a,aa_b):

        lista=list(seq)

        if seq[y-1] is aa_a:

            lista[y-1]= aa_b

            seq=".join(lista)


        else:

            print('wrong aa or position input ')

            n=str(input("enter 'y' to retry or enter 'x' to exit: "))


        #print (seq)

        return(seq)
```

```
result=open("result.txt","w+")

deleteContent(result)


snp_file=open('snp.txt')

snp=snp_file.read()

snp_list=snp.split()

pos_file=open('pos.txt')

pos=pos_file.read()

pos_list=pos.split()

uniprot_id_file=open('uniprot.txt')

uni=uniprot_id_file.read()

uniprot_id=uni.split()

count=0



#seq = input("Enter the sequence: ")

d = {'CYS': 'C', 'ASP': 'D', 'SER': 'S', 'GLN': 'Q', 'LYS': 'K',

     'ILE': 'I', 'PRO': 'P', 'THR': 'T', 'PHE': 'F', 'ASN': 'N',

     'GLY': 'G', 'HIS': 'H', 'LEU': 'L', 'ARG': 'R', 'TRP': 'W',

     'ALA': 'A', 'VAL':'V', 'GLU': 'E', 'TYR': 'Y', 'MET': 'M'}

pointer="

uniprot_id_a=""

if len(snp_list)==len(pos_list):

    for i in range(len(snp_list)):

        ra=uniprot_id[i]

        if uniprot_id_a != ra:
```

```
    uniprot_id_a=ra
    f=open("text.txt",'w')
    deleteContent(f)



    url=('http://www.uniprot.org/uniprot/'+uniprot_id_a+'.fasta')
    s = ur.urlopen(url)
    sl = str(s.read())
    sl=sl.split("\\n")
    count=count+1
    print(count)
    #print (sl)
    for line in sl:

        if line==sl[0] or line==sl[-1]:
            continue

        else:
            f.write(line)

            f.write('\n')
else:

    print('ya')
f.close()
fasta_file=open('text.txt')
```

```
        seq=('')
        for a in fasta_file:
            a=a.rstrip()


            seq=seq+a
        result.write('>'+str(snp_list[i]))
        pointer=pos_list[i]
        amino_a=str(pointer[2:5])
        amino_b=str(pointer[-3:])
        aa_a=d[amino_a.upper()]
        #print(aa_a)
        aa_b=d[amino_b.upper()]
        l=len(pos_list[i])
        r=l-8
        y=int(pointer[5:5+r])
        result.write("\n")
        #print(seq_pos_replace(seq,y,aa_a,aa_b))
        result.write(seq_pos_replace(seq,y,aa_a,aa_b))
        result.write("\n")


else:


    print("No. of SNPids are not equal to no. of positions")
    result.write("No. of SNPids are not equal to no. of positions")
print('Done')
fasta_file.close()
```

```
snp_file.close()

pos_file.close()

result.close()

uniprot_id_file.close()
```

--*--*----------------------------------