Analysis of missense mutations in adenosine deaminase using Pathogenic-Or-Not-Pipeline (PON-P)

Master's Thesis
Bioinformatics Master's Degree Programme
Institute of Medical Technology
University of Tampere, Finland
Sreevani Kotha
October 2010

ACKNOWLEDGEMENTS

I wish to express my gratitude to my supervisor, Prof. Mauno Vihinen, Bioinformatics Group Leader, without whose knowledge and assistance this study would not have been successful. Deepest gratitude is also due to the Janita Thusberg for her support and guidance. Special thanks to Jani Härkönen, who was helpful with the usage of latest version of PON-P and Ayodeji E. Olatubosun for formatting tables and editing. My heartily thanks to Martti Tolvanen for his encouragement, guidance and support. Not forgetting my best friends who always been there understanding and showing endless love, through the duration of my studies.

I wish to express my love and gratitude to my parents Mr. and Mrs. Kotha, who instilled in me from an early age the desire and skills to obtain the higher degrees. Finally, I thank God for giving me wisdom and knowledge to finish this project successfully.

October 2010,

Sreevani Kotha,
Bioinformatics Programme,
Institute of Medical Technology,
University of Tampere, Finland.

MASTER'S THESIS

Place	University of Tampere Bioinformatics Master's Degree Programme, Faculty of Medicine, Institute of Medical Technology, Tampere, Finland.
Author	Sreevani Kotha
Title	Analysis of missense mutations in adenosine deaminase using Pathogenic-Or-Not Pipeline
Pages	51 pp + Figures 12 + Tables 11 + Appendices 4pp
Supervisor	Prof. Mauno Vihinen
Reviewers	Prof. Mauno Vihinen + Martti Tolvanen
Time	October, 2010

Abstract

Background: Adenosine deaminase (ADA, E.C.3.5.4.4) is an enzyme that has an important role in immune functions and in the regulation of intracellular and extracellular concentrations of adenosine and adenosine receptor activity. The need of ADA is to breakdown the adenosine from food and also for the turnover of nucleic acids. ADA converts adenosine to inosine. Missense mutations differ from single-nucleotide polymorphism and these are rare things. It is a point mutation in which a single nucleotide is changed, which results in a codon that codes for a different amino acid. Mutations in ADA results in absence or deficiency of the adenosine deaminase enzyme in cells that prevents normal breakdown of deoxyadenosine. A buildup of this toxic compound hinders the development and of lymphocytes, which results in severe combined immunodeficiency.

Aims: The aim of this study is to analyze the effect of missense mutations using the tool Pathogenic-Or-Not Pipeline.

Methods: Different kinds of methods like tolerance predictions, stability change predictions, disorder predictions, aggregation predictions, and sequence conservation analysis methods were used to analyze the effect of missense mutations in ADA.

Results: Results conclude that not even single predictor output matches with those of others. Some of the predictors like Mupro conclude that the missense mutations have no affect on the stability of the protein whereas Cupsat gives the contrary results. Globplot, Iupred, MetaPrDos, PrDos, RONN shows the results in a way that no missense mutations leads to the disorder of the protein.

Conclusion: With single predictor it is not sufficient to achieve predictions good enough to follow the modular organization of a protein. Accuracy and sensitivity of the predictors should be well known. Combination of different predictors can minimize the errors in which PON-P plays a key role.

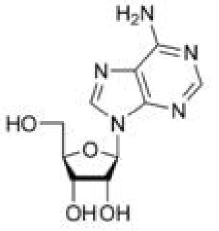
CONTENTS

ABBREVIATIONS

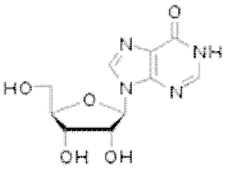
3D	Three-dimensional
AA	Amino acid
ADA	Adenosine deaminase
Aggrescan	For the prediction and evaluation of "hot spots" of aggregation in polypeptides
Amylpred	A web tool for a consensus prediction of amyloidogenic determinants
CASP	Critical Assessment of Techniques for Protein Structure Prediction
CUPSAT	Cologne University Protein Stability Analysis Tool
DisProt	The Database of Disordered Proteins
Drip-pred	Predicts the structurally disordered regions in proteins
FoldIndex	Tool to predict whether a given protein sequence is intrinsically unfolded
FoldUnfold	Web server for the prediction of disordered regions in protein chain.
GlobPlot	Intrinsic Protein Disorder, Domain and Globularity Prediction
IUpred	Web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content.
I-MUTANT	Neural Network based predictor of protein stability changes upon single point mutation from the protein structure.
iPTREE-STAB	Interpretable decision tree based method for predicting protein stability changes upon mutations
metaPRDOS	Prediction of disordered regions in proteins based on the meta approach
nsSNPanalyzer	Pathogenicity predictor
Pasta	Server for protein aggregation prediction
PrDOS	Prediction of disordered protein regions from amino acid sequence
PreLink	Prediction of unfolded segments in a protein sequence based on amino acid composition.
PON-P	Pathogenic-Or-Not-Pipeline
RONN	Regional order neural network software
SCID	Severe combined immunodeficiency
Scide	Identification of stabilization centers in proteins.
Scpred	Accurate prediction of protein structural class for sequences of twilight-zone similarity with predicting sequences
SRide	A server for identifying stabilizing residues in proteins
SIFT	Sorting Intolerant From Tolerant.
SNP	Single nucleotide polymorphism
SVM	Support vector machine
Waltz	An amyloid-prediction tool

1. Introduction

Adenosine deaminase (ADA EC 3.5.4.4) is a cytosolic enzyme, which is involved in purine metabolism, which affects lymphocyte development and function [Hirschhorn, 1999; Hershfield & Mitchell, 2001; Hershfield, 2004]. It has been object of considerable interest, mainly because in human a congenital defect in the enzyme causes severe combined immunodeficiency disease (SCID). ADA deficiency is 20% of all types of SCID [Aldrich *et al.*, 2000]. It is accounted as one of the most severe human immunodeficiencies which is associated with depletion of T cells, B cells, and natural killer cells, consequently resulting in impaired cellular immunity and thereby decreased production of immunoglobulins [Buckley *et al.*, 1997]. Research indicates that the metabolic basis for ADA-deficient immunodeficiency is very much related to the effect of its substrates, adenosine and 2′-deoxyadenosine [Hershfield *et al.*, 1995]. ADA is needed for the breakdown of adenosine from food and for the turnover of nucleic acids in tissues. It irreversibly deaminates adenosine, converting it to the related nucleoside inosine by the removal of an amino group. Inosine can then be deribosylated (removed from ribose) by another enzyme called purine nucleoside phosphorylase (PNP) converting it to hypoxanthine.



Adenosine



Inosine

ADA has two isoforms, ADA1 and ADA2. ADA1 is found in all parts of the body cells, specifically lymphocytes and macrophages. ADA2 was first identified in human spleen [Persico *et al.*, 2000]. ADAR is an RNA-specific ADA [Keegan *et al.*, 2004]. ADA deficiency is caused by mutations in ADA gene located on chromosome 20q13.2-q13.11. The function of the ADA enzyme is to eliminate a molecule called deoxyadenosine.

```
       X            D    R
   1 MAQTPAFDKP KVELHVHLDG SIKPETILYY GRRRGIALPA NTAEGLLNVI    50

                              C
                              D
                              V W       D              C
 51  GMDKPLTLPD FLAKFDYYMP AIAGCREAIK RIAYEFVEMK AKEGVVYVEV   100

     L
     Q                                            X      W
     W  L# P            X    #  Q  M           E  Q      Q
101  RYSPHLLANS KVEPIPWNQA EGDLTPDEVV ALVGQGLQEG ERDFGVKARS   150

          C
      M   H                      M D#                    P
151  ILCCMRHQPN WSPKVVELCK KYQQQTVVAI DLAGDETIPG SSLLPGHVQA   200

                H
                C   TRK                 I Q   S
201  YQEAVKSGIH RTVHAGEVGS AEVVKEAVDI LKTERLGHGY HTLEDQALYN   250

       PX                     L                  L     Q
251  RLRQENMHFE ICPWSSYLTG AWKPDTEHAV IRLKNDQANY SLNTDDPLIF   300

        R     T         #          V        #  #
301  KSTLDTDYQM TKRDMGFTEE EFKRLNINAA KSSFLPEDEK RELLDLLYKA   350

351  YGMPPSASAG QNL   363
```

Figure: 1 Missense mutations have been highlighted with yellow background above protein sequence [Piirilä *et al.*, 2006]

PDB code of Adenosine deaminase is 3IAR. The number of aminoacid residues in ADA is 363.

It converts deoxyadenosine, which is toxic to lymphocytes, and another molecule called deoxyinosine that is not harmful. ADA deficiency is inherited in an autosomal recessive manner [OMIM]. Its absence was first identified by Giblett and coworkers in 1972 as a cause of SCID [Giblett *et al.*, 1972]. Change in a single nucleotide results in a codon that codes for a different amino acid that results in a protein nonfunctional which is known as missense mutations. Missense mutations, in contrast to single-nucleotide polymorphisms are rather rare events. However, numerous single gene diseases have been attributed to missense mutations.

In the research more than 70 ADA mutations have been identified in individuals with ADA deficiency with immunodeficiency or in healthy individuals with "partial ADA deficiency" [Hirschhorn, 1999; Hershfield & Mitchell 2001; Vihinen *et al.*, 2001]. The distribution of missense mutations in ADA is 60%. The detailed study about the molecular mechanisms is going to be the next major steps in the field of mutation research. Studying at the protein level of missense mutations is required for the clear picture. Disease phenotypic expressions arise when an amino acid which affects the protein function, for example, a residue in the catalytic site of an enzyme or a residue which is involved in important interactions with the partner molecules. Protein molecules are rather robust, which allows insertions in numerous sites without any effect on protein function [Pajunen *et al.*, 2007; Poussu *et al.*, 2004]. Even minor changes in the size or chemical nature of an amino acid side chain can alter or prevent the function of the protein. Most of the known ADA mutations have been discovered through research in the genotype to phenotype relationship [Hirschhorn *et al.*, 1990; Santisteban *et al.*, 1993; Arredondo-Vega *et al.*, 1994; Ozsahin *et al.*, 1997].

In addition to the direct functional effects a substitution may have, a missense mutation may also lead to alterations in the protein structural properties, causing abnormal folding, structural instability, tolerance or aggregation in the protein. In the cases like molecular recognition and interactions disordered segments are important [Mészáros *et al.*, 2007].

Many disorder prediction methods have been developed along with many new prediction tools. The disorder prediction tools is much discussed in the review articles[Bourhis *et al.*, 2007; Dosztányi *et al.*, 2007; Dunker *et al.*, 2008; Uversky *et al.*, 2008]. Missense mutations that cause protein aggregation had associated with pathology [Khemtemourian *et al.*, 2008; Robinson, 2008; Yankner and Lu, 2009]. The tolerance prediction programs are based on different machine learning techniques such as neural networks, random forests or support vector machines [Bao and Cui 2005; Bromberg and Rost 2007; Calabrese *et al*., 2009; Capriotti *et al*., 2006; FerrerCosta *et al*., 2005; Li *et al*. 2009] or it may be rule-based [Ramensky *et al*., 2002], mathematical operations [Ngand Henikoff, 2001]. Missense mutations tend to affect the structurally important residues, which lead to loss of structural integrity or stability of the protein involved in protein function [Mooney and Klein, 2002; Wang and Moult, 2001; Yue *et al*., 2005].

This study was carried out on 40 missense mutations found in ADA with the help of different kinds of predictions which are assembled under a single tool known as PON-P. The aim is to see how the identified missense mutations affect ADA. In this study we are using different kinds of predictions such as stability, disorder, aggregation, tolerance predictions along with conservation analysis and structural considerations of ADA are also studied.

1.1 Aims of the study

The aims of this research work:

1) To identify the missense mutations of ADA.
2) Broaden my knowledge on different kinds of predictors according to their respective functions which include stability change predictions, aggregation predictions, disorder and tolerance predictions.
3) Studying, how all these predictions work under PON-P.
4) Analyzing the results obtained from different predictions.
5) Comparing the results obtained from the single set of predictions with the other set.
6) Overall analysis of all the predictions.
7) Conservational analysis, structural considerations of ADA.
8) Obtaining a clear view about the effects of missense mutations in ADA using pathogenic or not pipeline (PON-P) thus make them available to the entire research community.

1.2 Significance of the Study

Amino acid substitutions have diverse effects on structure and function of the protein, therefore we need a detailed analysis of mutations is essential. Sequence based analysis provides the necessary information about the critical sites that are conserved and also have a crucial role in protein structure and function. After the structure has been determined experimentally, mutation analysis of that particular protein can be conducted, which makes the analysis more reliable and complete. Till recently, the research has mainly concentrated on using just one or few methods, but rapid increase in mutation analysis studies which utilizes a many set of prediction methods to attain more reliable results [Burke *et al*., 2007; Lappalainen *et al*., 2008; Tavtigian *et al*.,2008; Thusberg and Vihinen, 2006, 2007; Worth *et al*., 2007].

Missense mutations may result in protein nonfunctional. Such kind of mutations lead to diseases such as epidermolysis bullosa, sickle-cell disease etc. Studying the effects of missense mutations are important for several reasons. Mutations are the source of new variation important for evolution, considered as one of the ways that evolution happens. So, to have a deeper knowledge about the missense mutations I focused my project on the missense mutations. This research work aims to analyze 40 different kinds of missense mutations in ADA using PON-P.

2. Review of the Literature

Bioinformatics methods are utilized at different steps of the analysis. The sites that are conserved in evolution often have a crucial role in protein structure or function. There are numerous sequence-based predictors available for the prediction of the effect of a mutation on various biochemical properties of a protein, such as aggregation propensity, disorder, or stability. When there is an experimentally determined structure available for the protein of interest, the mutation analysis can be taken to the structural level, making the analysis more reliable and complete. Alternatively, modeled structure can be used. Several recent studies have applied computational methods to predict potentially deleterious effects of nonsynonymous SNP's in humans [Chasman and Adams, 2001; Hyytinen *et al*., 2002; Lauand Chasman, 2004; Miller and Kumar, 2001; Ng and Henikoff, 2001; Sunyaev *et al*., 2001; Terp *et al*., 2002; Torkamani and Schork, 2007; Wang and Moult, 2001; Wood *et al*., 2007; Worth *et al*., 2007]. The research has mainly concentrated on using just one or a few methods in one study but for a while, there has been many new prediction programs are available and their combination gives better results.

In the current scenario, computational methods and wide range of information from the databases containing information on DNA and protein sequences and on protein, structure, function are very much used in order to know the effects of mutation on the protein structure and function [Sunyaev *et al*., 2001] .

2.1 Missense mutations

Many different types of mutations are present in human genome and they are classified in two main group including major gene rearrangements and point mutations. Point mutations with in a gene that result in a substitution of one amino acid to another in a protein are called missense mutations. Missense mutations affect the protein structure and function. Therefore change in single amino acid leds to the results in multiple effects. Mutations lead to the deadly phenotypical changes. Effects of missense mutations range from early mortality in prenatal development to no observable phenotypic changes [Graf *et al*., 2000]. Mutations have an effect on a protein structure and function.

A missense mutation is a non-synonymous protein coding single nucleotide polymorphism (nsSNP). A missense mutation can cause change in the stability, aggregation, order, pathogenicity of the protein. The study of these mutations gives us a wide range of knowledge about the substituted amino acid and also the consequences that leds to the change in the function and structure of the aminoacid [Karachi, 2009; Mooney, 2005]. If a missense mutation is present on an important site which is responsible for the protein function leads to a disease phenotype. Adenosine deaminase is one of the example of the missense mutations. Missense mutation which gains clinical importance, changes the physicochemical properties of the amino acid residue to the extent which affects the function [Stone & Sidow, 2005].

In the online Mendelian Inheritance in Man (OMIM) database, more than 2200 genes are known to have mutations causing genetic diseases [Amberger *et al*., 2009]. The most deleterious missense mutations affect protein function indirectly through effects on protein structural stability [Wang and Moult 2001, Ramensky *et al*., 2002]. Even the evolutionary properties of the mutated residue is also an important thing to be considered and also their effect on protein function [Chasman & Adams, 2001; Henikoff, 2001]

2.2 Adenosine deaminase

Human ADA is 41kDa protein encoded by the *ADA* gene was mapped to chromosome 20q 13.2-q13.11, cloned, and sequenced [Valerio *et al*., 1984]. The chief physiological function of ADA is very much related to lymphocytic proliferation and differentiation [Erel *et al*., 1998]

An absence of the enzyme adenosine deaminase was identified by Giblett and co-workers in 1972 as a cause of SCID [Giblett *et al*., 1972]. ADA deficiency is an inherited enzyme which is a rare metabolic disorder which causes immunodeficiency. ADA deficiency typically causes SCID in infants, who presents with growth failure, infections, lymphopenia, and defective cellular and humoral immune function [Hershfield *et al*., 1995]. Infants with SCID typically experience pneumonia, chronic diarrhea and widespread skin rashes and decrease in growth is seen compared to healthy children. Neurological problems such as developmental delay, movement disorders, and hearing loss can be seen. Immunodeficiency results from toxic effects of ADA substrates, including apoptosis induced by deoxyadenosine triphosphate (dATP) pool expansion [Seto *et al*., 1985; Gao *et al*., 1995; Benveniste *et al*., 1995].

By the time of diagnosis these individuals often have chronic pulmonary insufficiency and may have autoimmune phenomena (cytopenias, anti-thyroid antibodies), allergies, and elevated serum concentration of IgE. ADA is the only gene associated with ADA deficiency. Sequence analysis can identify most known ADA mutations, except for large deletions, which are detected by deletion/duplication analysis. More than 70 ADA mutations have been identified in individuals with adenosine deaminase deficiency or in healthy individuals with "partial ADA deficiency [Hirchhorn, 1999; Hershfield & Mitchell 2001; Vihinen *et al*., 2001]. The distribution is 60% missense, 20% splicing, 9% intra-exonic deletion, 7% nonsense and 3% deletions of one or multiple exons.

Mutations at the adenosine deaminase locus in humans result in the phenotypic changes. The mutations which block the complete activity of ADA and its enzyme activity in all cell types lead to SCID. Y97C and L106V missense mutations carried on the same allele, have been identified, which carried same allele of an immunodeficient patient [Jiang *et al*., 1997]. Three missense mutations (H15D, A83D, and A179D) found in the severe combined immunodeficiency diseases [Santisteban, Ines *et al*., 1995]. H15D is the first naturally occurring mutation of a residue that co-ordinates directly with the zinc ion. G to A transition was identified by direct sequencing of PCR-amplified genomic DNA, predicting a glycine to arginine substitution at codon 20 (G20R) [Yang et al., 1994]. R156C and S291L are the two missense mutations have been identified in two ADA-SCID patients [Hirschhorn, R. 1992]. E217K missense mutation has been identified at the catalytic site in two ADA alleles of a patient with neonatal onset ADA-SCID [Hirschhorn *et al*., 1992]. Missense mutations (G74C, V129M, G140E, R149W, Q199P) were identified in seven patients (3 with SCID and 4 with delayed-onset) to better define the phenotype-genotype relationship [Arrendondo-vega *et al*., 1998].

2.3 Bioinformatic predictions

a. Stability prediction

One of the fundamental property affecting function, activity and regulation of biomolecules of a protein is stability. Change in protein stability upon site-specific mutations is the admissible problem related to the protein stability and function [Daggett and Fersht, 2003]. As a result different methods have been developed to predict stability changes, based mainly on the development of different energy functions, suited for calculating the stability free energy changes in protein structures when mutating one residue at a time in the sequence [Prevost *et al*., 1991; Topham *et al*., 1997; Pitera and Kollman, 2000; Kwasigroch *et al*., 2001; Guerois *et al*., 2002; Zhou and zhou, 2001]

Decrease in stability and incorrect folding are very important consequences of pathogenic missense mutations. The charged or polar function regions are located in hydrophobic clefts [Dessailly et al., 2007]. The methods for predicting protein stability changes resulting from single amino acid mutations can be classified in to four categories: (1) physical potential approach; (2) statistical potential approach; (3) empirical potential approach; and (4) machine learning approach [Capriotti E. *et al*, 2004]. First three methods are similar in that they rely on energy functions [Guerois *et al*., 2002]. Machine learning approaches can gain more complex nonlinear functions of input mutation, protein sequence, and structure information. Machine learning approaches such as support vector machines (SVMs) and neural networks are more robust in their handling of outliers than linear methods. Furthermore, machine learning approaches are not limited to using energy terms; they can readily leverage all kinds of information relevant to protein stability. A new machine-learning approach was developed [Jianlin Cheng *et al*., 2006] based on SVMs to predict stability changes for single site mutations in two contexts, structure- dependent and sequence dependent information .

A computational program, PoPMuSiC [Gilis and Rooman 2000] that predicts the point mutations along with experimental work, regardless of the protein or peptide environment. Different kinds of stability predictors were designed to predict the point mutations affecting stability in the proteins (Table 1)

Table 1: Stability predictors

Name	Web address	Reference
PopMusic [New version]	http://babylone.ulb.ac.be/popmusic/	Dehouck Y. *et al*., 2009
PopMusic [Old version]	http://babylone.ulb.ac.be/popmusic/	Gilis and Rooman, 2000
AUTO-MUTE	http://proteins.gmu.edu/automute/AUTO-MUTE.html	Masso M. & Vaisman 2010
CUPSAT	http://cupsat.tu-bs.de/	Parthiban V *et al*., 2006
Dmutant	http://sparks.informatics.iupui.edu/hzhou/mutation.html	Zhou and Zhou, 2002
FoldX	http://foldx.crg.es/	Guerois R *et al*., 2002
I-Mutant 2.0	http://gpcr.biocomp.unibo.it/cgi/predictors/I-Mutant2.0/I-Mutant2.0.cgi	Capriotti *et al*., 2005
I-Mutant 3.0	http://gpcr.biocomp.unibo.it/cgi/predictors/I-Mutant3.0/I-Mutant3.0.cgi	Capriotti *et al*., 2008
iPTREE-STAB	http://210.60.98.19/IPTREEr/iptree.htm	Huang *et al*., 2007
MuPRO	http://www.ics.uci.edu/~baldig/mutation.html	Cheng, *et al*., 2006
Scide	http://www.enzim.hu/scide/ide2.html	Dosztányi, *et al*., 2003
Scpred	http://www.enzim.hu/scpred/pred.html	Dosztányi, *et al*., 2003
Sride	http://sride.enzim.hu/	Magyar *et al*.,2005

b. Aggregation predictors

In the biotechnological and medical sciences, the aggregation of protein has become a very important subject [Fink, 1998; Smith, 2003]. Protein aggregation is the aggregation of misfolded proteins, and many neurodegenerative disorders like the Alzheimer’s disease, spongiform encephalopeties, type II diabetes mellitus and Parkinson’s disease in humans are linked with the conversion of peptides and proteins from soluble functional structure in to amyloid fibrils [Chiti & Dobson, 2006].

β-sheet structures along with α-helices have an inherent property to form amyloid fibrils. Some organisms produce amyloids for the function without having deleterious effects [Fowler *et al.*, 2006]. Investigation of mutations which affects the functional sites of a protein, such as DNA, ligand and protein interaction sites are of very important. Many tools have been developed to predict the stability, aggregation, pathogenicity and disorder of a protein. The primary structures of a polypeptide relatively to a large extent is capable to aggregate and the very small changes have a huge impact on solubility.

Now the prediction of aggregation of a protein from the sequence is of much value. Some of the authors have proved recently that a very short and specific amino acid stretches function as facilitators or inhibitors of amyloidal fibril formation [Ivanovo, 2004; Ventura, 2004] which are known as aggregation "hot spots". This led to the identification of aggregation-prone segments in several unstructured and globular disease-linked polypeptides. Aggrescan [Oscar Conchillo-Sole *et al*., 2007] predicts the the aggregation of "hotspots" and evaluates differential aggregation behaviour of polypeptides. For the prediction of amyloidogenic stretches, a consensus algorithm, AMYLPRED [Kimon K Frousios et al., 2009] was developed. Another aggregation predictor is PASTA [Trovato et al., 2007] useful for the prediction of amyloid structure aggregation, edits a pair-wise energy function for residues facing one another with a β-sheet. Waltz [Oliveberg 2010] is a new amyloid predictor which distinguishes 'true' amyloids from amorphous aggregates.

c. Structural disorder

Many proteins have disordered regions; some of the proteins are disordered in itself. Different names are given to them as disordered proteins, unstructured proteins, intrinsically unfolded proteins etc. Protein disorder is very important for understanding the protein function as well as protein folding [Plaxco & Gross, 2001 and Verkhivker *et al*., 2003]. Disordered proteins are in limelight because these disordered protein regions often lead to difficulties in purification and crystallization of proteins, which led bioinformaticians to design tools for the disordered regions.

Previously the disordered proteins are that doesn’t have a stable secondary structure, large number of conformations with X-ray crystallography, nuclear magnetic resonance (NMR), circular dichroism(CD) and various hydrodynamic measurements [Tompa, 2002; Receveur-Brechot *et al.*, 2006]. The structural biologists took the advantage of disorder predictions to indicate tight domains to get the solutions of the 3D structure, and also to facilitate the tertiary structure and threading predictions by dissecting the target sequences in to a set of independent folded domains [Friedberg *et al*., 2004].

Missense mutations can cause disorder in ordered structures, thereby affecting protein function. Missense mutations initiate the protein aggregates which indirectly may lead to different kinds of diseases. The primary structure of a protein is used to determine the aggregation for larger extent, even a single mutation may lead to increase in aggregation and solubility. The first tool which was designed especially for prediction of protein disorder was PONDR (Predictor of Naturally Disordered Regions, http://www.pondr.com) (Romero *et al*., 1997; Garner *et al*., 1998, 1999) which is based on artificial neural networks. Thereafter several programs have been developed by using the sequence of a protein (Table 2)

Table 2: Disorder prediction programs

Program	**Web address**	**Predicts**	**Reference**
DisProt	http://www.ist.temple.edu/disprot/Predictors.html.	Lacking fixed 3D structure	Vucetic S., Brown C.J., Dunker A.K. and Obradovic, Z., (2003).
Drip-Pred	http://www.sbc.su.se/~maccallr/disorder/		MacCallum (2006)
FoldIndex	http://bioportal.weizmann.ac.il/fldbin/findex	Lower hydrophobicity and high net charge regions.	Prilusky *et al*., (2005)
FoldUnfold	http://skuld.protres.ru/~mlobanov/ogu/ogu.cgi	Disordered region in chain.	Galzitskaya OV *et al*., (2006)
GlobPlot	http://globplot.embl.de/	Regions which have high inclination to globularity**.**	Linding *et al*., (2003)
IUPred	http://iupred.enzim.hu/	Regions lacking well-defined 3D	Dosztanyi *et al*., (2005)

		structure.	
RONN	http://www.strubi.ox.ac.uk/RONN	Regions that lack well defined 3D structure.	Yang,ZR. *et al.,* 2005
Spritz	http://protein.cribi.unipd.it/spritz/	disordered regions that are Inherited	Vullo *et al*., (2006)
PrDos	http://prdos.hgc.jp	Disordered protein regions from amino acid sequence	Ishida *et al*., (2007)
MetaPrDos	http://prdos.hgc.jp/cgi-bin/meta/top.cgi	Natively disordered regions of a protein chain from its amino acid sequence.	Ishida, T., *et al*., 2008

d. Tolerance Prediction

The mutations which can cause disease may likes to change the structure and function of the protein [Steward *et al* 2003, Vitkup *et al*., 2003]. The positions which are buried in the protein are more prone to pathogenic mutations than positions which are on the surface of the protein. The structural residues are more affected than functional residues by the majority of pathogenic mutations [Wang and Moult, 2001; Mooney & Klein, 2002]. Thus this process leads to the overall disturbance of the protein structure. Many efforts have been made to design a nearly perfect predictor which can take care of the pathogenic mutations. Currently many methods are available to predict the pathogenicity of missense mutations but the predictor which is accurate, sensitive and also minimizes the errors is Important. All the predictors nearly have the same default parameters as an input. Sequence is given in an Fasta format or a PDB code that gives the output in two options i.e. “pathogenic” or “neutral”.

Table 3: Tolerance predictors

Program	**Available at**	**Based on**	**Reference**

nsSNP analyzer	http://www.brightstudy.ac.uk/das_help.html	Random forests	Bao *et al*., 2005
SIFT	http://sift.jcvi.org/	Homology comparisons	Ng and Henikoff., 2003
phD-SNP	http://gpcr.biocomp.unibo.it/~emidio/PhD-SNP/PhD-SNP.htm	SVM-based classifier	Capriotti *et al*., 2006
Pmut	http://mmb2.pcb.ub.es:8080/PMut/	Feed-forward neural network.	Ferrer-Costa *et al*., 2005
Polyphen	http://genetics.bwh.harvard.edu/pph	PSIC score, difference in fitness between the wildtype and mutant aminoacid.	Ramensky *et al*., 2002
SNAP	http://cubic.bioc.columbia.edu/services/SNAP/	Neural network based method	Bromberg *et al.*, 2007
Panther	http://www.pantherdb.org/	subPSEC (substitution position-specific evolutionary conservation score)	Thomas *et al.*, 2003

2.4 Sequence conservational Analysis

The level of conservation of the physicochemical properties between the wild type residue and substituted residue has an effect on the pathogenicity of the mutation. Sometimes the conserved residues are hydrophobic in nature and are located in core of a protein and these can be identified by multiple sequence alignment. Visualizing multiple sequence alignments gives the information about the sequence conservation of a protein [Berezin *et al*., 2004]. The highly conserved positions in multiple sequence alignment can be seen often in functional sites. Sequence conservation is necessary for the prediction of functionally and structurally important residues. There are many programs visualizing the MSAs, calculating conservation indices for each position in the alignment and color-coding the alignment for different level of sequence conservation [Thusberg *et al*., 2009]. A neural network prediction method can be seen to separate the buried and exposed residues in globular proteins [Fariseli & Casadio, 2001].

Table 4: MSA and conservational analysis tools

Server	Available at	Based on	Reference
ClustalW	http://www.ebi.ac.uk/Tools/clustalw2/index.html	Accurate results for similar sequences	Thompson *et al*., 1994
M-Coffee	http://www.tcoffee.org/Projects_home_page/ m_coffee_home_page.html	Runs many MSA methods and gives the single output	Moretti *et al*., 2007
Consurf	http://consurf.tau.ac.il/	Color coding scheme to the protein structure	Glaser *et al*., 2003; Landau *et al*., 2005
Multidisp	http://bioinf.uta.fi/cgi-bin/MultiDisp.cgi	An visualization method, gives the color code to the groups of every amino acid	Riikonen, Vihinen, (in preparation)
Matrix Plot	http://www.cbs.dtu.dk/services/MatrixPlot/	Mutual information plots for sequence alignments	Gorodkin *et al*., 1999
ConSeq	http://conseq.tau.ac.il/	Neural network prediction and Visualisation	Berezin, C.E *et al*., 2004
ConSSeq	http://www.cbi.cnptia.embrapa.br /SMS/STINGm/consseq/	Based on sequence alignments, homology derived structures of proteins	Higa, R.H *et al*., 2004

3. Materials and Methods

3.1 Stability change Predictors

Missense mutations may have an effect on the stability of the protein via over-packing, altered contacts between amino acid side chains, reduction in hydrophobic area, altered structural strain in the protein backbone introduced by proline residues or changes in the interactions.

1) **CUPSAT:** (Cologne University Protein Stability Analysis Tool)

   Single amino acid mutations can significantly change the stability of a protein structure [Street AG et al., 1999]. Random mutations at a specified position may aid in designing thermostable or thermosensitive proteins so that the functionality of a protein can be altered to suit favorable biological and industrial purposes. Several groups have already developed tools [Street *et al.*, 1999; Saven *et al.*, 2002; Mendes *et al.*, 2002; Bolon *et al.*, 2003; Looger *et al.*, 2003] for this purpose with moderate prediction accuracy. CUPSAT is a similar web tool to analyze and predict protein stability changes upon point mutations (Parthiban *et al.*, 2006). Major components that construct the prediction model are the atom potentials and torsion angle potentials. ΔΔG values are calculated. It takes the PDB ID, amino acid residue no. and wild type amino acid. http://cupsat.tu-bs.de/

2) **I-Mutant:** Support vector machines based predictor of protein stability changes upon single point mutation from the protein sequence and structure.

   I-Mutant 2.0 is based support vector machines (SVMs). The ΔΔG values associated with the mutation, the correlation of predicted with expected values, as taken from the experimental database is 0.71 and 0.62, depending on the structure and sequence-base prediction, respectively. Two alternatives are available, PDB code of the protein and the protein sequence needs to be pasted. http://gpcr2.biocomp.unibo.it/cgi/predictors/I-Mutant2.0/I-Mutant2.0.cgi. Four different outputs are retrieved.

3) **Mupro:** Single amino acid mutations can significantly change the stability of a protein structure. A new machine learning approach based on support vector

machines to predict the stability changes of single site mutations in taking structure dependence in to account. http://www.ics.uci.edu/~baldig/mutation.html

Predicts the sign of relative stability change (ΔΔG). If the energy change ΔΔG is positive, the mutation increases stability and is classified as a positive example, and vice versa.

4) **SCide:** It is a program to determine the stabilization centers from known protein structures. http://www.enzim.hu/scide/. These are the residues which are involved in the cooperative long-range contacts and which can be formed between various regions of a single polypeptide chain or different polypeptide chain [Dosztänyi, *et al.*, 2003]. The server takes a PDB file as an input, and the result is presented in graphical or text format.

5) **Scpred:** SCPRED method improves prediction accuracy for sequences that share twilight-zone pairwise similarity with sequences used for the prediction [Lukasz Kurgan, et al., 2008]. It uses a support vector machine classifier that takes several custom-designed features as its input to predict the structural classes. Scpred is useful for site-directed mutagenesis, as a restraint in prediction of the 3D structure or can help in prediction of the folding class of a protein with unknown structure. Sequence is required in this program.

6) **iPTREE-STAB:** Decision tree coupled with adaptive boosting algorithm, and classification and regression tree respectively are the basis for the prediction. 82% accuracy for discriminating the stabilizing and destabilizing mutants [Liang-Tsung Huang *et al.*, 2007] the program takes the information about the mutant and mutated residues, three neighboring residues on both sides of the mutant residue along with pH and T. http://210.60.98.19/IPTREEr/iptree.htm. In the output, it displays the predicted protein stability change upon mutation along with input conditions. In the case of discrimination, it shows the effect of the mutation to protein stability, whether stabilizing or destabilizing.

3.2 Aggregation predictors

Aggregations of the misfolded regions are responsible for the various diseases. Aggregation predictors allow us to study how mutations affect the protein aggregation propensity.

1) **Aggrescan:** A server for the prediction and evaluation of "hot spots" of aggregation in polypeptides. Aggrescan is web-based software for the prediction of aggregation-prone segments in protein sequences. http://bioinf.uab.es/aggrescan/. It analyzes the effect of mutations on protein aggregation propensities and their comparison of different proteins or protein sets. It is derived from the *in vivo* experiments and on the assumption that short and specific sequence stretches modulate protein aggregation. It identifies a series of protein fragments involved in the aggregation of disease-related proteins. It takes the FASTA format of the peptide sequences as input sequence.
2) **Waltz:** An amyloid-prediction tool. Waltz algorithm is a position specific prediction algorithm which identifies amyloid forming hexa peptides in amino acid sequences [Oliveberg, 2010]. The method shows ~84% sensitivity at ~92% specificity. It takes a single sequence or several sequences in FASTA format. http://waltz.vub.ac.be/)
3) **Pasta:** Prediction of amyloid structure aggregation. A web server for the analysis of amino acid sequences. http://protein.cribi.unipd.it/pasta/. Predicts the most aggregation-prone portions and the corresponding β-strand inter-molecular pairing for a given input sequence [Antonio Trovato *et al*., 2007]. A FASTA sequence is given as an input.

4) **Amylpred:** Amyloidoses are a group of usually fatal diseases, probably caused by protein misfolding and subsequent aggregation in to amyloid fibrillar deposits [Frousios, et *al.,* 2009]. A combination of different types of tools is more error free than individual prediction methods. It utilizes molecular graphics programs, like PyMOL, as well as the algorithm DSSP. A plain text, Swiss Prot, or FASTA format is given as an input. http://biophysics.biol.uoa.gr/AMYLPRED/input.html

3.3 Disorder predictors

Protein disorder is important for understanding protein function as well as protein folding pathway. Mutations may introduce disorder into usually ordered parts of a protein thereby causing alterations in the protein fold leading to possible changes in protein function.

1. **Disprot:** The database of disordered proteins (DisProt) provides information about both the proteins that lack fixed 3D structure in their native states as well as proteins that have local regions lacking a fixed 3D structure [Slobodan Vucetic *et al.*, 2004]. Disprot is designed as a companion to many online protein repositories like PDB, SWISS-PROT and TrEMBL, GenBank and PIR databases. Sequence is given as an input. E-mail address is required to send the results after running the program. http://www.ist.temple.edu/disprot/Predictors.html.
2. **Drip-pred:** Predicts the structurally disordered regions in proteins. Intrinsically disordered/unstructured proteins exist in a highly flexible conformational state, yet they carry out essential functions. http://www.sbc.su.se/~maccallr/disorder/cgi-bin/submit.cgi. This method also includes PSIPRED. The Kohonen's self organizing map (SOM) is used [MacCallum, 2005].
3. **Fold-Index:** A simple tool to predict whether a given protein sequence is intrinsically unfolded. It implements the algorithm [Uversky *et al.*, 2000] which is based on the average residue hydrophobicity and net charge of the sequence **[**Prilusky *et al.*, 2005]. It provides a single score for the entire sequence, predicting whether it is folded or not http://bioportal.weizmann.ac.il/fldbin/findex
4. **Fold unfold:** web server for the prediction of disordered regions in protein chain. Mean packing density of residues has been introduced to detect disordered regions in a protein sequence. Weak packing density would be responsible for the appearance of disordered regions [Oxana, Galzitskaya *et al.*, 2006]
5. **Globplot:** Intrinsic protein disorder, domain and globularity prediction. Non-globular sequence segments often contain short linear peptide motifs which are important for the protein function. This web service allows us to plot tendency within the query

protein for order/globularity and disorder. Globplot is mainly useful in domain hunting [Linding et al., 2003]. Plots indicate that instances of known domains may often contain additional N-or C-terminal segments. http://globplot.embl.de/

6. **IUPred:** Web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content [Dosztányi *et al.*, 2005]. IUPred presents a novel algorithm for predicting such regions from amino acid sequences by estimating their total pairwise interresidue interaction energy, based on the assumption that intrinsically unstructured protein sequences do not fold due to their inability to form sufficient stabilizing interresidue interactions [Dosztányi *et al.*, 2005]. The server takes a single amino acid sequence as an input and calculates the pair wise energy profile along the sequence. http://iupred.enzim.hu/. Probabilistic score ranges from 0 (complete order) to 1 (complete disorder), >0.5 (disorder).
7. **metaPRDOS:** Prediction of disordered regions in proteins based on the meta approach. This tool predicts based on the support vector machines from the prediction results of the seven independent predictors [Takashi Ishida et al., 2008]. Disorder prediction is one of the categories of the CASP experiments [Melamud & Moult, 2003]. A three letter PDB code is given as input. http://prdos.hgc.jp/cgi-bin/meta/top.cgi.
8. **Prdos:** Prediction of disordered protein regions from amino acid sequence. Predicts the disordered regions of a protein from its amino acid sequence. A plain text or FASTA format can be given as input. http://prdos.hgc.jp/cgi-bin/top.cgi . Two predicotrs are composed in this prediction system one is local amino acid sequence information and the other based on template proteins [Takashi Ishida *et al.*, 2007].
9. **Prelink:** Prediction of unfolded segments in a protein sequence based on amino acid composition.
10. **Ronn:** Regional order neural network (RONN) software. Nine methods for predicting protein disorder were tested against 80 proteins forming the smaller blind test set, each of which contains at least one significant region of disorder. [Zheng Rong Yang *et al.*, 2005]. Single protein sequence in plain text or FASTA format is given as input. http://www.strubi.ox.ac.uk/RONN

11. **Spritz:** Machine learning approach based on two support vector machines to discriminate disordered regions from sequence [Vullo, 2006]. E-mail address, along with plain sequence, selecting a short disorder or long disorder with false positive rate from 0.01 to 0.1 is given as input. http://distill.ucd.ie/spritz/

3.4 Tolerance predictors

Tolerance predictors are used to predict the pathogenicity of the missense mutations, based on the phylogenetic information, sequence conservation, multiple sequence alignment, sequence and structural information.

1. **nsSNP analyzers:** Describes the distribution of both types of nsSNPs using structural and sequence based features and evaluates the relative value of these attributes as predictors of function and uses machine learning methods [Dobson *et al*., 2006]. http://www.brightstudy.ac.uk/das_help.html
2. **SIFT:** SIFT can differentiate between the functionally neutral and pathogenic amino acids [Ng et *al*., 2003]. The number of substitutions that SIFT can predict on is expected to increase as more genomes are sequenced and more protein sequences become available. A FASTA format sequence is required as an input. http://sift.jcvi.org/
3. **Phd-SNP**: Predictor of human deleterious single nucleotide polymorphisms. It is based on a decision tree with the SVM-based classifier coupled to SVM-Profile. Three datasets are used for the construction of this method, first for training/testing the SVM system based on sequence information (HumVar), second for training/testing our SVM system based on profile information (HumVarProf) and the third, to be used when testing the robustness of the predictor (New Hum Var) [Capriotti *et al*., 2006]. Protein sequence or PDB code or sequence file, position of the mutation and the new substituted residue are given as an input. http://gpcr.biocomp.unibo.it/~emidio/PhD-SNP/PhD-SNP.htm
4. **PMut**: A web-based tool for the annotation of pathological mutations on proteins. It gives fast and accurate results in humans [Ferrer-Costa *et al*., 2005]. PMut retrieves

the information from a local database and it analyzes the given SNP in a specific protein. PDB code or FASTA sequence, positions or mutations are given as input. http://mmb2.pcb.ub.es:8080/PMut/ .

5. **Polyphen:** It is a tool which predicts possible impact of an amino acid substitution on the structure and function of a human protein using straightforward physical and comparative considerations. http://genetics.bwh.harvard.edu/pph. Protein identifier or AA sequence in FASTA format is required for the input [Ramensky *et al*, 2002]
6. **SNAP:** SNAP could predict gain of function as well as loss of function [Bromberg et al., 2007]. It requires protein sequence and list of substitutions as an input.

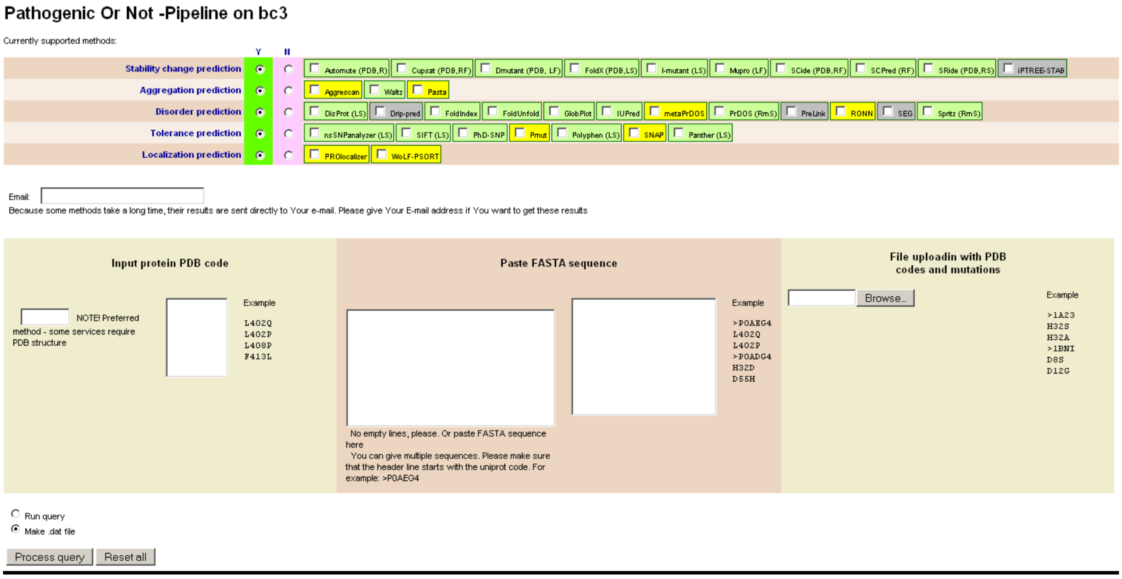
3.5 PON-P: Pathogenic-Or Not Pipeline

Portal, that is used for mutation effect analysis. Meta tools are tools which combine the analysis of several other prediction tools. Benefit of this is that the weaknesses of one program can be compensated by others and therefore the result is expected to be more reliable than obtained by individual programs.

Pathogenic Or Not- Pipeline is a meta tool that combines methods from the following categories: stability change prediction, aggregation prediction, disorder prediction and pathogenicity prediction. PON-P aims at providing a comprehensive collection of mutation analysis methods, facilitating their use. User should provide the PDB code or FASTA sequence or file uploading with PDB codes and mutations, along with e-mail address.A service where an input data can be simultaneously submitted provided by the user to select his required prediction methods, as well as parses the outputs of individual methods into a single output, will simplify the process and provide results faster and more conveniently [Thusberg & Vihinen, 2008]

Currently new version of Pon-P service (Fig.1) is in the process of development by our bioinformatics group and a test version can be obtained at

http://bioinf.uta.fi/cgi-bin/ponp/application_test.cgi



Pathogenic Or Not -Pipeline on bc3

Currently supported methods:

	Y	N	
Stability change prediction	●	○	□ Automute (PDB,R) □ Cupsat (PDB,RF) □ Dmutant (PDB, LF) □ FoldX (PDB,LS) □ I-mutant (LS) □ Mupro (LF) □ SCide (PDB,RF) □ SCPred (RF) □ SRide (PDB,RS) □ iPTREE-STAB
Aggregation prediction	●	○	□ Aggrescan □ Waltz □ Pasta
Disorder prediction	●	○	□ DisProt (LS) □ Drip-pred □ FoldIndex □ FoldUnfold □ GlobPlot □ IUPred □ metaPrDOS □ PrDOS (RmS) □ PreLink □ RONN □ SEG □ Spritz (RmS)
Tolerance prediction	●	○	□ nsSNPanalyzer (LS) □ SIFT (LS) □ PhD-SNP □ Pmut □ Polyphen (LS) □ SNAP □ Panther (LS)
Localization prediction	●	○	□ PROlocalizer □ WoLF-PSORT

Email:

Because some methods take a long time, their results are sent directly to Your e-mail. Please give Your E-mail address if You want to get these results

Input protein PDB code

NOTE! Preferred method - some services require PDB structure

Example

L402Q
L402P
L408P
F413L

Paste FASTA sequence

No empty lines, please. Or paste FASTA sequence here
You can give multiple sequences. Please make sure that the header line starts with the uniprot code. For example: >P0AEG4

Example

>P0AEG4
L402Q
L402P
>P0ADG4
H32D
D55H

File uploadin with PDB codes and mutations

Example

>1A23
H32S
H32A
>1BNI
D8S
D12G

○ Run query
● Make .dat file

Figure 2: PON-P service which is under preparation by Bioinformatics Group, Institute of Medical Technology, Tampere University.

3.6 Sequence conservation Analysis:

Conserved positions in a protein are highly essential for functional or structural reasons [Miller & Kumar, 2001; Mooney & Klein, 2002; Shen & Vihinen, 2004]. Disease causing mutations affecting highly conserved positions thus have a high likelihood of being pathogenic. The knowledge of the level and type of evolutionary conservation at the affected position is important for studying the pathogenicity of a missense mutation. There are several methods available for the detection of positional sequence conservation and identification of individual residues within a position [Ahola *et al*., 2004]. These methods calculate the conservation indices for each position in the alignment, and also add the color codes in to the alignment for different levels of sequence conservation.

Some methods are:

1. **CONSURF:** Catalytic activity, binding to ligand, DNA or other proteins, frequently under strong evolutionary resistance are the key positions which are important for maintaining the 3D structure of a protein and its function. Evolutionary conservation

and biological importance of a residue are often correlated [Landau *et al*., 2005]. Consurf is a web-based tool that automatically calculates evolutionary conservation scores and maps them on protein structures via a user-friendly interface. Structurally and functionally important regions in the protein typically appear as patches of evolutionarily conserved residues that are spatially close to each other. The consurf server is available at http://consurftest.tau.ac.il/

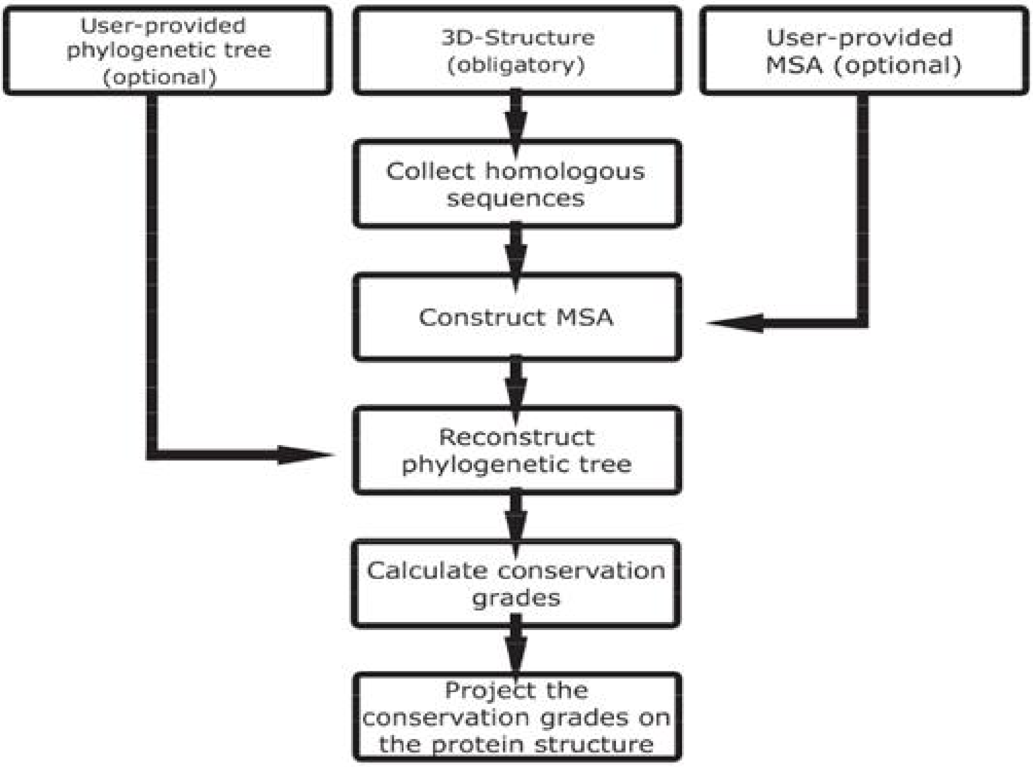

User-provided phylogenetic tree (optional)
3D-Structure (obligatory)
User-provided MSA (optional)
Collect homologous sequences
Construct MSA
Reconstruct phylogenetic tree
Calculate conservation grades
Project the conservation grades on the protein structure

Figure 3: A flowchart of ConSurf calculation **[**Landau *et al*., 2005]

2. **ConSeq:** ConSeq is a web server for the identification of functionally and structurally important residues. Conserved residues within the protein core are more conserved [Berezin *et al*., 2004]. ConSeq calculates the substitution rate at each position in the MSA, taking in to account the evolutionary relations between the homologues that uses maximum likelihood paradigm [Pupko *et al*., 2002] and it also applies the neural network prediction scheme to discriminate between buried and exposed residues in globular proteins [Fariselli & Casadio, 2001].

3. **ConSSeq:** A web-based application to analyze protein amino acids conservation-Consensus Sequence (ConSSeq). Homology-derived structures of proteins are used [Higa *et al.* 2004]. It is a part of the STING Millennium Suite (SMS) [Neshich *et al.*, 2003]. Homology derived structures of proteins (HSSP) provides two measures of variability: conservation weight and relative entropy [Sander and Schneider, 1991]. Relative entropy is inversely proportional to the conservation i.e. if the entropy= 0.00, then it is fully conserved and if the entropy is high, then it is said to be least conserved. http://sms.cbi.cnptia.embrapa.br/SMS/STINGm/consseq/

4. **Matrix plot:** Matrix plot is a program for making matrix plots such as mutual information plots of sequence alignments and distance matrices of sequences with known three-dimensional coordinates [Gorodkin *et al.*, 1999]. The matrixplot is available at http://www.cbs.dtu.dk/services/MatrixPlot.

3.7 Structural considerations

A physical and chemical properties may be altered when one of the residue is replaced by another residue. When the wild type residue is smaller than the substitution residue then there might be a major change in the structure. Rotamer analysis can be used to study the effects of the side chains of the mutant residue on the protein structure. PyMOL [DeLano, 2002], KiNG [Lovell et al., 2003], Discovery studio (Accelrys, San Diego, CA], or Swiss-PDB-Viewer [Guex and Peitsch, 1997] are used to model the new side chain. Substituted side chain is allowed to rotate during analysis. MolProbity [Davis et al., 2007] is used for the structural studies.

4. RESULTS

The protein sequence-structure [Voigt *et al*., 2000] and structure-function relationships gives us the way to understand the mutational effects on the structure of the protein.

In this study, 40 missense mutations have identified. H15D,G20R, G74C, G74D, G74V, R76W, A83D, Y97C, R101L, R101L, R101Q, R101W, P104L, L107P, P126Q, V129M, G140E, R142Q, R149Q, R149W, L152M, R156H, R156C, V177M, A179D, Q199P, R211C, R211H, A215T, G216R, E217K, T233I, R235Q, G239S, R253P, P274L, S291L, P297Q, L304R, M310T, A329V.

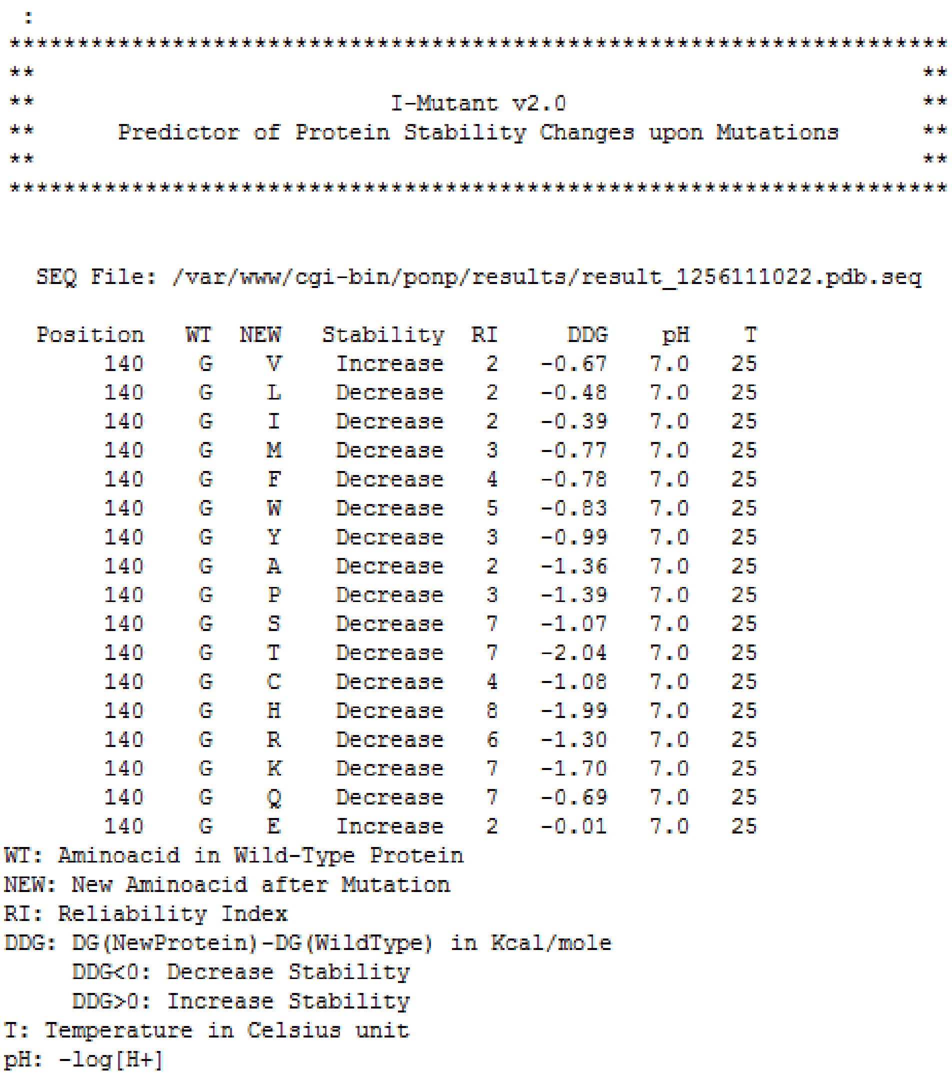
In the PON-P tool, three letter PDB code of ADA 3IAR is given; file containing the mutations is uploaded. In this study stability change prediction, aggregation prediction, disorder prediction, tolerance prediction methods are analyzed which includes different programs.

4.1 Stability change prediction

Missense mutations main effect has been shown to be on protein stability [De Cristofaro *et al*., 2006; Koukouritaki *et al*., 2007; Ode *et al*., 2007]. A residue is selected as a stabilizing residue if it has high surrounding hydrophobicity, high long-range order, and high conservation score and if it belongs to a stabilization center [Magyar *et al*., 2005]. Effects of mutations on the structural stability of proteins were studied by programs predicting stabilizing residues in proteins and by programs evaluating mutation induced stability changes in proteins, and this study focused on the latter one. There are several programs which study the mutational effect on the protein stability. The programs used under this prediction are CUPSAT, I-mutant, Mupro, Scide, ScPred, Sride, iPTree-STAB. Mupro uses feed-forward neural networks and SVMs. A score near 0 means unchanged stability. Score near -1 means high confidence in decreased stability. Score near +1 means high confidence in increased stability. In Mupro, a threshold of $\Delta\Delta G \leq -0.5$ kcal/mol is used for stabilizing mutations and $\Delta\Delta G \geq 0.5$ kcal/mol for destabilizing. Only Mupro gives an output stating that the given all missense mutations are stabilized with the $\Delta\Delta G \leq -0.5$ kcal/mol.

I-Mutant 2.0, machine learning-based and predict a stability change (positive or negative) of the whole protein for a given single-point mutation. If a residue is crucial for retaining the current structure, substitutions would result in large negative predictions. A threshold of ΔΔG ≤ -0.5 kcal/mol was used for stabilizing mutations, ΔΔG ≥ 0.5 kcal/mol destabilizing. In the figure 4, it can be noticed that the substitution of glycine with glutamic acid at the position 140, G140E exceptionally leads to Increase in stability with ΔΔ G value -0.01**.** In this study the missense mutations G140E, A329V are found to be destabilizing ADA.

3IAR 140



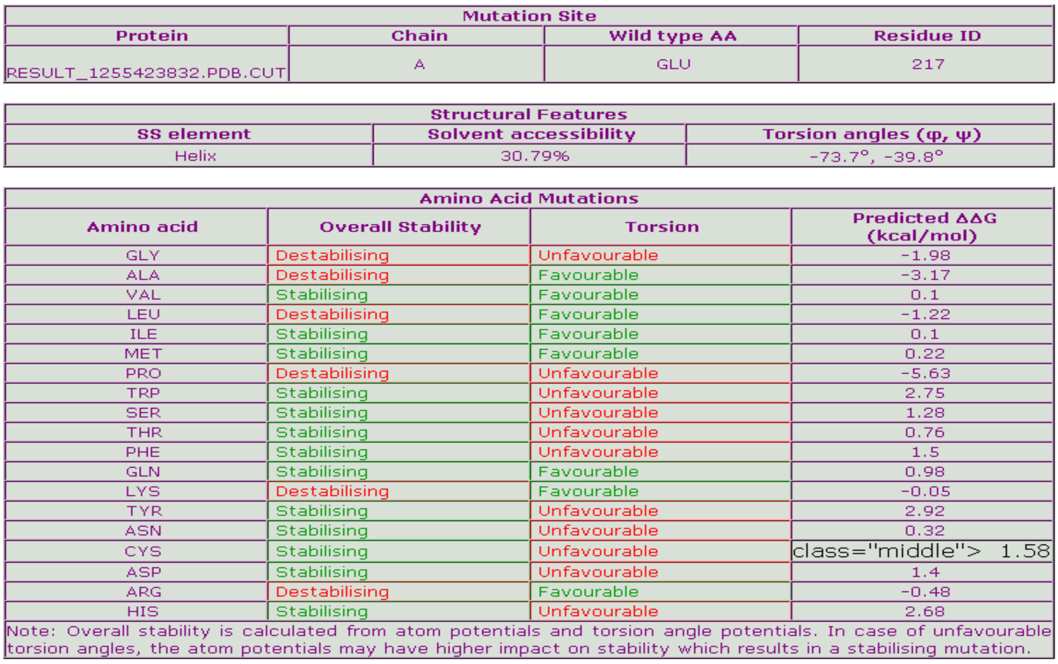
```
 :
**********************************************************************
**                                                                  **
**                           I-Mutant v2.0                          **
**        Predictor of Protein Stability Changes upon Mutations     **
**                                                                  **
**********************************************************************


  SEQ File: /var/www/cgi-bin/ponp/results/result_1256111022.pdb.seq

  Position   WT  NEW   Stability  RI     DDG     pH     T
       140    G    V    Increase   2   -0.67    7.0    25
       140    G    L    Decrease   2   -0.48    7.0    25
       140    G    I    Decrease   2   -0.39    7.0    25
       140    G    M    Decrease   3   -0.77    7.0    25
       140    G    F    Decrease   4   -0.78    7.0    25
       140    G    W    Decrease   5   -0.83    7.0    25
       140    G    Y    Decrease   3   -0.99    7.0    25
       140    G    A    Decrease   2   -1.36    7.0    25
       140    G    P    Decrease   3   -1.39    7.0    25
       140    G    S    Decrease   7   -1.07    7.0    25
       140    G    T    Decrease   7   -2.04    7.0    25
       140    G    C    Decrease   4   -1.08    7.0    25
       140    G    H    Decrease   8   -1.99    7.0    25
       140    G    R    Decrease   6   -1.30    7.0    25
       140    G    K    Decrease   7   -1.70    7.0    25
       140    G    Q    Decrease   7   -0.69    7.0    25
       140    G    E    Increase   2   -0.01    7.0    25
WT: Aminoacid in Wild-Type Protein
NEW: New Aminoacid after Mutation
RI: Reliability Index
DDG: DG(NewProtein)-DG(WildType) in Kcal/mole
     DDG<0: Decrease Stability
     DDG>0: Increase Stability
T: Temperature in Celsius unit
pH: -log[H+]
```

Figure 4: Substitution of wild amino acid (G) with other amino acids at position 140 and the result shows increase/decrease of protein stability with ΔΔG values.

Cupsat uses structural environment specific atom potentials and torsion angle potentials to predict ΔΔG, the difference in free energy of unfolding between wild-type and mutant proteins. This program gives an output of 30 missense mutations which destabilizes ADA. G20R,G74C,G74D,G74V,R76W,A83D,R101L,R101Q,R101W,P104L,P126Q,V129M, G140E,R142Q,R149W,L152M,R156H,R156C,A179D,R211C,R211H,A215T,G216R, R235Q,G239S,R253P,P274L,P297Q,L304R,M310T.



Mutation Site

Protein	Chain	Wild type AA	Residue ID
RESULT_1255423832.PDB.CUT	A	GLU	217

Structural Features

SS element	Solvent accessibility	Torsion angles (φ, ψ)
Helix	30.79%	-73.7°, -39.8°

Amino Acid Mutations

Amino acid	Overall Stability	Torsion	Predicted ΔΔG (kcal/mol)
GLY	Destabilising	Unfavourable	-1.98
ALA	Destabilising	Favourable	-3.17
VAL	Stabilising	Favourable	0.1
LEU	Destabilising	Favourable	-1.22
ILE	Stabilising	Favourable	0.1
MET	Stabilising	Favourable	0.22
PRO	Destabilising	Unfavourable	-5.63
TRP	Stabilising	Unfavourable	2.75
SER	Stabilising	Unfavourable	1.28
THR	Stabilising	Unfavourable	0.76
PHE	Stabilising	Unfavourable	1.5
GLN	Stabilising	Favourable	0.98
LYS	Destabilising	Favourable	-0.05
TYR	Stabilising	Unfavourable	2.92
ASN	Stabilising	Unfavourable	0.32
CYS	Stabilising	Unfavourable	class="middle"> 1.58
ASP	Stabilising	Unfavourable	1.4
ARG	Destabilising	Favourable	-0.48
HIS	Stabilising	Unfavourable	2.68

Note: Overall stability is calculated from atom potentials and torsion angle potentials. In case of unfavourable torsion angles, the atom potentials may have higher impact on stability which results in a stabilising mutation.

Figure 5: Cupsat Results for 3IAR, substitution of E by K at 217, results in destabilizing a negative ΔΔG value of -0.05

Scide: The results of Scide identification is presented in text or graphical format. A conclusion was drawn showing that 18 missense mutations destabilizes ADA. H15D, G20R, Y97C, R101L, R101Q, R101W, P104L, P126Q, V129M, L152M, V177M, A179D, G216R, E217K, T233I, R235Q, S291L, A329V.

Sride: The output of the server is a list of the sequences used to calculate the conservation score and the list of the SRs, together with the surrounding hydrophobicity (H_P), long range order (LRO) and conservation score values. Sride program concluded only 2 of 40 identified missense mutations, L152M, A179D destabilizes the enzyme ADA.

iPTREE-STAB: Predicts stability changes (ΔΔG) upon single amino acid substitutions from amino acid sequence. H15D, R76W, Q199P, T233I are the 4 missense mutations that destabilizes the protein ADA.

Scpred: In the output the the first one is your sequence; the second one contains digits of '0' or '1'. '0' stands for residues predicted to be not involved in stabilization centers, '1' stands for residues involved in stabilization centers. 14 missense mutations - H15D, A83D,Y97C, R101Q, R101W, P104L,V129M, L152M,V177M, A179D, R211C, R211H, P297Q, M310T destabilizes the enzyme ADA.

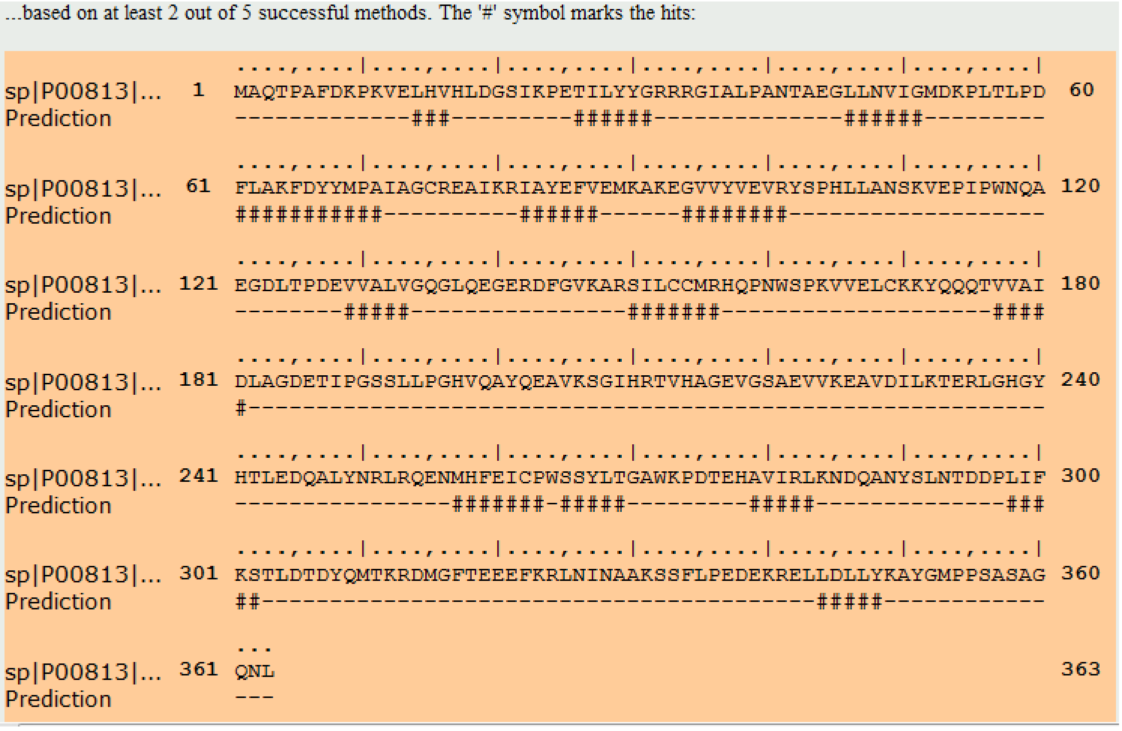
Table 5: The output summary of different stabilization programs

Program	**Mutations which destabilize the ADA enzyme**
Sride	L152M, A179D
iPTREE-STAB	H15D, R76W, Q199P, T233I
Mupro	Missense mutations have no effect on the stability of the enzyme
I-mutant	G140E, A329V
Cupsat	G20R,G74C,G74D,G74V,R76W,A83D,R101L,R101Q,R101W,P104L,P126Q,V129M,G140E,R142Q,R149W,L152M,R156H,R156C,A179D,R211C,R211H,A215T,G216R,R235Q,G239S,R253P,P274L,P297Q,L304R,M310T
Scide	H15D,G20R,Y97C,R101L,R101Q,R101W,P104L,P126Q,V129M,L152M,V177M,A179D,G216R,E217K,T233I,R235Q,S291L,A329V
ScPred	H15D,A83D,Y97C,R101Q,R101W,P104L,V129M,L152M,V177M, A179D,R211C,R211H,P297Q,M310T

4.2 Aggregation prediction

The aggregation predictor programs used for this study are Waltz, Pasta, Amylpred, and Aggrescan. The program Waltz resulted in two missense mutations, Q199P, S291L. Pasta program gave only one missense mutation (H15D) in the output which aggregates the enzyme.

Amylpred’s output has 12 missense mutations that aggregates the enzyme ADA.H15D,A83D,Y97C,R101L,R101Q,R101W,V129M,L152M,R156H,R156C, V177M,A179D. 18 missense mutations have been identified from the output which shows the aggregation propensity by Amylpred program, G74C,G74V,Y97C,R101L,R101Q,R101W,R149W,R211C,A215T, G216R,E217K, T233I, R235Q, G239S, R253P, P274L, S291L, A329V.



```
...based on at least 2 out of 5 successful methods. The '#' symbol marks the hits:

                      ....,....|....,....|....,....|....,....|....,....|....,....|
sp|P00813|...    1    MAQTPAFDKPKVELHVHLDGSIKPETILYYGRRRGIALPANTAEGLLNVIGMDKPLTLPD   60
Prediction            -------------###---------######--------------######---------

                      ....,....|....,....|....,....|....,....|....,....|....,....|
sp|P00813|...   61    FLAKFDYYMPAIAGCREAIKRIAYEFVEMKAKEGVVYVEVRYSPHLLANSKVEPIPWNQA  120
Prediction            ###########----------######------########-------------------

                      ....,....|....,....|....,....|....,....|....,....|....,....|
sp|P00813|...  121    EGDLTPDEVVALVGQGLQEGERDFGVKARSILCCMRHQPNWSPKVVELCKKYQQQTVVAI  180
Prediction            --------#####----------------#######--------------------####

                      ....,....|....,....|....,....|....,....|....,....|....,....|
sp|P00813|...  181    DLAGDETIPGSSLLPGHVQAYQEAVKSGIHRTVHAGEVGSAEVVKEAVDILKTERLGHGY  240
Prediction            #-----------------------------------------------------------

                      ....,....|....,....|....,....|....,....|....,....|....,....|
sp|P00813|...  241    HTLEDQALYNRLRQENMHFEICPWSSYLTGAWKPDTEHAVIRLKNDQANYSLNTDDPLIF  300
Prediction            ----------------#######-#####---------#####--------------###

                      ....,....|....,....|....,....|....,....|....,....|....,....|
sp|P00813|...  301    KSTLDTDYQMTKRDMGFTEEEFKRLNINAAKSSFLPEDEKRELLDLLYKAYGMPPSASAG  360
Prediction            ##-----------------------------------------#####------------

                      ...
sp|P00813|...  361    QNL                                                           363
Prediction            ---
```

Figure 6: Amylpred Results.

Amylpred [12], Aggrescan [18] output shows more number of missense mutations which aggregates ADA compared to Waltz [2] and Pasta [1] (Table 6)

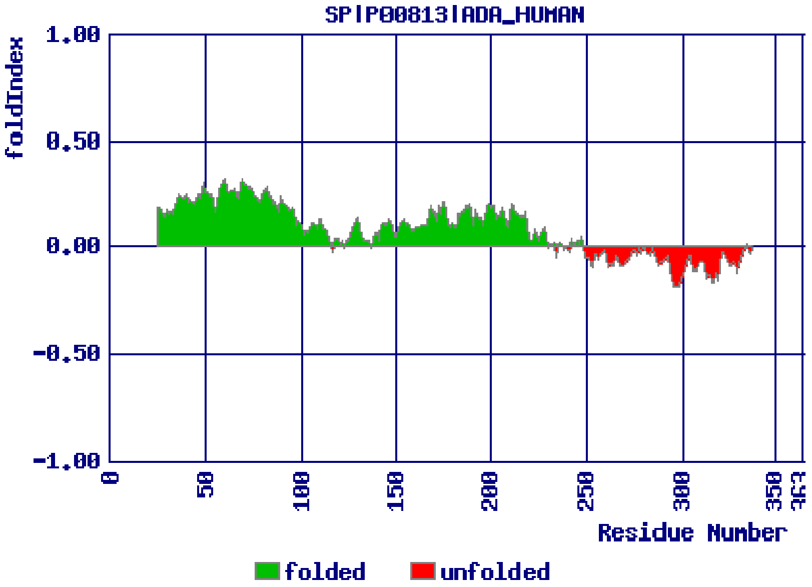
Table 6. Aggregation prediction output

Predictor (Tool)	**Missense mutations that aggregate ADA**
Waltz	Q199P, S291L
Pasta	H15D
Amylpred	H15D,A83D,Y97C,R101L,R101Q,R101W,V129M,L152M,R156H,R156C,V177M,A179D
Aggrescan	G74C, G74V, Y97C, R101L, R101Q, R101W, R149W, R211C, A215T, G216R, E217K, T233I, R235Q, G239S, R253P, P274L, S291L, A329V.

4.3 Disorder prediction

The ten different programs we used under the disorder prediction studies are Disprot, DRIP-PRED, Fold index, Fold Unfold, Glob plot, Inured, MetaPrDos, PrDos, RONN, and Spritz. The programs DRIP-PRED, Globplot, IuPred, MetaPrDos, PrDos, RONN shows that the 40 identified missense mutations do not change the order of the enzyme ADA which can be seen in the Table 7.

DisProt gives an output in which two missense mutations L304R, A329V found to affect the order of the functionally important regions. Foldindex give an information whether the given protein folds due to the missense mutations. In this study 7 out of 40 identified missense mutations affects the order of the enzyme.



```
Summary:
 Number Disordered Regions: 3
 Longest Disordered Region: 53
 Number Disordered Residues: 113
 Predicted disorder segment: [249]-[279] length: 31 score: -
0.06 ± 0.03
 Predicted disorder segment: [281]-[333] length: 53 score: -
0.09 ± 0.05
 Predicted disorder segment: [335]-[363] length: 29 score: -
0.02 ± 0.01
```

```
 1 MAQTPAFDKP KVELHVHLDG SIKPETILYY GRRRGIALPA NTAEGLLNVI
 51 GMDKPLTLPD FLAKFDYYMP AIAGCREAIK RIAYEFVEMK AKEGVVYVEV
 101 RYSPHLLANS KVEPIPWNQA EGDLTPDEVV ALVGQGLQEG ERDFGVKARS
 151 ILCCMRHQPN WSPKVVELCK KYQQQTVVAI DLAGDETIPG SSLLPGHVQA
 201 YQEAVKSGIH RTVHAGEVGS AEVVKEAVDI LKTERLGHGY HTLEDQALYN
 251 RLRQENMHFE ICPWSSYLTG AWKPDTEHAV IRLKNDQANY SLNTDDPLIF
 301 KSTLDTDYQM TKRDMGFTEE EFKRLNINAA KSSFLPEDEK RELLDLLYKA
 351 YGMPPSASAG QNL
```

Figure: 7 FoldIndex results. The seven missense mutations which cause disorder are under lined

FoldUnfold shows that weak expected packing density would be responsible for the appearance of disordered regions. 12 missense mutations show that a substitution of the wild amino acid would lead to disordered regions.

Spritz program gave an output with missense mutations P126Q, P274L altering the order of ADA.

Table 7: Disorder prediction results

Predictor(Tool)	Missense mutations which lead to disorder protein
DisProt	L304R, A329V
DRIP-PRED	Missense mutations has no affect on the order of protein
Foldindex	R253P, P274L, S291L, P297Q, L304R, M310T, A329V
FoldUnfold	G20R, P126Q, G140E, R142Q, R211C, R211H, A215T, G216R, E217K, P274L, S291L, M310T.
Globplot	Missense mutations has no affect on the order of protein
IuPred	Missense mutations has no affect on the order of protein
MetaPrDos	Missense mutations has no affect on the order of protein
PrDos	Missense mutations has no affect on the order of protein
RONN	Missense mutations has no affect on the order of protein
Spritz	P126Q, P274L

4.4 Tolerance prediction

Seven programs, nsSNPanalyzer, SIFT, PhD-SNP, Pmut, Polyphen, SNAP, Panther were used under tolerance prediction. From the Table 8, can be seen that all the seven programs gave an output in which most of the 40 identified missense mutations are not tolerated. It can be concluded quantitatively based on the majority of predictors that the missense mutations in ADA lead to pathogenicity.

Table 8: Showing the tolerance missense mutations

Predictor(Tool)	Intolerant (pathogenic) missense mutations
nsSNPanalyzer	H15D,G20R, G74C, G74D, G74V, R76W, A83D, Y97C, R101L, R101L, R101Q, R101W, P104L, L107P, P126Q, G140E,R149Q, R149W,R156H, R156C, V177M, A179D R211C, R211H,G216R, E217K,R235Q, G239S, R253P, S291L,L304R, M310T
SIFT	G20R, G74C, G74D, G74V, R76W, A83D, Y97C, R101L, R101L, R101Q, R101W, P104L, L107P, P126Q, V129M, G140E, R142Q, R149Q, R149W, L152M, R156H, R156C, V177M, A179D,R211C, R211H, A215T, G216R, E217K, T233I, R235Q, G239S,S291L, P297Q, L304R, M310T, A329V
phD-SNP	H15D, G20R, G74C, G74D, G74V, R76W, A83D, Y97C, R101L, R101L, R101Q, R101W, P104L, L107P, V129M, G140E, R142Q, R149Q, R149W, V177M, A179D, Q199P, R211C, R211H, A215T, G216R, E217K,T233I, R235Q, G239S, R253P, P274L, S291L, P297Q, L304R, A329V
Pmut	H15D,A83D,Y97C,R101L,R101W, G140E, R149W, R156C, E217K
Polyphen	G74D, V129M,R142Q, R149W, L152M, Q199P, T233I
SNAP	H15D,G20R, G74C, G74D, G74V, R76W, A83D, Y97C, R101L, R101L, R101Q, R101W, P104L, L107P, V129M, G140E, R149Q, R149W, L152M, R156H, R156C, V177M, A179D, Q199P, R211C, R211H, A215T, G216R, E217K, T233I, R235Q, G239S, R253P, P274L, S291L, P297Q, L304R, A329V

Table 9: Overall comparison of all the predictions shows that almost all (80%) the missense mutations affects the structure and function of the protein.

Mutation	**Stability change Prediction**	**Aggregation Prediction**	**Disorder Prediction**	**Tolerance Prediction**	**Conclusion (Missense mutation affect or do not affect the protein in structure or function)**
H15D	Stabilizes	50% ‘Aggregates’, 50% ‘Disaggregates’	Ordered	Intolerance	may or may not affect
G20R	Stabilizes	Disaggregates	Ordered	Intolerance	may or may not affect
G74C	De Stabilizes	Disaggregates	Ordered	Intolerance	Affects
G74D	De Stabilizes	Disaggregates	Ordered	Intolerance	Affects
G74V	De Stabilizes	Disaggregates	Ordered	Intolerance	Affects
R76W	De Stabilizes	Disaggregates	Ordered	Intolerance	Affects
A83D	De Stabilizes	50% ‘Aggregates’, 50% ‘Disaggregates’	Ordered	Intolerance	Affects
Y97C	De Stabilizes	50% ‘Aggregates’, 50% ‘Disaggregates’	Ordered	Intolerance	Affects
R101L	De Stabilizes	50% ‘Aggregates’, 50% ‘Disaggregates’	Ordered	Intolerance	Affects
R101L	Stabilizes	50% ‘Aggregates’, 50% ‘Disaggregates’	Ordered	Intolerance	may or may not affect
R101Q	Stabilizes	50% ‘Aggregates’, 50% ‘Disaggregates’	Ordered	Intolerance	may or may not affect
R101W	Stabilizes	50% ‘Aggregates’, 50% ‘Disaggregates’	Ordered	Intolerance	may or may not affect
P104L	De Stabilizes	Disaggregates	Ordered	Intolerance	Affects
L107P	De Stabilizes	Disaggregates	Ordered	Intolerance	Affects
P126Q	De Stabilizes	Disaggregates	Ordered	Tolerance	Affects
V129M	Stabilizes	Disaggregates	Ordered	Tolerance	Affects
G140E	De Stabilizes	Disaggregates	Ordered	Intolerance	Affects
R142Q	De Stabilizes	Disaggregates	Ordered	Tolerance	Affects
R149Q	De Stabilizes	Disaggregates	Ordered	Tolerance	Affects
R149W	De Stabilizes	Disaggregates	Ordered	Intolerance	Affects
L152M	Stabilizes	Disaggregates	Ordered	Tolerance	Affects
R156H	De Stabilizes	Disaggregates	Ordered	Intolerance	Affects
R156C	De Stabilizes	Disaggregates	Ordered	Intolerance	Affects
V177M	De Stabilizes	Disaggregates	Ordered	Intolerance	Affects

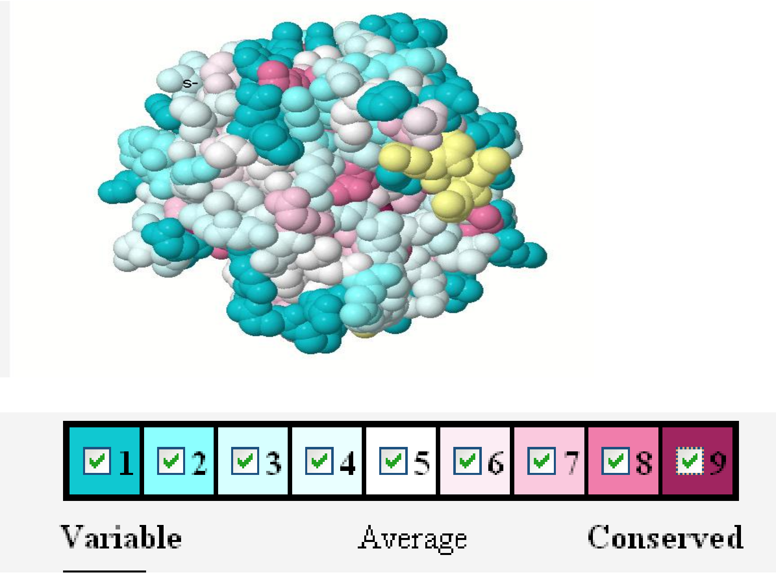
A179D	Stabilizes	Disaggregates	Ordered	Intolerance	Affects
Q199P	De stabilizes	50% ‘Aggregates’, 50% ‘Disaggregates’	Ordered	Tolerance	may or may not affect
R211C	Stabilizes	50% ‘Aggregates’, 50% ‘Disaggregates’	Ordered	Intolerance	may or may not affect
R211H	Stabilizes	Disaggregates	Ordered	Tolerance	Partially affects
A215T	De stabilizes	Disaggregates	Ordered	Intolerance	Affects
G216R	De stabilizes	Disaggregates	Ordered	Intolerance	Affects
E217K	De stabilizes	Disaggregates	Ordered	Intolerance	Affects
T233I	Stabilizes	Disaggregates	Ordered	Intolerance	Affects
R235Q	De stabilizes	Disaggregates	Ordered	Intolerance	Affects
G239S	De stabilizes	Disaggregates	Ordered	Intolerance	Affects
R253P	De stabilizes	Disaggregates	Ordered	Intolerance	Affects
P274L	De stabilizes	Disaggregates	Ordered	Tolerance	Affects
S291L	De stabilizes	Disaggregates	Ordered	Intolerance	Affects
P297Q	De stabilizes	Disaggregates	Ordered	Intolerance	Affects
L304R	De stabilizes	Disaggregates	Ordered	Intolerance	Affects
M310T	De stabilizes	Disaggregates	Ordered	Tolerance	Affects
A329V	De stabilizes	Disaggregates	Ordered	Intolerance	Affects

4.5 Sequence conservational Analysis

Disease causing mutations affects the sequence positions which are conserved in evolution, because these positions are essential for the structure and function of proteins. Invariance conservation, where all amino acids occurring at the corresponding position in a multiple sequence alignment are the same, conservation of physicochemical properties such as hydrophobicity or charge and the third one covariation, where upon mutation, a compensating mutation occurs at another sequence position.This study of conservational analysis is focused on the level of conservation of positions that are affected by mutations

a. Consurf results:

Human ADA has the PDB code 3IAR which we are using in this study as an input. Consurf calculations demonstrate the high level of conservation of the pore region as compared with the rest of the protein. 3IAR is presented as a space-filled model, and colored according to the conservation scores. The color coding bar shows the coloring scheme.



Highly conserved regions can be seen in purple color and they are buried.

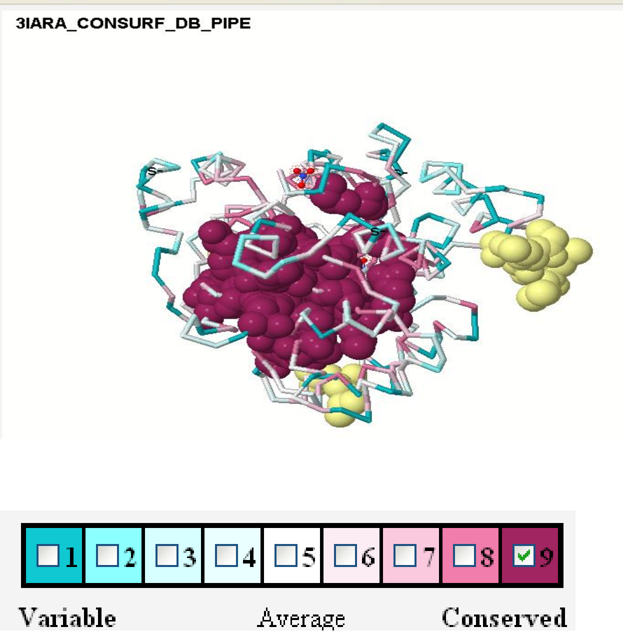
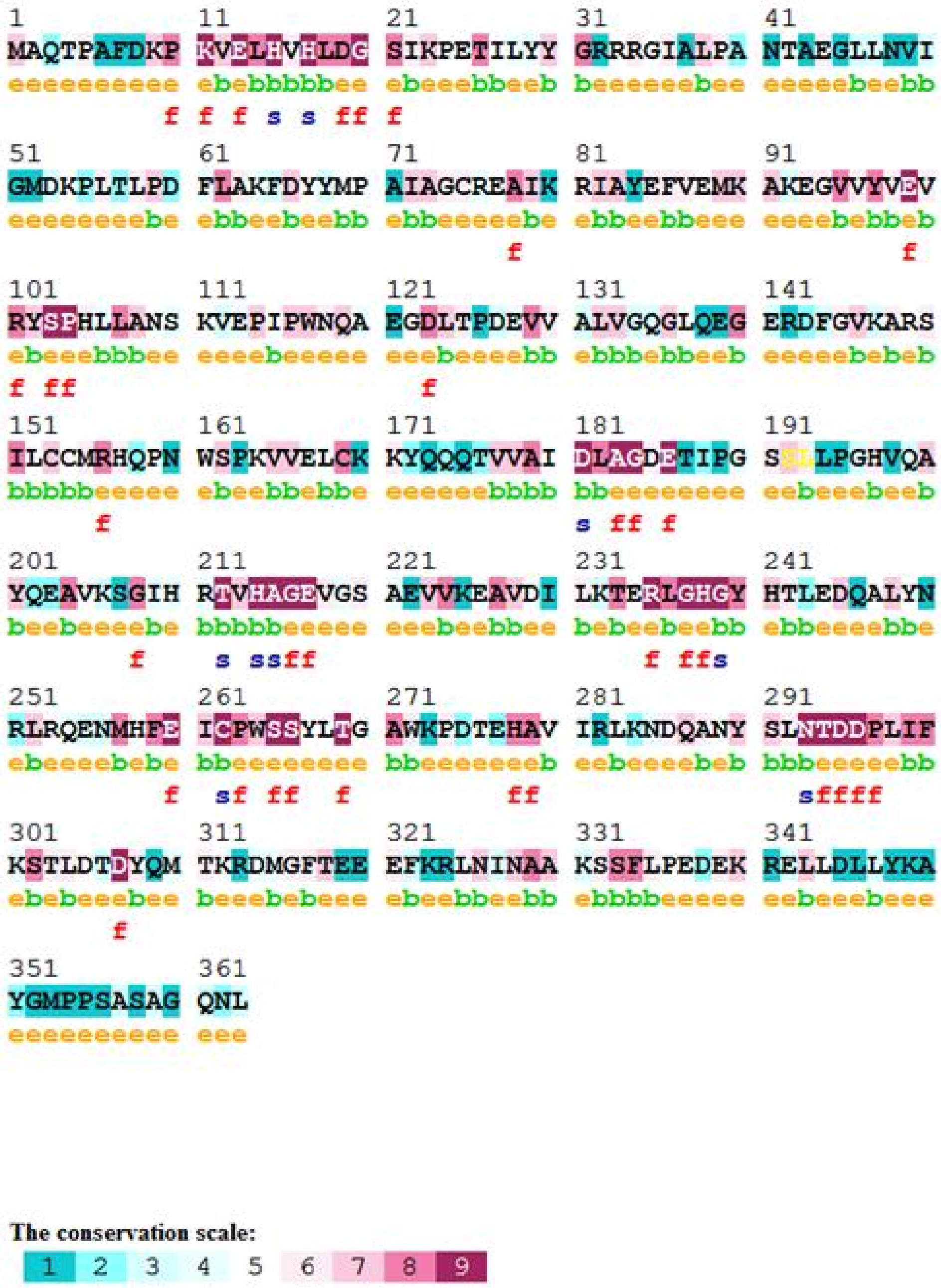


Figure 8: 3IAR structure showing different colors according to the conservation

b. ConSeq result:

ConSeq uses phylogenetic trees to calculate a conservation score for each residue. When our enzyme 3IAR is given, we got the output as color codes. Conservation scores are grouped into a nine color grade scale. The buried or exposed predicted status of each residue is marked in the first row below the sequence. Residues with color grades 8 and 9 are ‘exposed’ residues and predicted to be functional, whereas residues of color grade 9 are ‘buried’ and predicted to be structurally important. Both are indicated in the second row below the sequence, as ‘f’ and ‘s’, respectively.

In 3IAR having 363 residues, 34 residues are proved to be functionally important and 9 residues are proved to be structurally important residues.



```
1          11         21         31         41
MAQTPAFDKP KVELHVHLDG SIKPETILYY GRRRGIALPA NTAEGLLNVI
eeeeeeeeee ebebbbbbee ebeeebbeeb beeeeeebee eeeeebeebb
         f f f s s ff f
51         61         71         81         91
GMDKPLTLPD FLAKFDYYMP AIAGCREAIK RIAYEFVEMK AKEGVVYVEV
eeeeeeeebe ebbeebeebb ebbeeeeebe ebbeebbeee eeeebebbeb
                             f                      f
101        111        121        131        141
RYSPHLLANS KVEPIPWNQA EGDLTPDEVV ALVGQGLQEG ERDFGVKARS
ebeeebbbee eeeebeeeee eeebeeeebb ebbbebbeeb eeeeebebeb
f ff                    f
151        161        171        181        191
ILCCMRHQPN WSPKVVELCK KYQQQTVVAI DLAGDETIPG SSLLPGHVQA
bbbbbeeeee ebeebbebbe eeeeeebbbb bbeeeeeeee eebeeebeeb
     f                           s ff f
201        211        221        231        241
YQEAVKSGIH RTVHAGEVGS AEVVKEAVDI LKTERLGHGY HTLEDQALYN
beebeeeebe bbbbbeeeee eeebeebbee bebeebeebb ebbeeeebbe
       f    s ssff                     f ffs
251        261        271        281        291
RLRQENMHFE ICPWSSYLTG AWKPDTEHAV IRLKNDQANY SLNTDDPLIF
ebeeeebebe bbeeeeeeee bbeeeeeeeb eebeeeebeb bbbeeeeebb
         f  sf ff   f         ff               sffff
301        311        321        331        341
KSTLDTDYQM TKRDMGFTEE EFKRLNINAA KSSFLPEDEK RELLDLLYKA
ebebeeebee beeebebeee ebeebbeebb ebbbbeeeee eebeeebeee
      f
351        361
YGMPPSASAG QNL
eeeeeeeeee eee
```

The conservation scale:

1 2 3 4 5 6 7 8 9

Figure 9: ConSeq results of ADA (PDB code: 3IAR).
The sequence of the ADA protein is displayed with the evolutionary rates at each site colour-coded onto it. The residues of the query sequence are numbered starting from 1. The first row below the sequence lists the predictedburial status of the site (i.e. 'b'—buried versus 'e'—exposed). The second row indicates residues predicted to be structurally and functionally important: 's' and 'f', respectively.

c. ConSSeq:

This service presents PDB file sequence and consensus sequence (as found in HSSP) which is colored by conservation, color coded graphic bars of relative entropy, information about residues present in other humongous sequences, with their respective frequency. 3IAR is given as a query sequence to get the output. I summarized the results, of which 11 residues have higher conservation of which histidine, arginine, glutamic acid, arginine of positions 15, 104, 217 and 235, respectively, have 100% conservation and 7 residues of which proline and methionine has very low conservation %.

Table 10: Most Conserved residues

wild AA, position	Entropy	Conservation (%)
H 15	0.00	100
G 20	0.01	99
Y 97	0.02	99
R 101	0.04	98
R 104	0.01	100
R 156	0.10	95
A 215	0.05	98
G 216	0.02	99
E 217	0.00	100
R 235	0.01	100
G 239	0.10	94

Table 11: Least conserved (variable) residues

wild AA, position	Entropy	Conservation (%)
L 107	0.43	10
P 126	0.63	9
Q 199	0.66	10
R 211	0.58	11
P 274	0.56	6
M 310	0.66	6

d. Matrix Plot

MatrixPlot displays high quality matrix plots of any type of data given in a simple format. It takes the sequences having three dimensional co-ordinates. The Interatomic distances between the C-α atoms are indicated by the scales, which measures distance in angstrom. The outer bar indicates the four structural domains and a linker region. The inner bar indicates the secondary structure.. The scale of the matrix indicates the physical distances in Å.

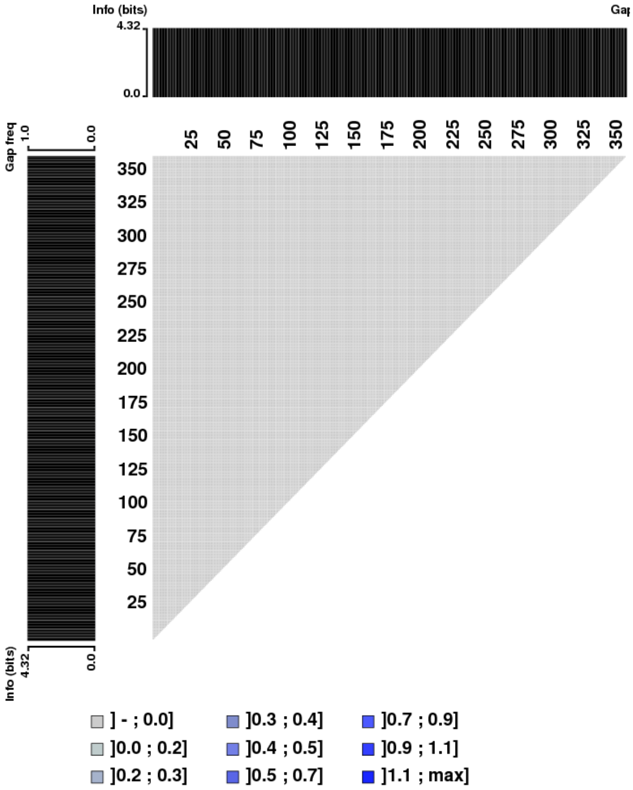


Figure 10: A combined plot of sequence information and gap frequencies is displayed along the edges.

(*The colors on the gif file are limited, they might be distorded, hence the black and white figure)

4.6 Structural considerations

The structures were visualized and mutated proteins are modeled by PyMOL, version 0.99 [DeLano, 2002]. The satisfied conformations for a mutated side chain have a total score of above -1.0 [Lovell *et al.*, 2000]. High score indicates that the side chain fits well in to the structure and vice versa. The created structure can be verified by MolProbity and bond angles are investigated using the PROBE procedure.

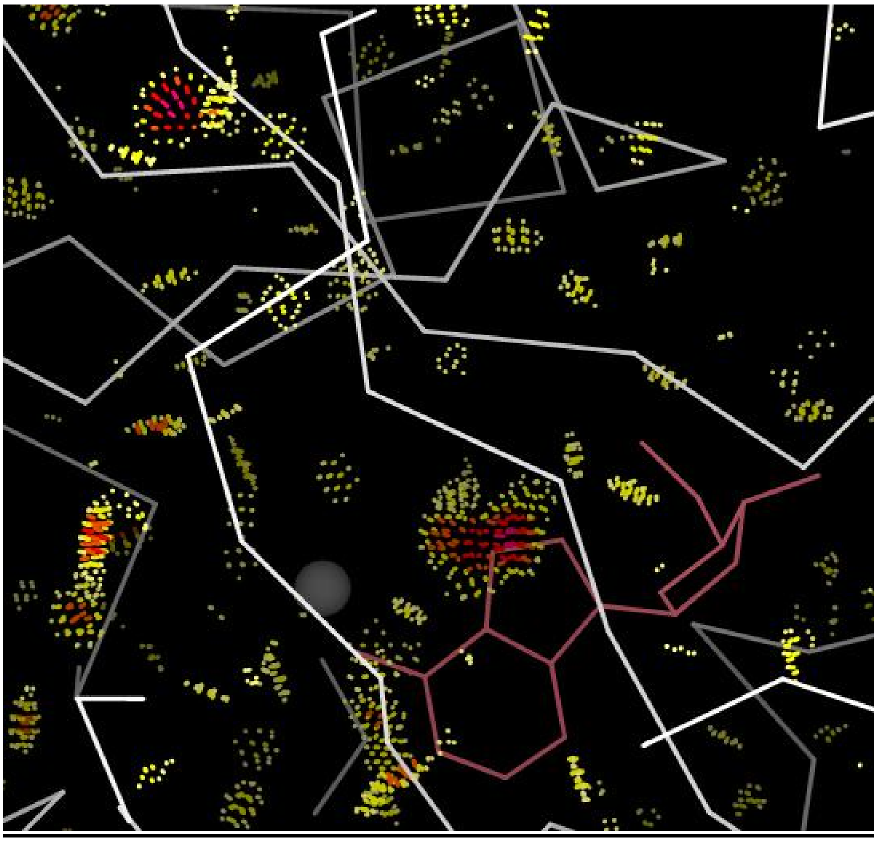


Figure 12: Substitution of R149 by Q causes small and bad overlap in spite of a positive probe value (7.850).

5. Discussion

Bioinformatics methods are very useful for the analysis of missense mutations, but every method used here checks the mutation from a different perspective whether it is about the conserved position, folding, stability, disorder or structure of the protein. The user must be familiar with the theory and limitations of each analysis and predictor. Careful choice and understanding of the predictors is an important aspect. From this study it was seen that no single predictor can reveal structural, functional, or pathogenicity of a protein completely. However, in combination they provide relatively better results. Thus, PON-P comes in to picture which combines features of several programs and the negativity of each predictor is compensated when combined.

Stability changes can be studied experimentally but is laborious, time consuming and often also costly. Therefore reliable computational prediction methods would prove valuable. Sorting out the different kinds of programs, identifying those that are capable of providing reliable results benefits not only the users but also for the development of new advanced computational tools. In this study I have used seven programs under stability change prediction, four programs under aggregation, ten programs under Disorder and seven programs under tolerance predictions to analyze the 40 missense mutations identified in ADA.

The accurate estimation of protein stability, aggregation, disorder, tolerance prediction changes induced by mutations still remains a significant challenge. The study indicates that there is great potential in prediction tools to be used efficiently in analysis of mutation effects and that this kind of evaluation is needed in order to further develop the prediction methods.

Reliability of different predictions promotes the usage of several methods that are based on different concepts, different physicochemical parameters, or different implementations. Many predictors give the results based on the different parameters. So, qualitative analysis is one of the methods to draw the conclusions. If many predictors of a certain prediction pose to increase in disorder or decrease in stability and the mutation causes the pathogenicity in the protein.

6. References

Ashkenazy, H., Erez, E., Martz, E., Pupko, P., and Ben-Tal, N. (2010). ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucl. Acids Res.,* 38(2), W529-W533.

Aldrich, M.B., Blackburn, M.R., Kellems, R. E. (2000). The importance of adenosine deaminase for lymphocyte development and function. *Biochem Biophys Res Commun.* , 272(2), 311–315.

Alessandro, V., Oscar B., Gianluca P., and Silvio C. E., Tosatto. (2006). Spritz: a server for the prediction of intrinsically disordered regions in protein sequences using kernel machines. *Nucleic Acids Research,* 34(2), W164–W168.

Amberger, J., Bocchini, C.A., Scott, A.F., Hamosh, A., McKusick's Online Mendelian Inheritance in Man (OMIM). *Nucleic Acids Research*., 37(1), D793–6.

Arredondo-Vega, F.X. , Santisteban, I. , Kelly S., Schlossman, C.M., Umetsu, D.T., Hershfield, M.S. (1994). Correct splicing despite mutation of the invariant first nucleotide of a 5' splice site: a possible basis for disparate clinical phenotypes in siblings with adenosine deaminase deficiency. *Am J Hum Genet.*, 54(5) 820–30.

Bourhis, J.M., Canard. B., and Longhi, S. (2007). Predicting protein disorder and induced folding: from theoretical principles to practical applications. *Curr Protein Pept Sci.,* 8(2), 135-49.

Bolon, D.N., Marcus, J.S., Ross, S.A., Mayo, S.L. (2003). Prudent modeling of core polar residues in computational protein design. *J Mol Biol*., 329(3) 611–622.

Bao, L., and Cui, Y. (2005). Prediction of the phenotypic effects of nonsynonymous single nucleotide polymorphisms using structural and evolutionary information. *Bioinformatics,* 21(10), 2185-90.

Benveniste, P., Cohen, A. (1995). p53 expression is required for thymocyte apoptosis induced by adenosine deaminase deficiency. *Proc Acad Sci.,* 92(18) , 8373-8377

Bromberg, Y., and Rost, B. (2007). SNAP: predict effect of nonsynonymous polymorphisms on function. *Nucleic Acids Res.*, 35(11) , 3823-35.

Buckley, R.H., Schiff, R.I., Schiff, S.E., Markert, M.L, Williams, L.W. (1997). Human severe combined immunodeficiency: Genetic, phenotypic, and functional diversity in one hundred eight infants. *J Pediatr*., 130(3), 378–387.

Burke, D.F., Worth, C.L., Priego, E.M., Cheng, T, Smink, L.J., Todd, J.A., Blundell TL.(2007). Genome bioinformatic analysis of nonsynonymous SNPs. *Bioinformatics* 8:301.

Capriotti, E., Fariselli, P., and Casadio, R. (2004). A neural network-based method for predicting protein stability changes upon single point mutations. *Bioinformatics*, 20(1) I63–I68.

Capriotti, E., Calabrese, R., and Casadio, R. (2006). Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. *Bioinformatics*, 22(22), 2729-2734.

Calabrese, R. E., Capriotti, P., Fariselli, P.L., Martelli and Casadio R. (2009). Functional annotations improve the predictive score of human disease related mutations in proteins. *Hum Mutat* 30(8), 1237-44.

Chasman, D., Adams, R.M. (2001). Predicting the Functional Consequences of Non-synonymous Single Nucleotide Polymorphisms: Structure-based Assessment of Amino Acid Variation. *J Mol Biol*, 307(2), 683-706.

Cheng, J, Randall, A., and Baldi, P. (2006). Prediction of Protein Stability Changes for Single-Site Mutations Using Support Vector Machines. *Proteins*, 62(4), 1125-1132.

Chiti, F., Dobson, C.M. (2006). Protein misfolding, functional amyloid, and human disease. *Annu. Rev. Biochem.*,75, 333–66.

Carine, B., Fabian, G., Josef, R., Inbal, P., Tal, P., Piero F., Rita C., and Nir, Ben-Tal. (2004). ConSeq: the identification of functionally and structurally important residues in protein sequences. *Bioinformatics*., 20(8), 1322-4.

Daggett, V., and Fersht, A.R. (2003). Is there a unifying mechanism for protein folding? *Trends Biochem. Sci.,* 28(1), 18-25.

Dahiyat, BI.(1999). In silico design for protein stabilization. *Curr Opin Biotech.*,10(4), 387–390.

DeGrado, W.F. (1999). *De novo* design and structural characterization of proteins and metalloproteins. *Ann Rev Biochem*, 68, 779–819.

Dosztányi, Z., Sandor, M., Tompa, P., and Simon, I. (2007). Prediction of protein disorder at the domain level. *Curr Protein Pept Sci.,* 8(2), 161-171.

Dobson, R.J, Munroe, P.B, Caulfield, M.J, and Saqi, M.A.S. (2006). Predicting deleterious nsSNPs: an analysis of sequence and structural attributes. *BMC Bioinformatics,* 7, 217.

Dunker , A.K., Silman, I., Uversky, V.N., and Sussman, J.L. (2008). Function and structure of inherently disordered proteins. *Curr Opin Struct Biol.,* 18(6), 756-64.

Dosztányi, Z., Magyar, C, Tusnády, G and Simon, I. (2003). SCide: identification of stabilization centers in proteins. *Bioinformatics.,* 19(7), 899-900.

Dosztányi, Z., Csizmok, V., Tompa, P., and Simon, I. (2005). IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics .,21*(16), 3433-3434.

Erel, O., Kocyigit, A., Gurel, M.S, Bulut, V., Seyrek, A. *et al*. Adenosine Deaminase activities in sera, lymphocytes and granulocytes in patients with cutaneous Leishmaniasis. *Mem Inst OswaldoCruz, Rio de Janeiro*, 93(4), 491-494.

Ferrer-Costa, C., Gelpí, JL., Zamakola, L., Parraga, I., de la Cruz, X., and Orozco, M.(2005). PMUT: a web-based tool for the annotation of pathological mutations on proteins. *Bioinformatics*, 21(14), 3176-3178

Ferrer-Costa, C., Orozco, M., de la Cruz, X. (2002). Characterization of disease-associated single amino acid polymorphisms in terms of sequence and structure properties. *J Mol Biol,* 315(4), 771-786.

Fink, A.L. (1998). Protein aggregation: folding aggregates, inclusion bodies and amyloid. *Fold Des.,*3(1), R9 –23.

Fowler, D.M. et al. (2006). Functional amyloid formation within mammalian tissue. *PLoS Biol.* 4(1), 100–107.

Friedberg, I., Jaroszewski, L., Ye, Y., Godzik, A. (2004). The interplay of fold recognition and experimental structure determination in structural genomics. *Curr Opin Struct Biol*, 14(3), 307–312.

Galzitskaya, O.V., Garbuzynskiy, S.O., and Lobanov, M.Y. (2006) . FoldUnfold: web server for the prediction of disordered regions in protein chain. *Bioinformatics,* 22(23), 2948-2949.

Garner, E., Cannon, P., Romero, P., Obradovic, Z., and Dunker, A. K. (1998). Predicting disordered regions from amino acid sequence: Common themes despite differing structural characterization. *Genome Inform. Ser. Workshop Genome Inform.,* 9:201-213.

Garner, E., Romero, P., Dunker, A. K., Brown, C., and Obradovic, Z. (1999). Predicting binding regions within disordered proteins. *Genome Inform. Ser. Workshop Genome Inform.,* 10:41-50.

Giblett, E.R, Anderson, J.E., Cohen, F., Pollara, B., Meuwissen , H.J. (1972). Adenosine deaminase deficiency in two patients with severely impaired cellular immunity. *Lancet,* 2(7786), 1067–9.

Gilis, D., and Rooman, M. (2000). PoPMuSiC, an algorithm for predicting protein mutant stability changes: application to prion proteins. *Protein Eng.*, 13(12) , 849–856.

Graf, W.D. (2000). Can Bioinformatics Help Trace the Steps from Gene Mutation to Disease? *Neurology* , 55(3), 331–3.

Guerois, R., Nielsen, J.E., and Serrano, L. (2002). Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J. Mol. Biol.,* 320, (2), 369–387.

Higa, R.H., Montagner, A.J, Togawa, R.C., Kuser, M.E.B., Yamagishi, A. L., Mancini, G. Pappas Jr., Miura, R.T., Horita, L.G., and G. Neshich. (2004). ConSSeq: a web-based application for analysis of amino acid conservation based on HSSP database and within context of structure. *Bioinformatics,* 20(12), 1983-1985.

Hirschorn, R. (1999) . Immunodeficiency disease due to deficiency of adenosine deaminase. In: Ochs HD, Smith CIE, Puck JM (eds) Primary Immunodeficiency iseases: A molecular and Genetic approach. *Oxford University Press, New York*, 121-39.

Hirschhorn, R., Ellenbogen, A., Tzall, S. (1992). Five missense mutations at the adenosine deaminase locus (ADA) detected by altered restriction fragments and their frequency in ADA-patients with severe combined immunodeficiency (ADA-SCID). *Am J Med Genet., 42* (2), 201-7.

Hershfield, M.S., and Mitchell, B.S. (2001). Immunodeficiency diseases caused by adenosine deaminase deficiency and purine nucleoside phosphorylase deficiency. In: Scriver C. R., A. L. Beaudet, M. S. Sly, and D. S. Valle (eds): The metabolic and molecular basis of inherited disease. McGraw-Hill, 2585-2625.

Hershfield, M.S. (2004). Combined immune deficiencies due to purine enzyme defects, in: Stiehm ER, Ochs HD, Winkelstein J, (eds). Immunologic Disorders in Infants and children. WB Saunders, Philadelphia, 480-504.

Hirschhorn, R., Tzall, S., Ellenbogen, A. (1990). Hot spot mutations in adenosine deaminase deficiency. *Proc Natl Acad Sci.,* 87(16), 6171–5.

Ishida, T., and Kinoshita, K. (2008). Prediction of disordered regions in proteins based on the meta approach. *Bioinformatics,* 24(11), 1344-1348.

Ivanova, M.I., Sawaya, M.R., Gingery, M., Attinger, A., Eisenberg, D. (2004). An amyloid-forming segment of {beta}2-microglobulin suggests a molecular model for the fibril. *Proc Natl Acad Sci.,* 101(29), 10584–10589.

Jiang, C., Hong, R., Horowitz, S.D., Kong, X., Hirschhorn, R. (1997). An adenosine deaminase (ADA) allele contains two newly identified deleterious mutations (Y97C and L106V) that interact to abolish enzyme activity. *Hum Mol Genet., 6* (13), 2271-8.

Khemtemourian, L., Killian, J.A., Hoppener, J.W., and Engel, M.F. (2008). Recent insights in islet amyloid polypeptideinduced membrane disruption and its role in cell death in type 2 diabetes mellitus. *Exp Diabetes Res.,* 421287.

Keegan, L.P., Leroy, A., Sproul, D., O'Connell, M.A. (2004). Adenosine deaminases acting on RNA (ADARs): RNA-editing enzymes. *Genome Biol.,* **5** (2), 209.

Kiel, C., Serrano, L., and Herrmann, C. (2004). A detailed thermodynamic analysis of ras/effector complex interfaces. *J. Mol. Biol.,* 340(5), 1039–1058.

Kurgan, L., Cios, K., and Chen, K. (2008). Accurate prediction of protein structural class for sequences of twilight-zone similarity with predicting sequences. *BMC Bioinformatics,* 9, 226.

Kwasigroch, J.M, Gilis, D, Dehouck, Y, and Rooman M. (2002) . PoPMuSiC, rationally designing point mutations in protein structures. *Bioinformatics*, 18(12), 1701-1702.

Lacroix, E., Viguera, A.R., Serrano, L. (1998). Elucidating the folding problem of alpha-helices: local motifs, long-range electrostatics, ionicstrength dependence and prediction of NMR parameters. *J Mol Biol.,* 284(1), 173-91.

Liang-Tsung Huang, M. Michael Gromiha, and Shinn-Ying Ho. (2007). Interpretable decision tree based method for predicting protein stability changes upon mutations. *Bioinformatics,* 23(10), 1292-1293.

Linding, R., Russell, R.B., Neduva, V., and Gibson, T.J. (2003). GlobPlot: exploring protein sequences for globularity and disorder. *Nucl.Acids Res.,* 31(13), 3701-3708.

Lappalainen, I., Thusberg, J., Shen, B., Vihinen, M. (2008). Genome wide analysis of pathogenic SH2 domain mutations. *Proteins,* 72(2), 779–792.

Looger, L.L., Dwyer, M.A., Smith, J.J., Hellinga, H.W. (2003). Computational design of receptor and sensor proteins with novel functions. *Nature,* 423, 185–190.

MacCallum, R.M. : Order/disorder prediction with self organizing maps.

Mendes, J., Guerois, R., Serrano, L. (2002). Energy estimation in protein design. *Curr Opin Struct Biol.,* 12(4), 441–446.

Meytal, L., Itay, M., Yossi, R., Fabian, G., Eric, M., Tal, P., and Nir, Ben-Tal. (2005) ConSurf: the projection of evolutionary conservation scores of residues on protein structures. *.Nucleic Acids Res.,* 33, W299-302.

Mooney, S.D., and Klein, T.E. (2002). The Functional Importance of Disease-Associated Mutation. *BMC Bioinformatics,* 3, 24.

Munoz, V., Serrano, L. (1997). Development of the multiple sequence approximation within the AGADIR model of alpha-helix formation: comparison with zimm-bragg and lifson-roig formalisms. *Biopolymers,* 41(5), 495-509.

Online 'Mendelian Inheritance in Man' (OMIM) 102700.

Ozsahin, H., Arredondo-Vega, F.X., Santisteban, I., Fuhre, H., Tuchschmid, P., Jochum, W., Aguzzi, A., Lederman, H.M., Fleischman, A., Winkelstein, J.A., Seger, R.A., Hershfield, M.S. (1997). Adenosine deaminase deficiency in adults. *Blood,* 89(8), 2849–55.

Pauline, C. Ng, and Henikoff , S. (2001). Predicting deleterious amino acid substitutions. *Genome Res,* 11(5), 863-874.

Persico, A.M., Militerni, R., Bravaccio C., *et al.* (2000). Adenosine deaminase alleles and autistic disorder: case-control and family-based association studies . *Am. J. Med. Genet.,* **96** (6), 784-90.

Pitera, J.W., and Kollman, P.A. (2000). Exhaustive mutagenesis in silico:multicoordinate free energy calculations on proteins and peptides. *Proteins*, 41(3), 385-397.

Prevost, M., Wodak, S.J., Tidor, B., and Karplus, M. (1991). Contribution of the hydrophobic effect to protein stability: analysis based on simulations of the I1e-96-Ala mutation in baranse. *Proc.Natl Acad.Sci .,* 88(23), 10880-10884.

Pauline C. Ng., and Henikoff, S. (2003). SIFT: predicting amino acid changes that affect protein function. *Nucl.Acids Res.,* 31(13), 3812-3814.

Pajunen, M., Turakainen, H., Poussu, E., Peränen, J., Vihinen, M., Savilahti, H. (2007). High-precision mapping of protein protein interfaces: an integrated genetic strategy combining en masse mutagenesis and DNA-level parallel analysis on a yeast two-hybrid platform. *Nucleic Acids Res*., 35(16), e103.

Ramensky, V., Bork. P., and Sunyaev, S. (2002). Human nonsynonymous SNPs: server and survey. *Nucleic Acids Res,* 30(17), 3894-3900.

Receveur-Bre´chot, V., Bourhis, J.M, Uversky, V.N, Canard, B., Longhi, S. (2006). Assessing protein disorder and induced folding. *Proteins*, 62(1), 24–45.

Romero, P., Obradovic, Z., Kissinger, C. R., Villafranca, J. E., and Dunker, A. K. (1997). Identifying disordered proteins from amino acid sequences. *Proc. IEEE Int. Conf. Neural Networks,* 1: 90-95.

Saven J.G. (2002). Combinatorial protein design. *Curr Opin Struct Biol*., 12(4), 453–458.

Street, A.G., Mayo, S.L. (1999). Computational protein design. *Struct Fold Des*., 7, R105–R109.

Santisteban, I., Arredondo-Vega, F.X., Kelly, S., Debre, M., Fischer, A., Pérignon, J.L., Hilman, B., elDahr, J., Dreyfus, D.H., Gelfand, E.W. (1995). Four new adenosine deaminase mutations, altering a zinc-binding histidine, two conserved alanines, and a 5' splice site. *Hum Mutat.,* 5(3), 243-250.

Santisteban, I., Arredondo-Vega, F.X., Kelly, S., Mary, A., Fischer, A., Hummell, D.S., Lawton, A., Sorensen, R.U., Stiehm, E.R., Uribe, L., et al. (1993). Novel splicing, missense, and deletion mutations in seven adenosine deaminase-deficient patients with late/delayed onset of combined immunodeficiency disease. Contribution of genotype to phenotype. *J Clin Invest.*, 92(5), 2291–302.

Schymkowitz, J., Borg, J., Stricher, F., Nys R., Rousseau, F., and Serrano, L. (2005). The FoldX web server: an online force field. *Nucleic Acids Res.*, 33, W382-8

Steward, R.E, MacArthur, M.W., Laskowski, R.A., Thornton, J.M. (2003). Molecular basis of inherited diseases: a structural perspective. *Trends Genet.*, 19(9), 505–513.

Smith A. (2003). Protein misfolding. *Nature.*, 426**,** 883 –8883.

Sunyaev, S., *et al.* (2001) . Prediction of deleterious human alleles. *Hum Mol Genet.*, 10(6), 591-7.

Tavtigian, S., Greenblatt, M., Lesueur, F., Byrnes, G. (2008). In silico analysis of missense substitutions using sequence alignment based methods. *Hum Mutat.*, 29(11), 1327–1336.

Thusberg, J., and Vihinen, M. (2009). Pathogenic or Not? And If So, Then How? Studying the Effects of Missense Mutations Using Bioinformatics Methods. *Hum Mutat.*, 30(5), 703-14.

Thusberg, J., Vihinen, M. (2006). Bioinformatic analysis of protein structure–function relationships: case study of leukocyte elastase (ELA2) missense mutations. *Hum Mutat.*, 27(12), 1230–1243.

Thusberg, J., Vihinen, M. (2007). The structural basis of hyper IgM deficiency-CD40L mutations. *Protein Eng Des Sel.*, 20(3), 133–141.

Tompa, P. (2002). Intrinsically unstructured proteins. *Trends Biochem Sci*., 27(10), 527–533.

Topham, C.M., *et al.* (1997). Prediction of the stability of protein mutants based on structural environmental-dependent amino acid substitution and propensity tables. *Protein Eng.*, 10(1), 7-21.

Uversky, V.N., Oldfield, C.J., and Dunker, A.K. (2008). Intrinsically disordered proteins in human diseases: introducing the D2 concept. *Annu Rev Biophys.*, 37, 215-46

Ventura, S., Zurdo, J., Narayanan, S., Parreno, M., Mangues, R., Reif, B., Chiti, F., Giannoni, E., Dobson, C.M., Aviles, F.X., Serrano, L. (2004). Short amino acid stretches can mediate amyloid formation in globular proteins: the Src homology 3 (SH3) case. *Proc Natl Acad Sci.*, 101, 7258 –77263.

Vitkup, D., Sander, C., Church, G.M. (2003). The amino-acid mutational spectrum of human genetic disease. *Genome Biol.,* 4(11), R72.

Vihinen, M., Arredondo-Vega, F.X., Casanova, J.L., Etzioni, A., Giliani, S., Hammarstrom, L., Hershfield, M.S, Heyworth, P.G, Hsu, AP, Lahdesmaki, A., Lappalainen, I., Notarangelo, L.D., Puck, J.M., Reith, W., Roos, D. Schumacher, R.F., Schwarz, K., Vezzoni, P., Villa, A., Valiaho, J., Smith, C.I. (2001). Primary immunodeficiency mutation databases. *Adv Genet.,* 43, 103–88.

Valerio, D., McIvor, R.S., Williams, S.R., Duyvesteyn, M.G., van Ormondt, H., *et al.* (1984). Cloning of human adenosine deaminase cDNA and expression in mouse cells. *Gene*., 31(1-3), 147–53.

Voigt, C.A., Kauffman, S., and Wang, Z.G. (2000). Rational evolutionary design: the theory of *in vitro* protein evolution. *Adv Protein Chem.,* 55, 79-160.

Wang, Z., Moult, J. (2001). SNPs, protein structure, and disease. *Hum Mutat* ., 17(4), 263-270.

Worth, C.L, Bickerton, G.R, Schreyer, A., Forman, J.R., Cheng, T.M., Lee, S., Gong, S., Burke, D.F, Blundell, T.L. (2007). A structural bioinformatics approach to the analysis of nonsynonymous single nucleotide polymorphisms (nsSNPs) and their relation to disease. *J Bioinform Comput Biol.,* 5(6), 1297–1318.

Yang, D.R, Huie, M.L, Hirschhorn, R. (1994). Homozygosity for a missense mutation (G20R) associated with neonatal onset adenosine deaminase-deficient severe combined immunodeficiency (ADA-SCID). *Clin Immunol Immunopathol.,* 70 (2), 171-5.

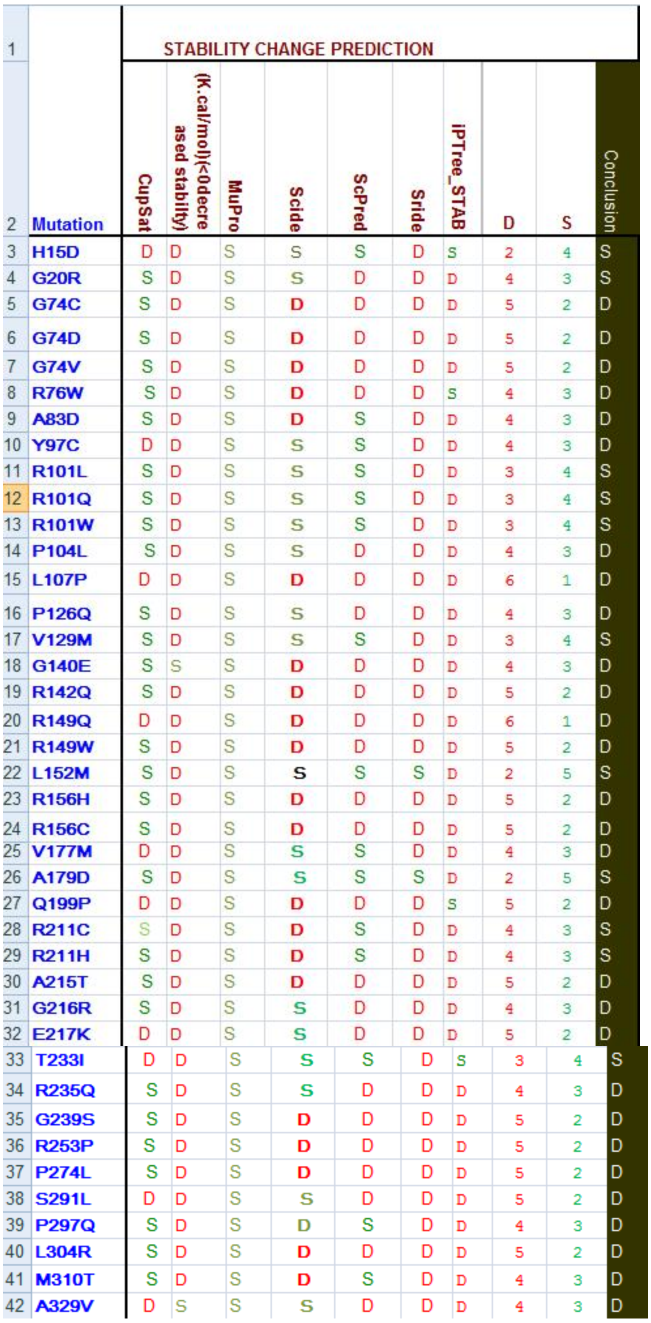
Yang, Z.R., Thomson, R., McNeil P., and Esnouf, R.M. (2008). RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. *Bioinformatics*., 21(16), 3369-3376.

Zhou, H., and Zhou,Y. (2002) . Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structute selection and stability prediction. *Protein Sci.*, 11(11), 2714-2726.

7. APPENDICES

Table :1 Output of stability predictors that shows the affect of missense mutations. (‘D’ = destabilizes;’S’ = stabilizes)

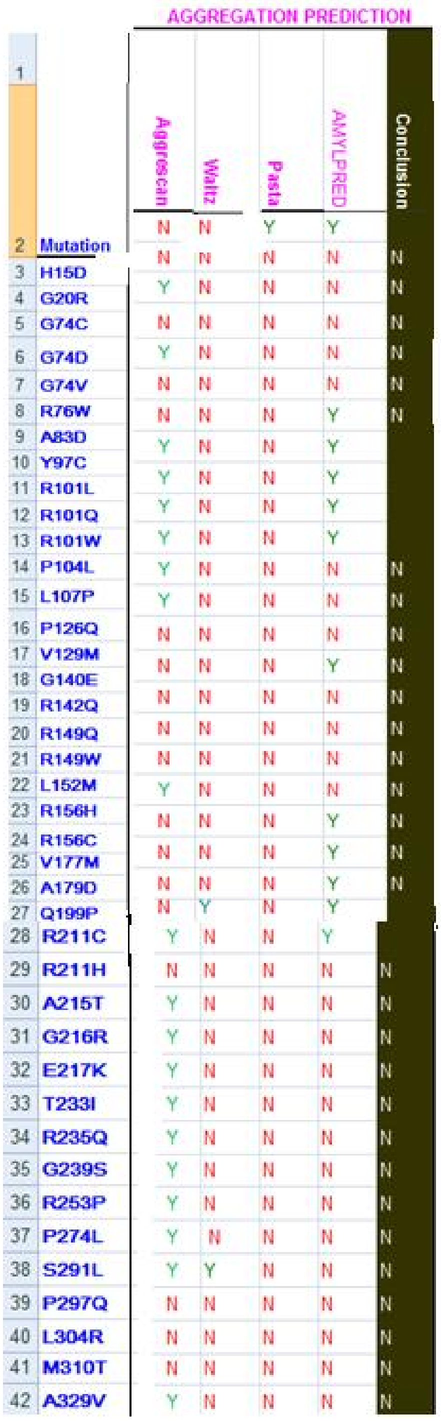
A. Stability change prediction



1		STABILITY CHANGE PREDICTION									
2	Mutation	CupSat	(K.cal/mol)(<0decre ased stability)	MuPro	Scide	ScPred	Sride	iPTree_STAB	D	S	Conclusion
3	H15D	D	D	S	S	S	D	S	2	4	S
4	G20R	S	D	S	S	D	D	D	4	3	S
5	G74C	S	D	S	D	D	D	D	5	2	D
6	G74D	S	D	S	D	D	D	D	5	2	D
7	G74V	S	D	S	D	D	D	D	5	2	D
8	R76W	S	D	S	D	D	D	S	4	3	D
9	A83D	S	D	S	D	S	D	D	4	3	D
10	Y97C	D	D	S	S	S	D	D	4	3	D
11	R101L	S	D	S	S	S	D	D	3	4	S
12	R101Q	S	D	S	S	S	D	D	3	4	S
13	R101W	S	D	S	S	S	D	D	3	4	S
14	P104L	S	D	S	S	D	D	D	4	3	D
15	L107P	D	D	S	D	D	D	D	6	1	D
16	P126Q	S	D	S	S	D	D	D	4	3	D
17	V129M	S	D	S	S	S	D	D	3	4	S
18	G140E	S	S	S	D	D	D	D	4	3	D
19	R142Q	S	D	S	D	D	D	D	5	2	D
20	R149Q	D	D	S	D	D	D	D	6	1	D
21	R149W	S	D	S	D	D	D	D	5	2	D
22	L152M	S	D	S	S	S	S	D	2	5	S
23	R156H	S	D	S	D	D	D	D	5	2	D
24	R156C	S	D	S	D	D	D	D	5	2	D
25	V177M	D	D	S	S	S	D	D	4	3	D
26	A179D	S	D	S	S	S	S	D	2	5	S
27	Q199P	D	D	S	D	D	D	S	5	2	D
28	R211C	S	D	S	D	S	D	D	4	3	S
29	R211H	S	D	S	D	S	D	D	4	3	S
30	A215T	S	D	S	D	D	D	D	5	2	D
31	G216R	S	D	S	S	D	D	D	4	3	D
32	E217K	D	D	S	S	D	D	D	5	2	D
33	T233I	D	D	S	S	S	D	S	3	4	S
34	R235Q	S	D	S	S	D	D	D	4	3	D
35	G239S	S	D	S	D	D	D	D	5	2	D
36	R253P	S	D	S	D	D	D	D	5	2	D
37	P274L	S	D	S	D	D	D	D	5	2	D
38	S291L	D	D	S	S	D	D	D	5	2	D
39	P297Q	S	D	S	D	S	D	D	4	3	D
40	L304R	S	D	S	D	D	D	D	5	2	D
41	M310T	S	D	S	D	S	D	D	4	3	D
42	A329V	D	S	S	S	D	D	D	4	3	D

Table : 2 Output of aggregation predictors that shows the affect of missense mutations.
('Y' = yes; 'N' = no)

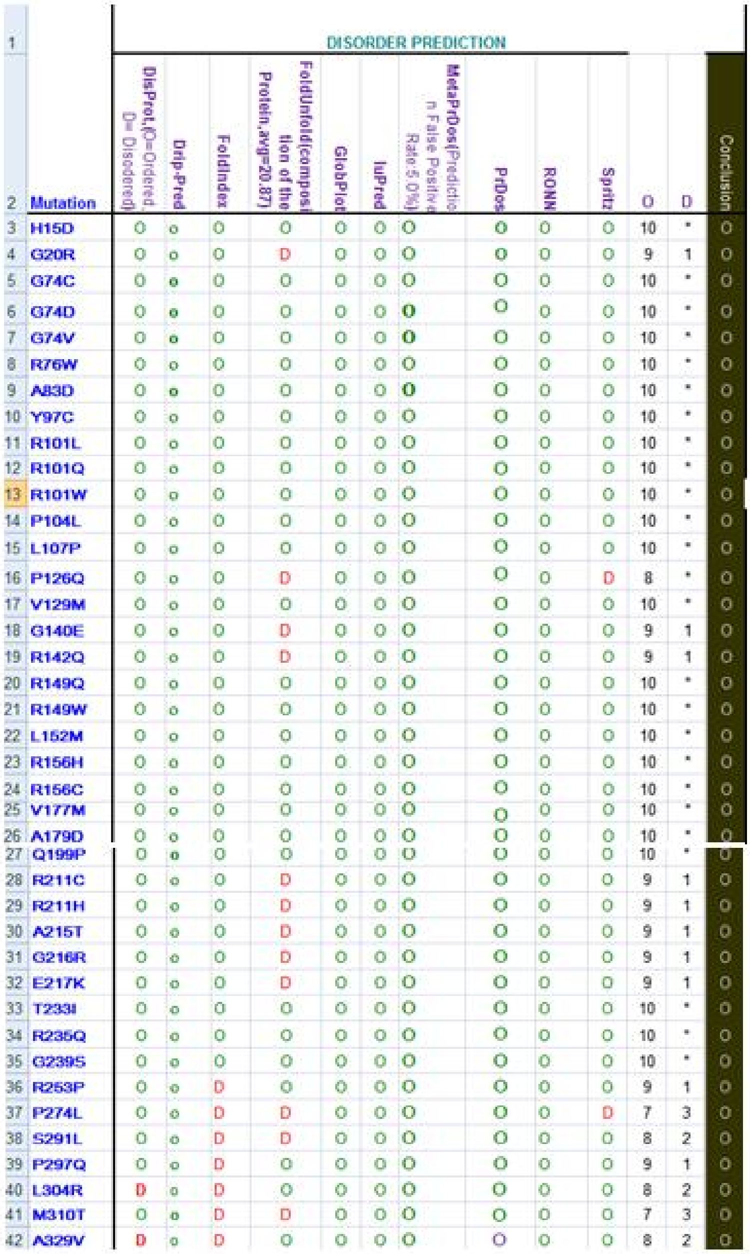
B. Aggregation Prediction



1		AGGREGATION PREDICTION				
2	Mutation	Aggrescan	Waltz	Pasta	AMYLPRED	Conclusion
		N	N	Y	Y	
3	H15D	N	N	N	N	N
4	G20R	Y	N	N	N	N
5	G74C	N	N	N	N	N
6	G74D	Y	N	N	N	N
7	G74V	N	N	N	N	N
8	R76W	N	N	N	Y	N
9	A83D	Y	N	N	Y	
10	Y97C					
11	R101L	Y	N	N	Y	
12	R101Q	Y	N	N	Y	
13	R101W	Y	N	N	Y	
14	P104L	Y	N	N	N	N
15	L107P	Y	N	N	N	N
16	P126Q	N	N	N	N	N
17	V129M	N	N	N	Y	N
18	G140E					
19	R142Q	N	N	N	N	N
20	R149Q	N	N	N	N	N
21	R149W	N	N	N	N	N
22	L152M	Y	N	N	N	N
23	R156H	N	N	N	Y	N
24	R156C					
25	V177M	N	N	N	Y	N
26	A179D	N	N	N	Y	N
27	Q199P	N	Y	N	Y	
28	R211C	Y	N	N	Y	
29	R211H	N	N	N	N	N
30	A215T	Y	N	N	N	N
31	G216R	Y	N	N	N	N
32	E217K	Y	N	N	N	N
33	T233I	Y	N	N	N	N
34	R235Q	Y	N	N	N	N
35	G239S	Y	N	N	N	N
36	R253P	Y	N	N	N	N
37	P274L	Y	N	N	N	N
38	S291L	Y	Y	N	N	N
39	P297Q	N	N	N	N	N
40	L304R	N	N	N	N	N
41	M310T	N	N	N	N	N
42	A329V	Y	N	N	N	N

Table : 3 Output of disorder predictors that shows the affect of missense mutations.
(‘O’ = orders ; ‘D’ = disorders)

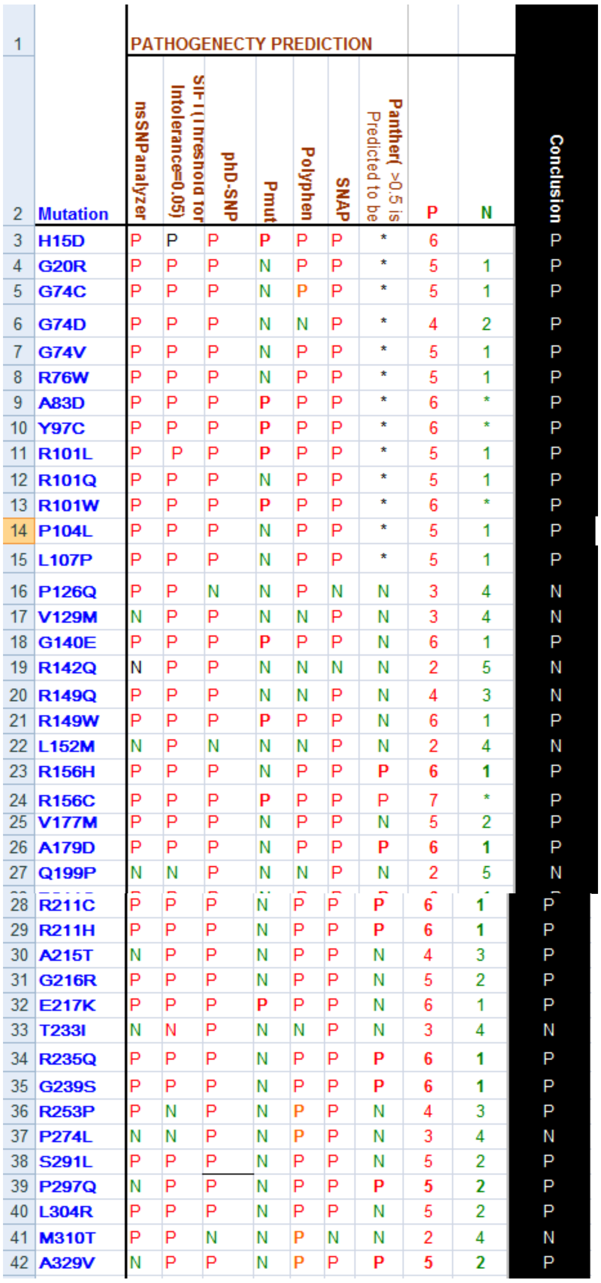
C.Disorder Prediction



1		DISORDER PREDICTION												
2	Mutation	DisProt,(O=Ordered, D= Disodered)	Drip-Pred	FoldIndex	FoldUnfold(composition of the Protein,avg=20.87)	GlobPlot	IuPred	MetaPrDos(Prediction False Positive Rate:5.0%)	PrDos	RONN	Spritz	O	D	Conclusion
3	H15D	O	o	O	O	O	O	O	O	O	O	10	*	O
4	G20R	O	o	O	D	O	O	O	O	O	O	9	1	O
5	G74C	O	o	O	O	O	O	O	O	O	O	10	*	O
6	G74D	O	o	O	O	O	O	O	O	O	O	10	*	O
7	G74V	O	o	O	O	O	O	O	O	O	O	10	*	O
8	R76W	O	o	O	O	O	O	O	O	O	O	10	*	O
9	A83D	O	o	O	O	O	O	O	O	O	O	10	*	O
10	Y97C	O	o	O	O	O	O	O	O	O	O	10	*	O
11	R101L	O	o	O	O	O	O	O	O	O	O	10	*	O
12	R101Q	O	o	O	O	O	O	O	O	O	O	10	*	O
13	R101W	O	o	O	O	O	O	O	O	O	O	10	*	O
14	P104L	O	o	O	O	O	O	O	O	O	O	10	*	O
15	L107P	O	o	O	O	O	O	O	O	O	O	10	*	O
16	P126Q	O	o	O	D	O	O	O	O	O	D	8	*	O
17	V129M	O	o	O	O	O	O	O	O	O	O	10	*	O
18	G140E	O	o	O	D	O	O	O	O	O	O	9	1	O
19	R142Q	O	o	O	D	O	O	O	O	O	O	9	1	O
20	R149Q	O	o	O	O	O	O	O	O	O	O	10	*	O
21	R149W	O	o	O	O	O	O	O	O	O	O	10	*	O
22	L152M	O	o	O	O	O	O	O	O	O	O	10	*	O
23	R156H	O	o	O	O	O	O	O	O	O	O	10	*	O
24	R156C	O	o	O	O	O	O	O	O	O	O	10	*	O
25	V177M	O	o	O	O	O	O	O	O	O	O	10	*	O
26	A179D	O	o	O	O	O	O	O	O	O	O	10	*	O
27	Q199P	O	o	O	O	O	O	O	O	O	O	10	*	O
28	R211C	O	o	O	D	O	O	O	O	O	O	9	1	O
29	R211H	O	o	O	D	O	O	O	O	O	O	9	1	O
30	A215T	O	o	O	D	O	O	O	O	O	O	9	1	O
31	G216R	O	o	O	D	O	O	O	O	O	O	9	1	O
32	E217K	O	o	O	D	O	O	O	O	O	O	9	1	O
33	T233I	O	o	O	O	O	O	O	O	O	O	10	*	O
34	R235Q	O	o	O	O	O	O	O	O	O	O	10	*	O
35	G239S	O	o	O	O	O	O	O	O	O	O	10	*	O
36	R253P	O	o	D	O	O	O	O	O	O	O	9	1	O
37	P274L	O	o	D	D	O	O	O	O	O	D	7	3	O
38	S291L	O	o	D	D	O	O	O	O	O	O	8	2	O
39	P297Q	O	o	D	O	O	O	O	O	O	O	9	1	O
40	L304R	D	o	D	O	O	O	O	O	O	O	8	2	O
41	M310T	O	o	D	D	O	O	O	O	O	O	7	3	O
42	A329V	D	o	D	O	O	O	O	O	O	O	8	2	O

Table : 4 Output of pathogenecity predictors that shows the affect of missense mutations. ('P' = pathogenic; 'N' = nonpathogenic)

D.Pathogenicity Prediction



1		PATHOGENECTY PREDICTION									
2	Mutation	nsSNPanalyzer	SIFT(Threshold for Intolerance=0.05)	phD-SNP	Pmut	Polyphen	SNAP	Panther(>0.5 is Predicted to be	P	N	Conclusion
3	H15D	P	P	P	P	P	P	*	6		P
4	G20R	P	P	P	N	P	P	*	5	1	P
5	G74C	P	P	P	N	P	P	*	5	1	P
6	G74D	P	P	P	N	N	P	*	4	2	P
7	G74V	P	P	P	N	P	P	*	5	1	P
8	R76W	P	P	P	N	P	P	*	5	1	P
9	A83D	P	P	P	P	P	P	*	6	*	P
10	Y97C	P	P	P	P	P	P	*	6	*	P
11	R101L	P	P	P	P	P	P	*	5	1	P
12	R101Q	P	P	P	N	P	P	*	5	1	P
13	R101W	P	P	P	P	P	P	*	6	*	P
14	P104L	P	P	P	N	P	P	*	5	1	P
15	L107P	P	P	P	N	P	P	*	5	1	P
16	P126Q	P	P	N	N	P	N	N	3	4	N
17	V129M	N	P	P	N	N	P	N	3	4	N
18	G140E	P	P	P	P	P	P	N	6	1	P
19	R142Q	N	P	P	N	N	N	N	2	5	N
20	R149Q	P	P	P	N	N	P	N	4	3	N
21	R149W	P	P	P	P	P	P	N	6	1	P
22	L152M	N	P	N	N	N	P	N	2	4	N
23	R156H	P	P	P	N	P	P	P	6	1	P
24	R156C	P	P	P	P	P	P	P	7	*	P
25	V177M	P	P	P	N	P	P	N	5	2	P
26	A179D	P	P	P	N	P	P	P	6	1	P
27	Q199P	N	N	P	N	N	P	N	2	5	N
28	R211C	P	P	P	N	P	P	P	6	1	P
29	R211H	P	P	P	N	P	P	P	6	1	P
30	A215T	N	P	P	N	P	P	N	4	3	P
31	G216R	P	P	P	N	P	P	N	5	2	P
32	E217K	P	P	P	P	P	P	N	6	1	P
33	T233I	N	N	P	N	N	P	N	3	4	N
34	R235Q	P	P	P	N	P	P	P	6	1	P
35	G239S	P	P	P	N	P	P	P	6	1	P
36	R253P	P	N	P	N	P	P	N	4	3	P
37	P274L	N	N	P	N	P	P	N	3	4	N
38	S291L	P	P	P	N	P	P	N	5	2	P
39	P297Q	N	P	P	N	P	P	P	5	2	P
40	L304R	P	P	P	N	P	P	N	5	2	P
41	M310T	P	P	N	N	P	N	N	2	4	N
42	A329V	N	P	P	N	P	P	P	5	2	P