Dottorato di Ricerca in Informatica
Università di Bologna e Padova

Ciclo XXIII

Settore scientifico-disciplinare di afferenza: INF/01
Indirizzo: Bioinformatica

Investigating the role of single point mutations in the human proteome: a computational study

Shalinee Tiwari

March 2011

Coordinatore	Relatore	Tutore
Prof. Simone Martini	Prof.ssa. Rita Casadio	Prof. Lucinao Margara
------------------------	-------------------------	--------------------------

Dedicated to ‘my biggest critic, yet my biggest supporter’
my loving husband Dr. Priyank Shukla

Investigating the role of single point mutations in the human Proteome: a computational study

Shalinee Tiwari

Synopsis

In the post genomic era with the massive production of biological data the understanding of factors affecting protein stability is one of the most important and challenging tasks for highlighting the role of mutations in relation to human maladies. The problem is at the basis of what is referred to as molecular medicine with the underlying idea that pathologies can be detailed at a molecular level. To this purpose scientific efforts focus on characterising mutations that hamper protein functions and by these affect biological processes at the basis of cell physiology. New techniques have been developed with the aim of detailing single nucleotide polymorphisms (SNPs) at large in all the human chromosomes and by this information in specific databases are exponentially increasing. Eventually mutations that can be found at the DNA level, when occurring in transcribed regions may then lead to mutated proteins and this can be a serious medical problem, largely affecting the phenotype. Bioinformatics tools are urgently needed to cope with the flood of genomic data stored in database and in order to analyse the role of SNPs at the protein level. In principle several experimental and theoretical observations are suggesting that protein stability in the solvent-protein space is responsible of the correct protein functioning. Then mutations that are found disease related during DNA analysis are often assumed to perturb protein stability as well. However so far no extensive analysis at the proteome level has investigated whether this is the case. Also computationally methods have been developed to infer whether a mutation is disease related and independently whether it affects protein stability. Therefore whether the perturbation of protein stability is related to what it is routinely referred to as a disease is still a big question mark. In this work we have tried for the first time to explore the relation among mutations at the protein level and their relevance to diseases with a large-scale computational study of the data from different databases. To this aim in the first part of the thesis for each mutation type we have derived two probabilistic indices (for 141 out of 150 possible *SNPs*): the perturbing index (*Pp*), which indicates the probability that a given mutation effects protein stability considering all the "in vitro" thermodynamic data available and the disease index (*Pd*), which indicates the probability of a mutation to be disease related, given all the mutations that have been clinically associated so far. We find with a robust statistics that the two indexes correlate with the exception of all the mutations that are somatic cancer related. By this each mutation of the 150 can be coded by two values that allow a direct comparison with data base information. Furthermore we also implement computational methods that starting from the protein structure is suited to predict the effect of a mutation on protein stability and find that overpasses a set of other predictors performing the same task. The predictor is based on support vector machines and takes as input protein tertiary structures. We show that the predicted data well correlate with the data from the databases. All our efforts therefore add to the SNP annotation process and more importantly found the relationship among protein stability perturbation and the human variome leading to the diseasome.

Key words: perturbing index, disease index, disease class index, disease related mutations, correlation coefficients, protein function, SNP annotation.

Acknowledgements

First and foremost my sincerest gratitude to my PhD thesis supervisor *Prof. Rita Casadio* for accepting me as a PhD student under her benevolent guidance and allowing me to be a part of her renowned 'Bologna Biocomputing Group'. I am thankful to her for her scientific supervision and for supporting me while writing my thesis. I am grateful to my internal committee members *Prof. Luciano Margara* and *Prof. Gianluigi Zavattaro.*

I thank *Prof. Simone Martini* (PhD coordinator), for his wonderful coordination, and, prompt and concerning replies to my queries. I am thankful to all the members of 'Bologna Biocomputing Group', in particular *Dr. Pier Luigi Martelli* and *Dr. Marco Vassura* for guiding me and helping me in thesis writing, *Dr. Piero Fariselli* for his invaluable suggestions in the predictor development, *Dr. Ivan Rossi* and *Dr. Gianluca Tasco* for their hardware and technical support.

I would like to thank my external referees, *Prof. Osman Ugur Sezerman* (Sabanci University, Turkey) and *Prof. Gustavo Parisi* (Universidad Ncaional de Quilmes, Argentina) for accepting to review this thesis and giving me their valuable comments & suggestions.

Beyond everything I am indebted to my parents for their unflagging love, support and all the sacrifices they have made for me. I am obliged to my father *Mr. Mukul Tiwari* for his encouraging and inspiring attitude. He spared no efforts to provide me the best, and I don't have suitable words to thank my mother *Mrs. Hemlata Tiwari* for her love, care & support. Mom-Dad I love you and thanks for your everlasting and unconditional love. I am grateful to my sister *Ms. Sakshi Tiwari* for being my best friend and one of my best counsellors. I thank my brother *Mr. Avinav Tiwari* for giving me mental and moral support through out this undertaking. I owe my deepest gratitude to the most important man in my life, my husband *Dr. Priyank Shukla,* without whom this thesis would not have been possible. I am also thankful to my friends whom I met in Bologna. Thank you *Dhruv* and *Deepak* for your moral support and help.

Finally, I would like to acknowledge financial support received in the form of two scholarships; 'Brains-in' PhD residential scholarship from Institute of Advanced Studies (2009- 2010) and the University of Bologna's foreign student PhD scholarship (2008-2010).

At last but not the least, above all of us, regards to omnipresent GOD, for answering my prayers.

Shalinee Tiwari
Bologna, March 2011

Index

Original communication based on this thesis

Casadio R.*, Vassura M., Tiwari S. ,Fariselli P. and Martelli PL: *Correlating disease related mutations to their effect on protein stability: a large scale analysis in the human Proteome* (Submitted)

List of Abbreviations

3D	Three-Dimensional
BLAST	Basic Local Alignment Search Tool
COSMIC	Catalogue of Somatic Mutations in Cancer
DNA	Deoxyribonucleic acid
DSSP	Define Secondary Structure of Protein
GO	Gene Ontology
MCC	Matthews Correlation Coefficient
MCP model	McCullogh-Pitts model
OMIM	Online Mendelian Inheritance in Man
PDB	Protein Data Bank
Pd	Disease Index
Pp	Perturbing Index
PSSM	Position- Specific Scoring Matrix
RBF	Radial Basis Function
RMSE	Root Mean Square Error
RNA	Ribonucleic acid
SNP	Single Nucleotide Polymorphism
SVM	Support Vector Machines
UniProt	Universal Protein Resource

List of Figures

List of Tables

Chapter1. Introduction

In the post-genomic era a great shift of interest is occurring in computational studies aiming at deciphering the complexity of our genetic code. Methods derived from computer science help in developing tools that are extremely useful in analysing genomes and proteomes and modelling all the complexities of the area on the basis of all the molecular interactions that supports life in the cells. DNA in each cell encodes for proteins that are considered to be the gear of living tissues in which they carry out all the chemical reactions necessary for life. Proteins are endowed with three-dimensional structures that are relevant for their functions. When mutations occur at the DNA level, proteins may also be affected to the point that their stability is perturbed. A crucial problem is to understand to which extent this mechanism is responsible for maladies. In this work we are therefore trying to look inside the protein world in the context of small variations in their sequences. A brief introduction on protein structure is therefore at hand.

1.1 The protein universe

Proteins are polymers of 20 different amino acid molecules, linked covalently by peptide bonds, generating long backbones where the protein identity consists in the specific position along the sequence of the residue side chains. A protein is therefore the sequence of its residues with a variable length spanning from 40 up to thousands of residues in the over 10 millions of proteins from over 1000 species that are presently known in the data bases.

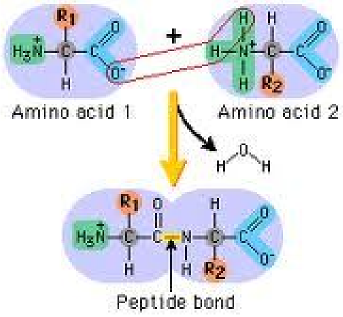
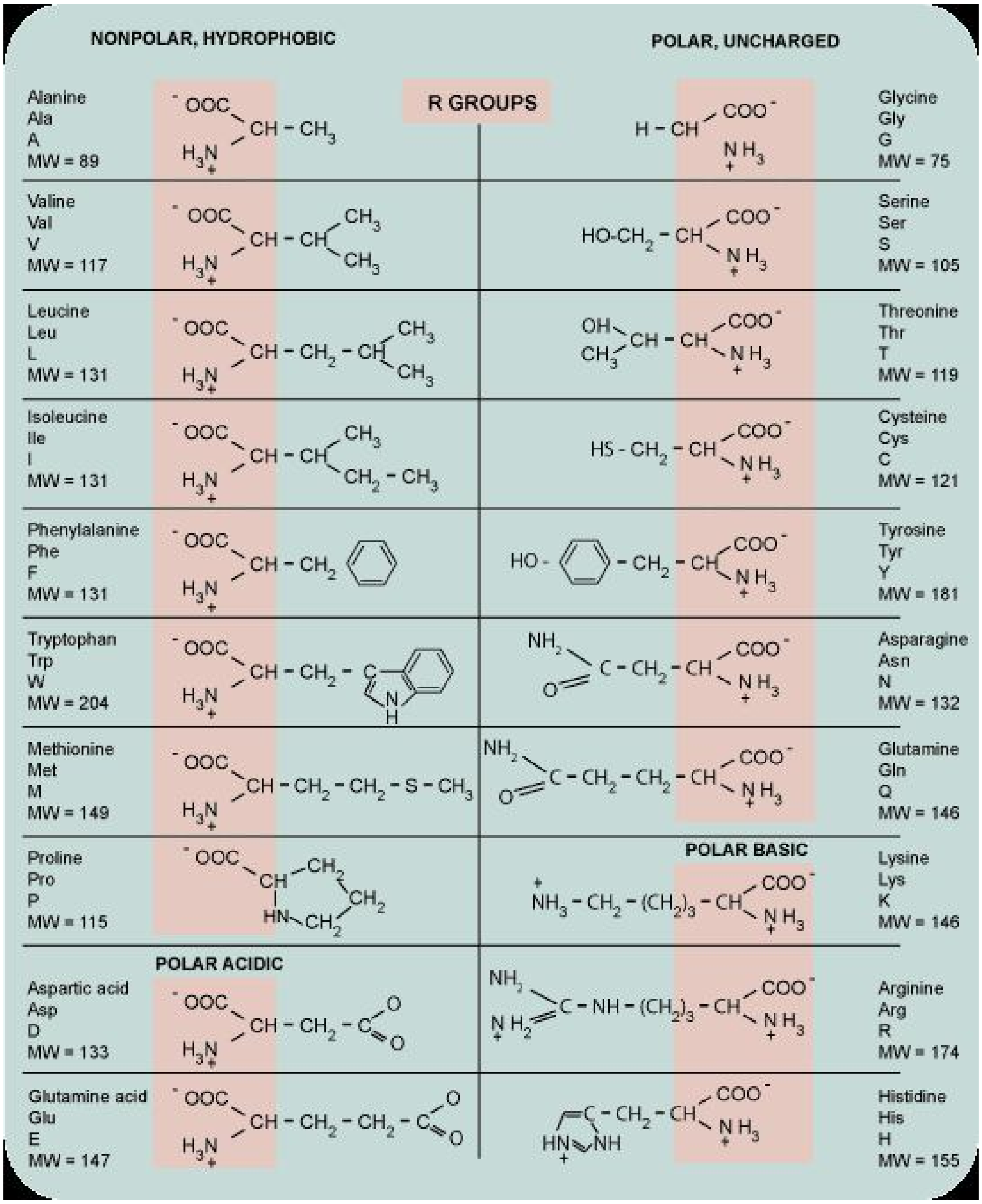


Figure1.1 A typical covalent bond formed between two amino acids. *R1 and R2 are the different side chains of each amino acid.*

One end of every polypeptide, called amino terminal or N-terminal, has a free amino group. The other end, with its free carboxyl group, is called the carboxyl terminal or C-terminal. A typical protein contains 200-300 amino acids but some proteins like peptides are smaller and yet other proteins like titans (about 30,000 residues) are much larger than the average protein size. Function is therefore depending on the specific protein sequence that in turn and in the cell space is organised in a three-dimensional structure. Routinely more proteins can take part into what is called a biological process and several thousands of biological processes are the basis of what is routinely referred to as a cell life.

1.2 Amino acids

Twenty amino acids are found in proteins (Figure 1.2). The amino acids differ from each other at the level of the portion of molecule called side chain [‘*R*’ in Figure 1.1]. Different residues in the protein chain are responsible of the specific structure, function and physicochemical properties of the different proteins.



.

***Figure 1.2. 20 common amino acid residues grouped according to the physico-chemical properties** like Polar acidic, Polar basic, Nonpolar hydrophobic and polar uncharged. Each Box shows the name of the amino acid (AA), the three-letter code of AA, the 1- letter code of AA and its Molecular weight of the amino acid along with the corresponding structure.*

1.3 The protein structure

There are four main levels of protein structure (Figure 1.3).

(1) **Primary Structure.** The primary structure of proteins refers to the linear chain and order of amino acid residues present in the proteins. The sequence of amino acid residues in a polypeptide is dictated by the codons in the messenger (mRNA) molecules.

(2) **Secondary Structure.** A local organisation of the protein backbone stabilised by hydrogen bonding that is basically classified into α helix and β strand, depending on the geometrical parameters.

(3) **Tertiary Structure.** The tertiary structure refers to the complete three-dimensional (3D) structure of the protein chain. The α helices and β sheets get folded into a compact globule. Folding of secondary structure into tertiary structure is driven by many interactions like hydrogen bonding, hydrophobic forces, electrostatic forces and Van der Waals forces.

(4) **Quaternary Structure.** Many proteins contain two or more different polypeptide chains, which are held together by the association of same non-covalent forces that helped in stabilization of tertiary structure. The quaternary structure of the protein is formed by the monomer-monomer interaction of the proteins. The proteins with multiple polypeptide chains are called oligomeric proteins. The quaternary proteins with identical subunits are termed as homo-oligomers, whereas proteins containing many different polypeptides are termed as hetero-oligomers.

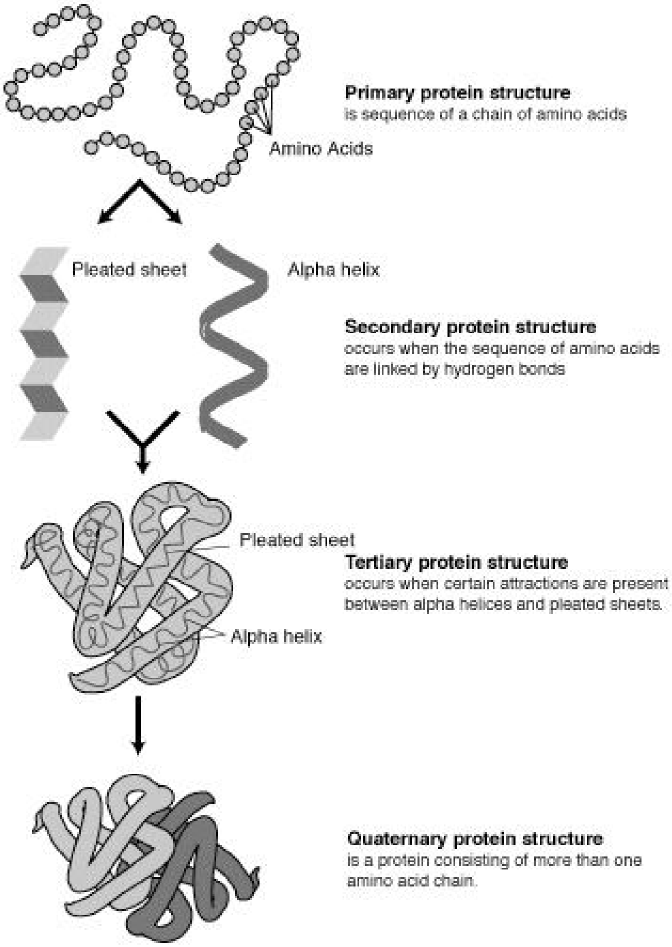

Primary protein structure
is sequence of a chain of amino acids
Amino Acids
Pleated sheet
Alpha helix
Secondary protein structure
occurs when the sequence of amino acids
are linked by hydrogen bonds
Pleated sheet
Tertiary protein structure
occurs when certain attractions are present
between alpha helices and pleated sheets.
Alpha helix
Quaternary protein structure
is a protein consisting of more than one
amino acid chain.

***Figure 1.3 The protein from its primary to its quaternary structure*.**

1.3.1 The relation of structure and function

The protein structure specifies its function. Large portion of the structure that generally referred to as protein domains are basically related to given protein functions, and interestingly due to evolutions even distantly related proteins may have in common similar domains when they perform the same functions in different species. Summing up, function is generally conserved when the structure is conserved, even when the sequence varies.

1.4 Mutations

Mutations are changes in the nucleotide sequence of the genetic material. Mutation is an error, which can occur any time while copying the genetic material during cell division, or by some external means like exposure to UV radiations or ionising radiations or by exposure to some chemicals or viruses. Any physical or chemical agent that causes mutation is termed as Mutagens. A mutation in any residue of a protein may affect the structure and therefore the functionality of the protein, which may further affect the metabolism of the organism and may lead to some disease.

1.4.1 Single Nucleotide Polymorphism (SNP)

A Single Nucleotide Polymorphism (SNP) is a single base mutation in DNA, the most simple form and most common source of genetic polymorphism in the human genome (90% of the human genome polymorphisms). Two types of nucleotide base substitutions give rise to SNPs:

- A transition substitution between purines (A, G) or between pyrimidines (C, T). This type of substitution constitutes two thirds of all SNPs.

- A transversion substitution that occurs between a purine and a pyrimidine.

1.4.2 Sequence Variation

Sequence variation caused by SNPs can be measured in terms of nucleotide diversity, the ratio of the number of base differences between two genomes over the number of bases compared. This is approximately 1/1000 base pairs between two equivalent chromosomes [Jin Li et al (1999)]. SNPs are not uniformly distributed over the entire human genome, neither over all chromosomes and neither within a single chromosome. There are one third of SNPs found within the coding regions to non-coding region SNPs. It has also been shown that sequence variation is much lower for the sex chromosomes. Within a single chromosome, SNPs can be concentrated for a specific region, usually implying a region of medical or research interest.

1.4.3 SNPs in coding regions

A SNP in a coding region may have two different effects on the resulting protein:

- ***Synonymous SNP:*** when the substitution causes no amino acid change to the protein it produces. This is also called a silent mutation.

- ***Non-Synonymous (nsSNP):*** when the substitution results in an alteration of the encoded amino acid. A missense mutation changes the protein by causing a change in codon. A nonsense mutation results in a misplaced termination codon. One half of all coding sequence SNPs result into non-synonymous codon changes.

1.5 The thermodynamics of the folding and unfolding of proteins

The primary structure of a protein is a linear polypeptide chain. After synthesis on a ribosome, the polypeptide chain quickly transforms from an unstructured, random conformation (the unfolded state) to its unique native conformation (folded) state, in which the protein carries out its function. Despite of this relatively easy concept of protein folding, defining the folding pathway from a protein's primary structure to its tertiary structure has proven to be an immensely challenging task to solve [Fitzkee et al. (2005); Kang and Kini (2009)]. However, much is known about the forces that distinguish unfolded and folded forms. The thermodynamic stability of a protein is defined by the free-energy difference between the unfolded state and the folded state [Pace 1990] as follows:

$$\Delta\Delta G = \Delta G_{unfold} - \Delta G_{fold} \qquad \textit{1.1}$$

The energetic difference between the folded and unfolded states is due to differences in the balance of intramolecular interactions, interactions with the surrounding medium, and entropic factors. The folded state is more stable (with lower ΔG) than the unfolded state (with greater ΔG). Many studies have confirmed that a high proportion of globular proteins are marginally stable under physiological conditions since the free energy of folded conformations are generally only 5 to 10 kcal/mol higher than for unfolded [Privalov and Khechinashvili (1974); Pace (1975); Ruvinov et al. (1997); Vogl et al. (1997); Giver et al. (1998)]. Experimental and theoretical work has provided mechanistic details of the forces that govern the folding and unfolding of proteins. These forces include hydrogen bonding interactions [Mirsky and Pauling 1936], electrostatic interactions [Thornton (1981); Cho and Raleigh (2006)], Van der Waals interactions [Chen and Stites (2001)], and hydrophobic interactions [Dill (1990)].

The major destabilizing force among all is the conformational entropy [Pace et al. (1996)]. These various interactions that arises from the functional side chains of amino acids, contribute with various strengths to the protein stability and protein folding process [Dill (1990); Yang et al. (2007); Dill et al. (2008)]. The final structure of the protein monomer or complex is a result of an understated balance between stability posing and stability opposing forces and thus a single mutation may shift this balance and can significantly perturb the whole protein structure and thus its function [Matthews (1987); Alber (1989)]. Intramolecular interactions define the overall structure and stability of a protein, as well as regions that can undergo conformational rearrangements. Also functional properties, such as catalysis, allosteric regulation and ligand binding, arise from the same interactions that define stability. Although much is known about individual forces, understanding the precise interplay of intermolecular interactions is still a problem [Mounce et al. (2009)].

1.6 Biomedical relevance of SNPs

One of the fundamental problems in protein research is the understanding of protein stabilization by complex physical interactions. This problem has attracted the attention also in molecular medicine since when it has been noticed that some diseases such as amyloidisis are likely to be related to changes in protein stability due to mutations occurring in the protein sequence [Pepys et al (1993)]. The protein stability mechanism has been studied by using site-directed mutagenesis followed by thermodynamic measurements and structure determination [Matthews et al (1993)]. In the biomedical context SNPs are very important, as they are the most common source of variation in the human genome. Due to the redundant nature of the RNA triplet code that encodes proteins, many of these SNPs may not cause an amino acid change in the encoded protein (synonymous mutations). However, when a SNP causes an amino acid change (a non-synonymous mutation) there may be an effect on the structure and function of the encoded protein. Loosing of function may lead to disease. It would therefore be extremely useful to be able to predict which mutations are likely to cause disease. Identifying those SNPs that promote susceptibility or protection to complex diseases will aid early diagnosis, prevention and help in finding treatments.

A SNP may affect the function of a protein mainly in three ways [Worth et al (2007)].

- Firstly, a SNP may affect the functional residues of a protein *i. e.* the active site or protein-protein interaction site, impairing the protein's ability to carry out its normal function and hence affecting the molecular pathway within which the protein functions.

- Secondly, destabilizing the protein fold (increasing the ratio of unfolded protein to folded protein) or stabilizing it (decreasing the ratio of unfolded protein to folded protein).

- Thirdly by causing protein aggregation.

Chapter 2. The state of the art: reviewing literature

2.1 In-sights

The question of how changes protein stability after a single point mutation has been a matter of investigation for scientists for a long time. Starting the sixties, several theoretical and computational methods have been developed to compute stability change promoted by residue mutations in the protein fold. The earliest approaches involved free energy calculations with detailed models coupled to semi-empirical potentials [Basch et al, 1987; Tidor and Karplus, 1991]. Broadly there are two distinct *in-silico* approaches to predict the stability of proteins after point mutations, energy function and machine learning based methods.

2.1.1 Energy function-based approaches

Methods that utilize energy functions can be further subdivided into two types: physical based and knowledge based. The first ones based on physical pair wise interaction potentials directly allow the calculations of ΔΔG values upon substitutions by adopting classical force fields [Alexander Benedix et al., 2009; Masso and Vaisman, 2008; Parthiban et al, 2006; Yin et al, 2007; Pokala et al., 2005; Rohl et al, 2004; Gao et al, 1989; Bash et al, 1987]. This approach, while very useful for small protein sets, is however computationally expensive and cannot be applied on large-scale investigation [Capriotti et al, 2006]. Other methods have been recently developed and they are briefly described in the following sections.

2.1.1.1 PoPMuSiC vol 1.0

The first version of PoPMuSiC is a knowledge-based predictor [Gilis and Rooman, 2000] designed to predict protein stability upon single point mutation. It basically uses the backbone dihedral potentials derived from the protein structures. Proline mutation is not allowed since it is suspected to modify the backbone structure. The threshold value for non-local interactions is above 15 residues. 141 protein structures are used to extract the torsion angle potentials and distances between the amino acid residue atoms. The method takes as input the protein structure and its schematic representation is shown in Figure 2.1. Potentials are derived from observed frequencies of sequence and structure patterns in a dataset of 141 high-resolution protein X-ray structures with < 25% sequence identity or no structural similarity. Two types of potentials, called torsion potentials, consider only local interactions along with the sequence and C^{μ}-C^{μ} potentials. These are distance potentials dominated by non-local, hydrophobic interactions, purely based on the propensities of pair amino acid residues (a_m, a_j) at positions m and j along the sequence at a fixed 3-D distance, $d_{mj,}$ between the side chain centroids C^{μ}. PoPMuSiC computed values well correlate with the measured folding free energy changes since the Pearson correlation coefficient is 0.80.

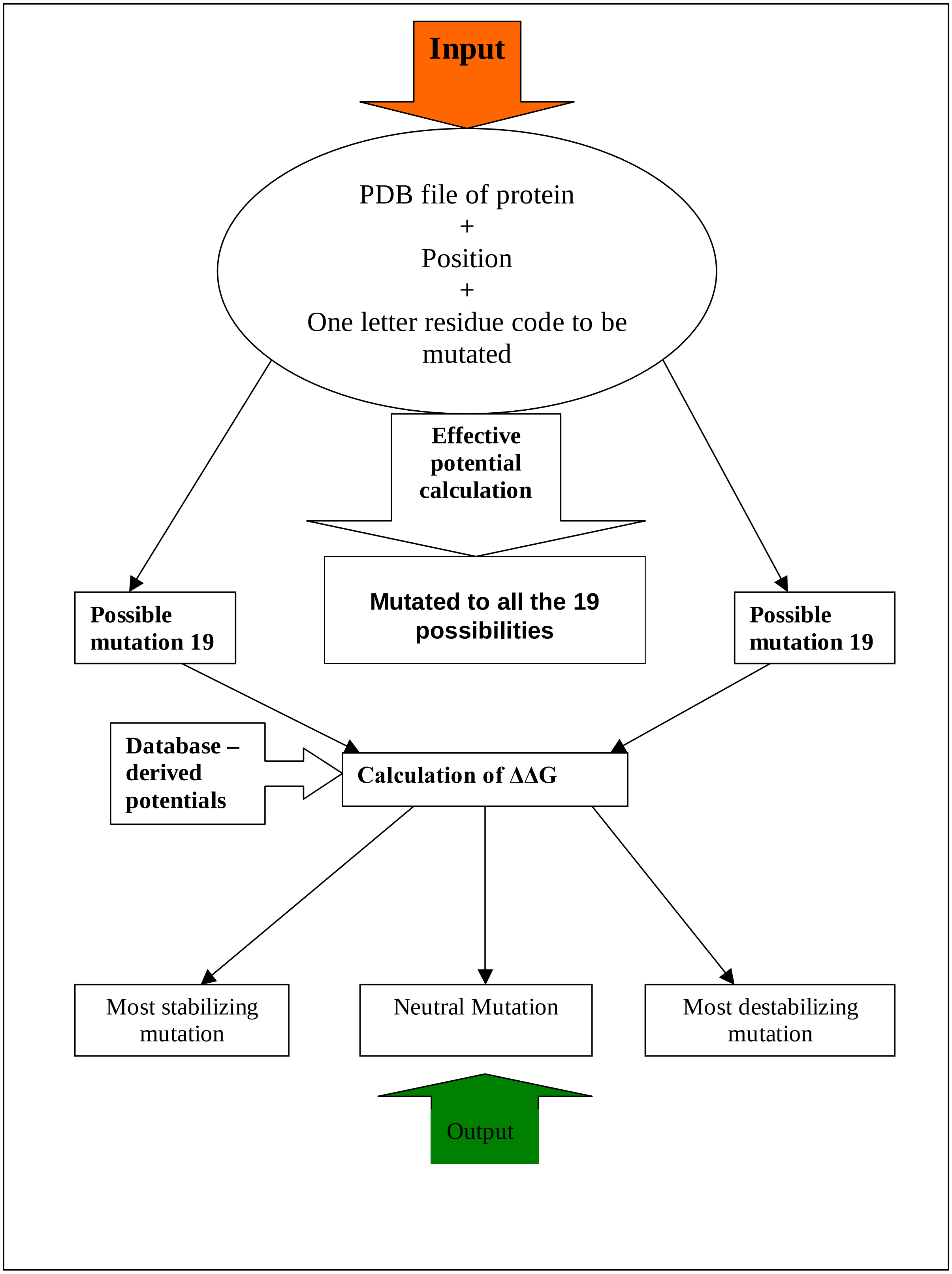

Input
PDB file of protein
+
Position
+
One letter residue code to be mutated
Effective potential calculation
Possible mutation 19
Mutated to all the 19 possibilities
Possible mutation 19
Database – derived potentials
Calculation of ΔΔG
Most stabilizing mutation
Neutral Mutation
Most destabilizing mutation
Output

Figure 2.1 The flow chart of PoPMuSiC 1.0 *[Gilis and Rooman, 2000]*

2.1.1.2 Fold-X

Fold-x [Guerois et al, 2002] uses an empirical force field and computes the energetic effect of point mutations as well as the interaction energy of protein complexes (including Protein-DNA). FoldX mutates protein and DNA side chains using a probability-based rotamer library, while exploring alternative conformations of the surrounding side chain. It can be used for a number of purposes, including prediction of the effect of point mutations or human SNPs on protein stability or protein complexes, protein design to improve stability or modify affinity or specificity and homology modelling. The energy function of FoldX includes terms, which are found to be important for protein stability, where the energy of unfolding (ΔG) of a target protein is calculated as

$$\Delta G = \Delta G_{vdw} + \Delta G_{solvH} + \Delta G_{solvP} + \Delta G_{hbond} + \Delta G_{wb} + \Delta G_{el} + \Delta S_{mc} + \Delta S_{sc} \quad 2.1$$

ΔG_{vdw} = Sum of Van der Waals contributions of all atoms
ΔG_{solvH} = Difference in solvation energy for apolar groups of residues in folded and unfolded states
ΔG_{solvP} = Difference in solvation energy for polar groups of residues in folded and unfolded states
ΔGhbond= Extra stabilizing free energy by water bridges
ΔG_{el} = Electrostatic contribution of charged groups
ΔS_{mc} = Entropy cost for fixing backbone in the folded state
ΔS_{sc} = Entropy cost for fixing a side chain in a particular conformation.

The program is freely available (http://fold-x.embl-heidelberg.de) and it was trained on a dataset of 339 mutations in 9 proteins. Testing was done on a dataset of 667 mutations in 82 proteins. The method reaches a Pearson correlation of 0.83 with a standard deviation of 0.81 kcal/mol.

2.1.2 Machine Learning methods

In Computational Biology, machine-learning methods are often adopted in order to perform a classification of data described with complex features. They are supervised methods since they are able to extract the relevant information from a set of data (the training set) for which the correct classification in known. Machine learning methods include Neural Networks (NNs), Support Vector Machines (SVMs), Hidden Markov Models (HMMs). Each one of these tools is endowed with a set of trainable parameters, whose values are determined by means of training algorithms that are suited to minimise the error rate of the classification of the data included in the training set. Machine learning methods are able to generalise the information extracted from the training set to new data with unknown classification and can be therefore used as predictors. Statistical scoring indexes such overall accuracy and Matthew’s correlation coefficient have to be evaluated on testing sets that contains data completely unrelated to those included in the training set. The size and quality of training set is crucial in determining the performances of a prediction method. Owing to their generality, most of machine learning tools can extract information starting from different types of input. In the field of the prediction of the effect of protein mutations on protein stability, different tools have been developed. They analyse information derived from either the protein structure or sequence. Structure-based predictors are usually more efficient, while sequence-based methods have a larger applicability, since the do not require the protein to be endowed with a known 3-D structure. Here we will briefly review only methods that have been developed at the Biocomputing Group since they are among the state of the art predictors and later on we will focus in improving over their scores.

2.1.2.1 Multi Layer Neural Networks

Artificial neural networks are connectionist models, in which the computation is performed by many elementary units, called neurons. They can be activated or not, depending on the signals that they receive from other neurons (including the input units). The connection between two neurons (synapse) is unidirectional and it is modulated by a numerical synaptic weight. Figure 2.2 shows the McCullogh-Pitts (MCP) model of an artificial neuron. The neuron receives weighted signals from other neurons (of from the input) and integrates them. If the addition of weighted inputs exceeds a threshold value, typical of each neuron, the neuron is activated. The value of the activation is:

$$v_k = X_1W_{k1}+X_2W_{k2}+\ldots+X_iW_{ki} - T_k \qquad \mathit{2.3}$$

Where X = Signal from of the neurons (or from input)
W= Weight
T= Threshold

The trainable parameters of a NN are the synaptic weights between neurons and the threshold value of each neuron.

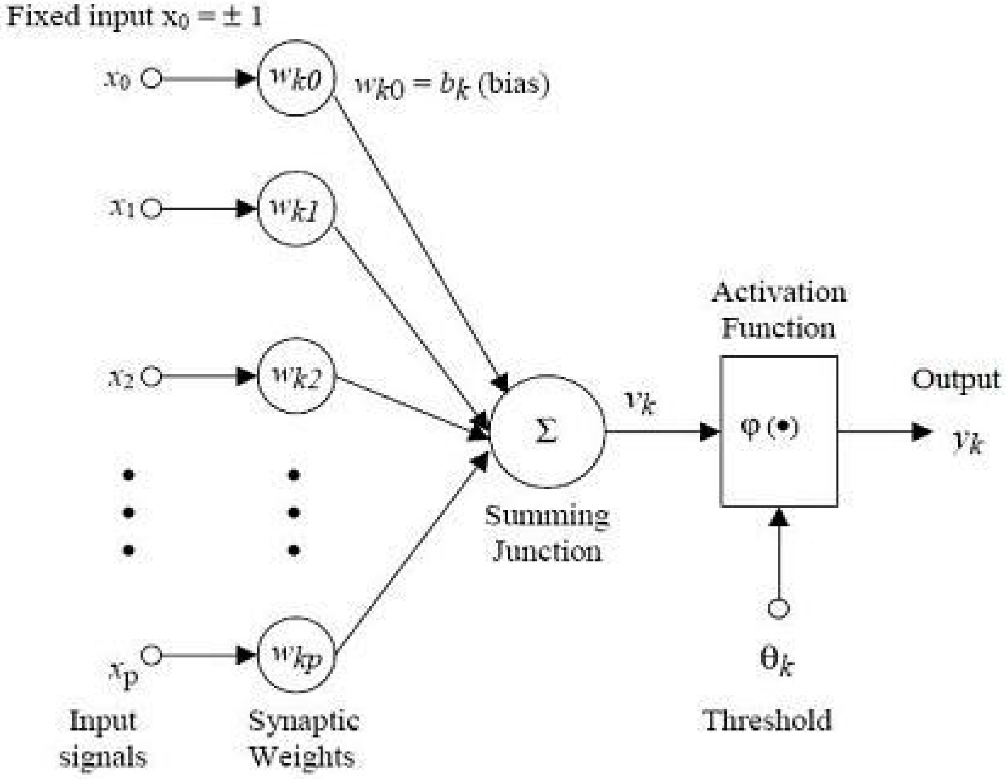


Figure 2.2 The MCP neuron model

Usually, the output of the neuron, denoted as y_k, is modulated by an activation function denoted with φ

(1) Step Function. The value of step function is 0 if the activation is lower than 0 that is if the sum of inputs is lower than the threshold value of the neuron (Tk) and 1 otherwise:

$$\varphi(v) = \begin{cases} 1 & \text{if } v \geq 0 \\ 0 & \text{if } v < 0 \end{cases}$$

2.5

(2) Piecewise Linear Function. The function is continue and linear in different sub domains

$$\varphi(v) = \begin{cases} 1 & v \geq \tfrac{1}{2} \\ v & -\tfrac{1}{2} > v > \tfrac{1}{2} \\ 0 & v \leq -\tfrac{1}{2} \end{cases}$$

2.6

(3) Sigmoid Function: This function may take the range between 0 and 1 as in the case of the logistic function, or between –1 and 1, as in the case of hyperbolic tangent.
The equation of the hyperbolic tangent is:

$$\varphi(v) = \tanh\left(\frac{v}{2}\right) = \frac{1 - \exp(-v)}{1 + \exp(-v)}$$

2.7

The NN topology is defined by the pattern of the synaptic connections between neurons. In a feed forward NN neurons are organized in layers connected in unidirectional way. The flux of information proceeds from the input layer towards the output layer. A steepest descent based algorithm that aims at minimizing the error rate on the training data is used for training the parameters of a feed forward NN. The algorithm is known as “backpropagation” (Rumelhart et al, 1986).

2.1.2.2 Support Vector Machines

Support Vector Machines (SVM) are a supervised method that perform a binary linear classification by searching for a separating hyperplane that maximises the margin, that is the distance between the hyperplane itself and the nearest data in each class. The Figure 2.3 shows a simple example of hyperplanes that correctly classify the different circles. The line in plane (a) performs a better classification than that on plane (b) because the separation margin is much larger. The margin is determined only by a subset of the data, called support vectors. SVMs can also solve non-linearly separable problems by introducing kernel techniques that represent, in an implicit way, the mapping of the original data into a feature space in which linear separation is easier. The most adopted kernels are the polynomial and the radial basis function (RBF) kernels. Predictive performance of SVM-based methods represents the state-of-the art in many real world applications [Cristianini and Shawe-Taylor, 2007].

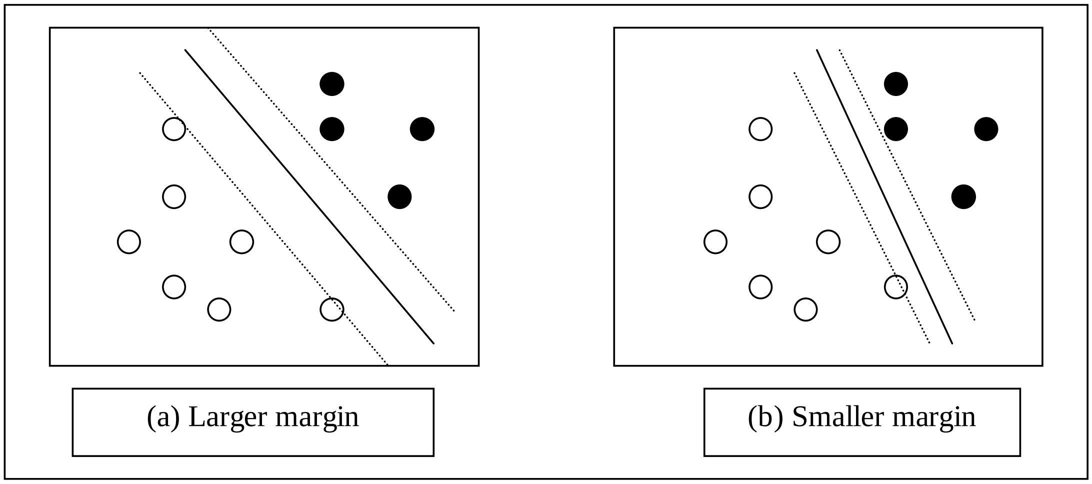


Figure 2.3 A scheme of the discriminating capability of Support Vector Machines (SVMs).

2.1.2.3 IMutant 2.0

Capriotti et al (2005) adopted SVMs to exploit information from protein structures. Previous implementations by the same group, based on neural network were able to compute only the sign of the free energy change (minus or plus) after a single point mutation. In the SVM implementation 42 and 43 input vectors code for sequence and structure, respectively. By this the first two vectors code for pH and temperature at which the stability energy change of the protein after the mutation was experimentally determined (data are available in the ProTherm data base, (http://gibk26.bse.kyutech.ac.jp/jouhou/Protherm/protherm.html). The next 20 values represent the mutated residue (–1 and 1 for the residue to be deleted and substituted residues respectively and 0 for the remaining other residues). The other 20 values encode the residues, which are in contact with the target (mutated) residue in the sphere of 0.9 nm centered on the coordinates of the residue that undergoes mutation in the case of structure-based prediction. In the case of sequence-based prediction these 20 values comprise the types of residues present in a window of length 19 stretching from N-terminus to C-terminus. For a structure based prediction RSA value (Relative Solvent Accessible Area) calculated with the help of DSSP program is also included as the 43rd vector. I Mutant 2.0 was trained on dataset retrieved from ProTherm (December, 2004). It was trained/tested with a cross-validation procedure. It reaches accuracy of 80% and a Person correlation value of 0.71 (with a standard error of 1.30 kcal/mol) when experimental ΔΔGs are correlated to the predicted ones starting from the protein structure. When prediction is sequenced based accuracy drops to 77% and Pearson correlation is 0.62 (with a standard error of 1.45 Kcal per mole). [Capriotti, et al, 2005]. The output includes a table of 19 rows* 6 columns (19 residues that are different from the one present in the sequence at a selected position), value and sign of ΔΔG prediction (Increase/decrease), original and mutated one letter code of residue, solvent accessible surface area, reliability index (RI) temperature and pH at which the prediction was carried out is obtained. By definition ΔΔG values for stabilizing mutation is greater than zero and for destabilizing mutation is less than zero [Capriotti et al, 2006)].

2.1.2.4 I-Mutant3.0

I Mutant 3.0 [Capriotti, et al, 2008] discriminate three classes: stabilizing, destabilizing and neutral single point mutations. It can be based on sequence and it reaches accuracy of 56% and average Matthew’s correlation coefficient of 0.27. It can also considers structures and then the overall accuracy is 61% with an average Matthew’s mean correlation coefficient of 0.35.

2.1.2.5 PHD-SNP

Considering a protein sequence the mutation can or cannot be disease-related. Considering a data based of annotated SNPs, routinely associated to monogenetic diseases. Capriotti et al [2006] developed a method based on decision tree with a SVM-based classifier (SVM-Sequence) coupled to a SVM-Profile trained on sequence profile information. Scoring was as high as 74% accuracy with Matthew's correlation coefficient equal to 0.46.

Table 2.1 gives a brief description about the information about the literature-based performance of different predictors at a glance.

Table 2.1 Literature based performance of different predictors

Predictors	***Accuracy***	***Correlation Coefficient***	***Method***	***Reference***
PoPMuSiC v1.0	-	**0.80** (Pearson correlation coefficient)	Knowledge based potential	Gilis and Rooman, (2000)
Fold-x	**61%**	**0.83** (Pearson correlation coefficient)	Knowledge based potential	Guerois *et al*, (2002)
I Mutant 2.0	**77%** (sequence based prediction) **80%** (structure based prediction)	**0.62** (sequence based prediction -Pearson correlation coefficient) **0.71** (structure based prediction- Pearson correlation coefficient)	Machine learning method (Support Vector Machine)	Capriotti, *et al*, (2005)
I Mutant 3.0	**56%** (sequence based prediction) **61%** (structure based prediction)	**0.27** (sequence based prediction- average MCC) **0.37** (structure based prediction- average MCC)	Machine learning method (Support Vector Machine)	Capriotti, *et al*, (2008)
PHD-SNP	**74%**	**0.46** (Matthew's correlation coefficient)	Machine learning method (Support Vector Machine)	Capriotti *et al*, (2006)

Chapter 3. Materials and Methods

3.1 Datasets

The data used in this work can broadly be classified as thermodynamic data and SNPs related data. The thermodynamic data refers to the online data with prior calculated thermodynamic properties of proteins like Gibbs free energy change (ΔΔG). The key aspect of the research in genetics is associating sequence variations with heritable phenotypes. As explained earlier the most common types of variations are SNPs, which occur approximately once every 100 to 300 bases. As these SNPs are expected to facilitate large-scale genetic studies, there has recently been a great interest in SNP discovery and detection and databases related to it. SNP related data includes the accurate information regarding the residue substitution and diseases related to it.

3.1.1 Thermodynamic data

The thermodynamic data for this study was derived from ProTherm, released on 31 March 2010. It contains 4044 mutations in 212 proteins collected from 10341 experiments [Kumar et al., 2006]. After excluding two membrane proteins we downloaded the data of 3089 experiments, which correspond to 141 SNPs in 129 proteins. The dataset comprises 55% data from eukaryotes, and rest 45% data from prokaryotes. The dataset contains entries from different organisms including 26% data from Homo sapiens.

The majority of ΔΔG values (63% values) lie in the range of 1 kcal/mol (Figure 3.1). The average ΔΔG value in the dataset was calculated to be –0.79 kcal/mol. We classified the mutation as non-perturbing if the associated | ΔΔG | value is less than or equal to 1 kcal/mol and perturbing otherwise. Therefore

If $|\Delta\Delta G| \leq 1$ kcal/mol → mutation non-perturbing protein stability

$|\Delta\Delta G| > 1$ kcal/mol → mutation perturbing protein stability

In case that multiple experiments for the same mutation in the same protein are present with different ΔΔG values the following rules were followed:

(1) The majority rule was applied to decide if the mutation is non-perturbing or perturbing

(2) If the number of perturbing experiments is the same of the non-perturbing ones the average ΔΔG value was considered

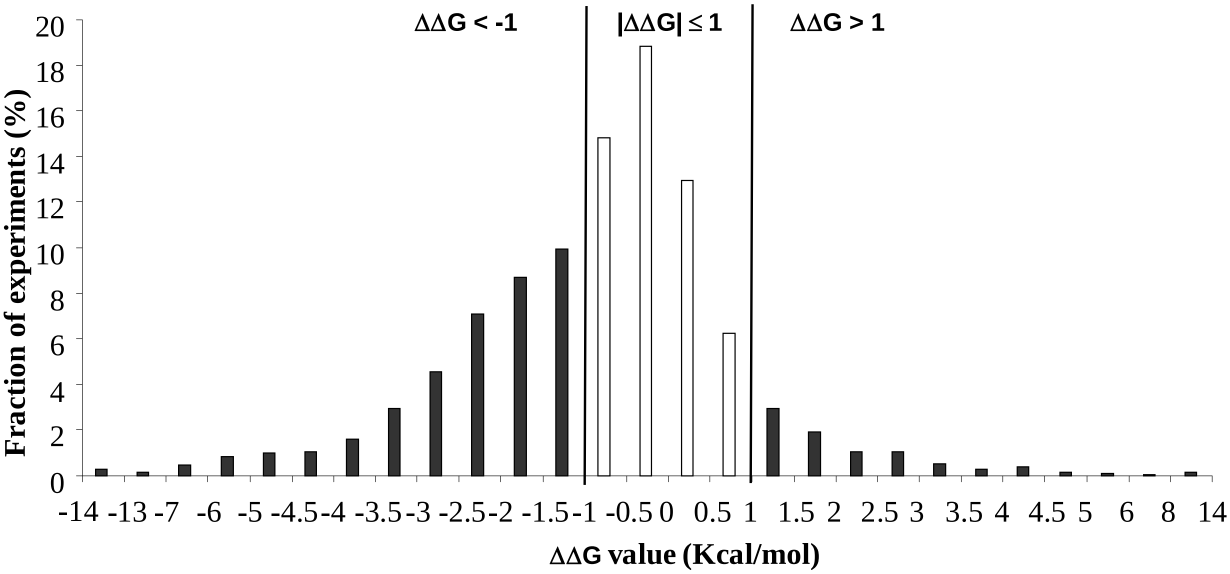


***Figure 3.1 Distribution of ΔΔG values corresponding to 141 SNP types found in ProTherm.** Each bar represents the fraction of experiments (total is 3089, Table 3.1) with ΔΔG values in the corresponding range. Mutations corresponding to ΔΔG < -1 Kcal/mol and ΔΔG>1 Kcal/mol are labeled as perturbing the protein stability (see text for details). ProTherm is at http://gibk26.bse.kyutech.ac.jp/jouhou/Protherm/protherm.html.*

3.1.2 SNP data

The SNP and disease related data were derived from two databases: UniProt KB and COSMIC (Table 3.1). The UniProt KB (Universal Protein Resource) release is 2010_04/23 March 2010 (http://www.uniprot.org/docs/humsavar) [Yip et al, 2008]. The entries without any disease annotations, annotated as polymorphism, were considered to be neutral mutations. We downloaded 35,557 mutation examples in 7,892 proteins. Out of them 13,456 were disease related mutations and 22,101 were neutral mutations.

Another set of SNPs, related to somatic cancer, was downloaded from COSMIC (www.sanger.ac.uk/genetics/CGP/cosmic/), release v46, 8 March 2010 [Forbes et al., 2010]. We extracted 5,063 mutation examples in 2111 proteins, 5, 061 mutation examples were disease related and only 2 mutations were neutral. There were 2,110 proteins containing only disease related mutations and a single protein with both disease related and neutral mutations.

Comparing COSMIC dataset with Uniprot KB we found 3,639 mutations, which were not present in the dataset that was extracted from UniProt KB. So after merging and sorting the two datasets derived from UniProt KB and COSMIC we ended up with 39,196 mutations in 8,277 proteins. These 39,196 mutations include 17,093 disease related mutations and 22,103 neutral mutations. Out of the complete set of proteins there are 831 proteins containing only disease related mutations, 5,455 proteins containing only neutral mutations and 1,991 proteins endowed with disease related and neutral mutations. For comparison with thermodynamic data we reduce this dataset to the mutations corresponding to the 141 types of SNPs available in the ProTherm database. Our reduced dataset contains 38,242 mutations, which include 16,476 disease related mutations and 21,766 neutral mutations. The reduced and final dataset consists of 8,224 proteins out of which 832 proteins have only disease related mutations, 5,439 proteins contain only neutral mutations and 1,953 proteins were endowed with both disease related and neutral mutations. Out of these only 37 mutations in 29 proteins were found to be residing in active site (ACT_SITE) according to UniProt KB annotations.

Table 3.1 SNP mutation data sets

SNP-Data Base	**SNP (no.)**			**Protein (no.)**			
	Disease related	Neutral	Total	Disease (only)	Neutral (only)	Disease & Neutral	Total
SNP-UniProt KB	13,456	22,101	35,557	446	6,261	1,185	7,892
SNP-COSMIC	5,061	2	5,063	2,110	0	1	2,111
SNP-UniProt+COSMIC (All)	17,093	22,103	39,196	831	5,455	1,991	8,277
SNP-Final (with 141 SNP types)	16,476	21,766	38,242	832	5,439	1,953	8,224

SNP mutations reported for globular proteins, as annotated in UniProtKB and COSMIC data sets, are sorted out by disease related, neutral and total mutations, with the corresponding number of proteins where the mutations were found. Our SNP-Final dataset is obtained by merging both UniProt and COSMIC mutations (SNP-UniProt + COSMIC) and by restricting to the 141 SNP types whose thermodynamic data are available in ProTherm. Human non-synonymous single nucleotide polymorphisms without annotation involvement in disease are considered non-damaging (Neutral).

Figure 3.2 shows a comparison between the residue composition of our dataset (in gray color) to that of mutated (yellow bars), disease related (red bars) and neutral (green bars) subsets. The expected frequency of mutation for each residue (violet bars) in human genome is nearly the same as the overall mutated residues (yellow bars) except in case of ‘Arginine (R)’. Vitkup et al explained the high mutation frequency of residue Arginine (R), in the year 2003 as consequences of high mutability of 5’ –CpG dinucleotides in the corresponding codons [Vitkup et al, 2003]. The frequency of mutation is different for the different residue types and different from the frequency of a given residue type in the database set (grey bars). This last value is however very similar to that of the human proteome (gray and pale blue bars, respectively) that comprises a much larger number of proteins (8,277 and 77,748 protein chains, respectively). Also for most of the residues, the frequency of occurrence in the disease related mutation set is similar to that in the neutral set. Exceptions include Glycine (G) and Cysteine (C), whose frequency of occurrence is twofold higher in the set of disease related mutations than in that of neutral cases and Valine (V) with the reversed occurrence. Summing up the 39,196 mutations included in our data set, and annotated in 8,277 proteins represent a significant sample of the mutational residue spectrum that can be randomly computed on the basis of the actual human Proteome.

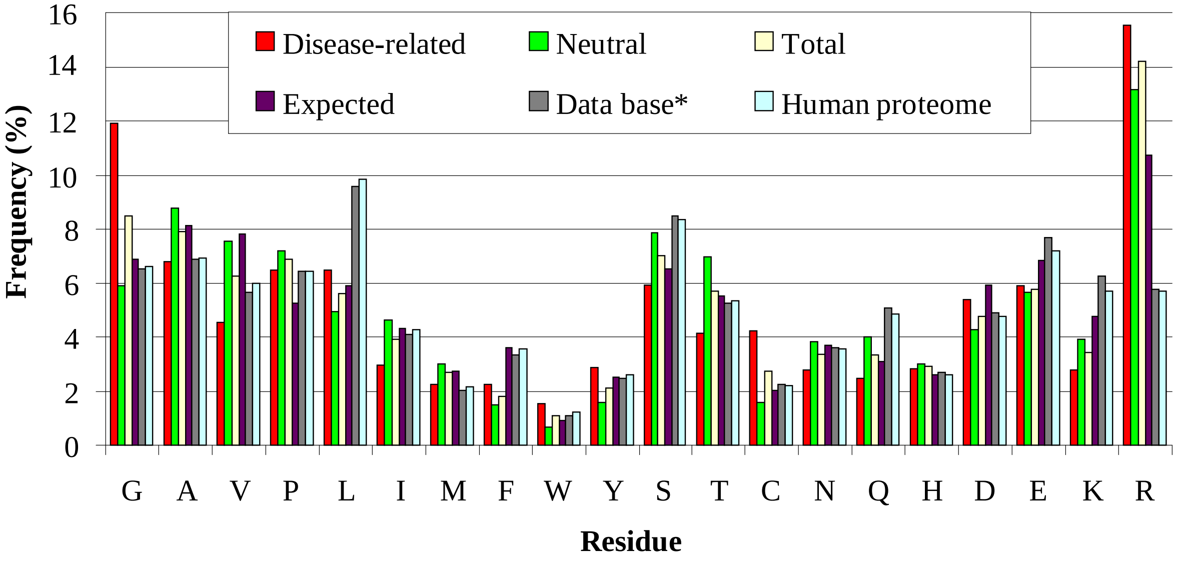


***Figure 3.2 Frequency of residues in the database.** Frequency of residues in disease related, neutral and total SNP mutation sub-sets of SNP-UniProt+COSMIC (Table 3.2) and in the corresponding total proteins (data base*). The frequency of residues in the human proteome (from EnSEMBL ver 5, based on the build of Genome Reference Consortium 37, with a total of 34,369,508 residues and 77,748 sequences) and that of the expected mutations rate on the human genome are also shown. Residues are grouped as follows: apolar (GAVPLIM), aromatic (FWY), polar (STCNQH), charged (DEKR), with residues in the same group listed at increasing order of size.*

Some of the proteins in our dataset were endowed with as many as above 1000 mutations (table 3.2). One protein named Cellular tumour antigen (p53), with length 393 residues, is associated with a total of 1285 mutations out of which 4 are neutral and 1282 are disease related mutations. About 69% of cases of our complete dataset restrain 1-3 mutations per protein (Table 3.2).

Disorders, as annotated in the Online Mendelian Inheritance in Man database (OMIM, www.ncbi.nlm.nih.gov/omim), were grouped into 22 disorder classes based on the affected physiological system, as previously described [Go et al., 2007]. We distinguish genetic cancer from somatic cancer ending up with 23 disorder classes, and when necessary neutral mutations are grouped in a specific class.

3.1.3 GO Terms

We downloaded GO terms from the current release (April 19, 2010) of the Gene Ontology vocabulary (GO) (http://www.geneontology.org/). It was containing 30282 terms, clustered in biological process, cellular component and molecular function. In our analysis we consider all the GO terms of each protein for each main root. 7270 proteins of our data set are endowed with GO terms: 6028 proteins with cellular components, 6455 proteins with molecular functions, and 5829 proteins with biological processes.

Table 3.2 Proteins with the highest number of mutations.

UniprotKB	**Protein name**	***L**	**°Tot**	**°D**	**°N**
P04637	*Cellular tumor antigen p53*	393	1285	1281	4
P00451	*Coagulation factor VIII*	2351	456	450	6
Q8WZ42	*Titin*	34350	354	131	223
P60484	*Phosphatidylinositol-3,4,5-trisphosphate 3-phosphatase and dual-specificity protein phosphatase PTEN*	403	319	318	1
P42771	*Cyclin-dependent kinase inhibitor 2A, isoforms 1/2/3*	156	265	249	16
P35555	*Fibrillin-1*	2871	219	202	17
P10275	*Androgen receptor*	919	212	209	3
P12883	*Myosin-7*	1935	193	184	9
Q13315	*Serine-protein kinase ATM*	3056	190	137	53
P00439	*Phenylalanine-4-hydroxylase*	452	186	185	1
P42336	*Phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha isoform*	1068	172	169	3
P06280	*Alpha-galactosidase A*	429	157	157	0
P29400	*Collagen alpha-5(IV) chain*	1685	152	145	7
P51587	*Breast cancer type 2 susceptibility protein*	3418	151	91	60
P00740	*Coagulation factor IX*	461	143	140	3
O00255	*Menin*	615	139	137	2
P25054	*Adenomatous polyposis coli protein*	2843	134	122	12
P40692	*DNA mismatch repair protein Mlh1*	756	128	110	18
P35222	*Catenin beta-1*	781	127	125	2
P22304	*Iduronate 2-sulfatase*	550	127	127	0
P00480	*Ornithine carbamoyltransferase, mitochondrial*	354	118	112	6
P02461	*Collagen alpha-1(III) chain*	1466	116	107	9
P15056	*Serine/threonine-protein kinase B-raf*	766	115	114	1
P34059	*N-acetylgalactosamine-6-sulfatase*	522	110	103	7
P21359	*Neurofibromin*	2839	109	103	6
P43246	*DNA mismatch repair protein Msh2*	934	106	93	13
P30613	*Pyruvate kinase isozymes R/L*	574	103	102	1
P15289	*Arylsulfatase A*	507	100	92	8

*28 proteins in our dataset endowed with at least 100 mutations (°Tot), sorted out as disease-related (°D), and neutral (°N). *L: protein length in residues.*

3.1.4 The 3D structure-data set

To train, test, cross validate and compare our predictor of protein stability with other predictors employed in the same task we retrieved the 3D structures of mutated proteins from UniProt KB released on July, 2009. The set of mutations on proteins with known 3D structure comprises 7347 mutation examples spread in 1187 protein chains (Table 3.3).

Table 3.3 The 3D-data set of proteins with neutral and disease-related mutations

	Polymorphic (only)	**Disease related & Polymorphic mutations**	**Disease related (only)**
No. of mutations (Total 7347)	2338 (32%)	-	5009(68%)
No. of PDB chains (Total 1187)	803 (67%)	209 (18%)	175 (15%)
No. of residues (Total 328702)	218578 (67%)	69548 (21%)	40576 (12%)

The dataset includes 160 types of mutations (mutation couples) and all possible 150 types of SNPs. Out of 7347 mutations 2338 (32%) mutations are related to polymorphism and 5009 (68%) belong to disease related mutations. Out of 1187 proteins 803 (67%) have exclusively neutral mutations and 175 (15%) only disease related mutations, while 209 (18%) are endowed with both disease related and neutral mutations.

3.2 Computing expected mutations

The base-line distribution of the residue mutations is computed by considering the frequency of occurrence of the four nucleotides namely A (Adenine), T (Thymine), C (Cytosine) and G (Guanine) and of the 64 codons in the human genome as reported in the Codon Usage Database (http://www.kazusa.or.jp/codon/). For each codon, all the 27 possible single-nucleotide mutations are generated. The probability of each mutation is estimated according to a matrix that reports the mutation rates of each nucleotide depending on the all 16 possible 5' and 3' neighborhoods [Hess et al., 1994]. The distribution of the residue mutations is computed as the sum of the probabilities of the corresponding codon mutations.

3.3 Defining different indices

We define two indices in this study namely: the perturbing index and the disease index for each SNP mutation of type X→ Y (where X and Y are wild type and mutant type residues respectively). For the computation of these indices we added a pseudo-count of 1 to all the classes under consideration. This smoothing scheme assumes an a-priori equi-probability of the classes correcting the probabilities of the mutation types under-represented in the data set [Durbin et al., 1998].

3.3.1 The Perturbing index (Pp)

We define the perturbation index as the probability of any mutation type X→ Y to result into protein destabilization. For each mutation type X→ Y we computed the associated perturbation probability or perturbation index (Pp) as in equation 3.1.

$$\text{Pp }(X \rightarrow Y) = \text{Np }(X \rightarrow Y)/\text{N }(X \rightarrow Y) \qquad \textit{3.1}$$

Where Np (X→ Y) = Number of X→ Y mutations with $|\Delta\Delta G|>1$ kcal/mol
N (X→Y) = total number of mutations X→Y

3.3.2 The Disease index (Pd)

The disease index may be defined as the probability of any mutation type X→ Y to be actually responsible for causing any disease. For each mutation type X→Y we computed the disease probability or Disease index (Pd) as in equation 3.2.

$$\text{Pd}(X \rightarrow Y) = \text{Nd }(X \rightarrow Y)/\text{N}(X \rightarrow Y) \qquad \textit{3.2}$$

Where Nd(X→Y) = total number of X→ Y mutations conductive to any kind of disease
N(X→ Y) = Total number of X→ Y mutations

3.4 Computing standard errors

For evaluating the Pp(X→Y) and Pd(X→Y) Standard Errors (SEPp(X→Y) and SEPd(X→Y)) a binomial distribution can be reasonably applied (equation 3.4).

$$\text{SEPx }(X \rightarrow Y) = ([\text{Px}(X \rightarrow Y) \cdot (1 - \text{Px }(X \rightarrow Y))/\ \text{N}(X \rightarrow Y)])^{½} \qquad \textit{3.4}$$

Where x is either p (perturbing index) or d (disease index).

However samples of data generated considering the diseases class partition have distributions, which depart from the traditional parametric distributions. In this case classical hypothesis-testing procedures based on strong parametric assumptions cannot be used to estimate the confidence intervals. Therefore we adopt also a bootstrap method that makes no assumption about the different data distributions. In the case of SEPp(X→Y) and SEPd(X→Y) we verify that values obtained with Equation 3.4 and the bootstrap procedure are similar. The standard errors reported in this work for all the different data sets are obtained with bootstrapping as follows. One thousand re-samples were generated starting from the different original data sets (Pp, Pd) by a random extraction with replacement. The size of each resample is set to be 90% of the complete dataset and the value of each index for each sample is computed. The standard deviation across the 1000 re-samples is then the standard error [Moore and McCabe, 2004].

3.5 Computing the regression line and the weighted residuals

For computing a regression line we adopted a Bivariate Least Median Square (BLMS) method previously described [Río et al., 2001]. The method is suited to compute the correct regression line when outliers are possible in the data set and it is based on a regression procedure that takes into account errors on both axes. The regression coefficients of the regression line are computed by minimizing the median of the weighted residuals (Ri) for each data pair (xi, yi) that are defined as in equation 3.5.

$$Ri= (yi-yi^*)^2/Wi \qquad \textit{3.5}$$

Where yi = experimental variable
yi*= prediction of the experimental variable yi
Wi = the weighting factor that corresponds to the variance of the i*th*-residual.

The method is based on an iterative process performed with the Monte Carlo simulation method to obtain the BMLS straight line as the best line of a group of robust straight lines generated by taking into account the errors in both axes. 1000 iterations were chosen for the Monte Carlo simulation stage. When necessary lines parallel to the regression one are drawn at a y distance equal to the root mean square of the residuals (root mean square error, RMSE (equation 3.6))

$$RMSE =(\sum(yi- yi^*)^2/(n - 2))^{½} \qquad \textit{3.6}$$

Where n is the number of data pairs.

3.6 Computing correlation and its statistical significance

We checked the normality of the data sets with the Jarque-Brera (1980, 1987) Lagrange Multiplier test (LM) and its significance from tables listing values for LM statistics computed with a Monte Carlo simulation [Wuertz and Katzgraber, 2009]. For detecting correlation we computed the Pearson product-moment correlation coefficient (r) (equation 3.7).

$$r(x,y) = cov (x,y) / (\sigma(x) \cdot \sigma(y)) \qquad \textit{3.7}$$

Where cov = covariance
σ = Sample standard deviation.

In order to estimate the statistical significance of r, we computed the associated *P-value* that estimates the probability that the value of the correlation coefficient is due to chance (the null hypothesis is that data are not correlated). Given r and the number of data (n), we compute the value (equation 3.8). A two-tailed test is performed for estimating the P-value. Routinely in our analysis P-values >0.05 is considered as the indication of non-significant correlation.

$$t = [r(x,y)\cdot(n-2)^{1/2}]/[1-r (x,y)^2]^{½} \qquad \textbf{\textit{3.8}}$$

Where t = distributed on a Student's t distribution with n-2 degrees of freedom.

Large error components reduce the apparent correlation coefficient and disattenuation (the estimation of the correlation in a manner that accounts for measurement errors contained within the estimates of the correlation parameters) is evaluated as follows:

$$r^*(x,y)=r\,(x,y)\ [(1-\ ASE(x)^2/\sigma(x)^{\,2})\ (1-\ ASE(y)^2/\sigma(y)^2)]^{\frac{1}{2}} \qquad \textbf{\textit{3.9}}$$

Where r(x,y), σ(x), σ(y) are as in Equation 3.7, ASE is Average Standard Error (for SE definition please see Equation 3.4) of the data [Spearman 1904/1987; Francis et al., 1999]. The r* significance was evaluated by computing a P*-value as described above.

In order to have correlation coefficients less sensitive to non-normal distributions and non linear dependence, correlation was also estimated with the Spearman's rank correlation coefficient (that assumes that the variables under consideration were measured on at least an ordinal (rank order) scale) and the Kendall tau rank correlation coefficient (that represents the difference between the probability that the observed data are in the same order for the two variables versus the probability that the observed data are in different orders for the two variables) [Hill and Lewicki, 2007)]. Significance tests were performed accordingly and as previously described [Myers and Well, 2003)].

3.6.1 Spearman's rank correlation coefficient

The Spearman’s rank correlation coefficient or Spearman’s rho is often thought to be Pearson correlation coefficient between ranked variables. It assumes that the variables under consideration were measured on at least an ordinal (rank order) scale. It calculates the soundness of the relationship between two variables using a monotonic function. If there are no repeated data values, a perfect Spearman correlation of +1 or −1 occurs when each of the variables is a perfect monotone function of the other. It is abbreviated as ‘ρ’.

3.6.2 Kendall tau rank correlation coefficient

It is a measure of rank correlation that is the similarity of the orderings of the data when ranked by each of the quantities. It is commonly known as Kendall's tau (τ) coefficient [Nelsen R.B et al (2001)]. It is used to measure the association between two measured quantities. It is used as a statistical test in statistical hypothesis test to establish if the two variables are statistically dependent on each other or not. Since it is a non-parametric test it does not depends on any assumptions on the distribution of variables X and Y (X and Y are independent variables). The Kendall tau τ coefficient is defined as in equation 3.10

$$\tau = \frac{(\text{Number of concordant pairs}) - (\text{Number of discordant pairs})}{\frac{1}{2}\, n\, (n-1)} \qquad \textbf{\textit{3.10}}$$

The denominator is the total number of pairs, so the coefficient must be in the range $-1 \leq \tau \leq 1$.

- If the agreement between the two rankings is perfect (i.e., the two rankings are the same) the coefficient has value 1.
- If the disagreement between the two rankings is perfect (i.e., one ranking is the reverse of the other) the coefficient has value −1.
- If *X* and *Y* are independent, then we would expect the coefficient to be approximately zero.

3.7 Statistical evaluation of GO terms by means of a P-value analysis

To assess whether a GO term is significant for a disease class, we performed a statistical test by computing P-values as previously described [Bartoli et al, 2009]. If N is the number of sequences in a given class which correspond to the same specific GO term, the P-value is the probability of finding N or more proteins that have a given annotation by chance, given the dimension of the class, the dimension of the data base and the overall number of sequences in the original data base with the given annotation. For each GO term within a cluster the corresponding P-value is evaluated as in equation 3.11.

$$P_{GO} = \sum_{i=n}^{\min(K,N)} \left(\frac{\binom{K}{i}\binom{D-K}{N-i}}{\binom{D}{N}} \right) \qquad 3.11$$

Where D = total number of protein sequences containing disease related mutations with at least one associated GO term

N = number of sequences in each disease class with at least one associated GO term

K =number of sequences with disease associated mutations in the entire database which have associated the same specific GO term. To account for the multiplicity of the GO terms in each class, we applied the Bonferroni correction to the computed P-values [Moore and McCabe, 2006].

3.8 Matthews correlation coefficient (MCC)

In the second part of the thesis to assess the performance of different predictors in predicting the stability changes caused by single point mutation in proteins different parameters are computed: accuracy, Matthews correlation coefficient (MCC), sensitivity and specificity of stabilizing and destabilizing mutations. . The Matthews correlation coefficient is defined as in equation 3.12.

$$\text{MCC} = TP*TN - FP*FN / ((TP+FP)(TP+FN)(TN+FP)(TN+FN))^{½} \qquad \textit{3.12}$$

Where TP = True Positives
TN = True Negatives
FP = False Positives
FN = False Negatives

For our predictor the true positive is the number of destabilizing mutations predicted as destabilizing, false positive is the measure of stabilizing mutations predicted as destabilizing mutations, true negative is the measure of stabilizing mutation predicted as stabilizing mutations and false negative is the measure of destabilizing mutation predicted as stabilizing mutations.

Sensitivity and Specificity are intended as statistical measures of the performance of a binary classification test. Sensitivity measures the proportion of actual positives, which are correctly identified, whereas Specificity measures the proportion of negatives, which are correctly identified. Theoretical optimal predictions can achieve 100% sensitivity and 100% specificity. The predictors predict the result in binary form as stabilizing or destabilizing mutations. In order to assess the performance of different methods Sensitivity [TN/(TN+FP)] and Specificity [TN/(TN+FN)] of destabilizing mutations and Sensitivity [TP/(TP+FN)] and Specificity [TP/(TP+FP)] for stabilizing mutations are calculated. The overall accuracy of the predictor is then calculated as:

$$Q= (TN+TP)/(TN+TP+FN+FP) \qquad 3.13$$

Chapter 4. Feature encoding for a NEW predictor

A new predictor was implemented to predict whether a single residue mutation perturbs the stability of a protein structure. Perturbing mutations are defined as the mutations that affect the Gibbs free energy of folding for more than 1Kcal/mol. The predictor is based on SVMs. We tested several kernels and we found that the most convenient for the problems at hand is the one based on Radial Basis Functions.

$$\text{RBF kernel} = \exp\left[-G\| x_i - x_j \|^2\right] \qquad 4.1$$

The results described are therefore relative to RBF kernels only. Our predictor fetches both local and global information from the respective mutated residue, protein sequence and protein structure. The global information refers to the properties, which are based on whole protein sequence of the respective mutated residue for example length of the protein chain, residue composition of the protein chain etc. Local information refers to the properties based on the local environment around the mutated residue like physiological properties of the mutated residue, types of residues in contact with the mutated residue in a 3-diamentional environment etc.

The predictor carries out the classification task by identifying two labels. If $|\Delta\Delta G| \leq 1$ kcal/mol then the mutation is predicted to be neutral and perturbing otherwise. Based on this threshold the predictor gives result as '–1' (if $|\Delta\Delta G| < 1$) for perturbing mutations and '1' ($|\Delta\Delta G| \geq 1$) for non-perturbing mutations. The input vector includes different features extracted from the protein structure. The features are:

4.1 Average contact information

By saying the contact tendency of any residue we mean to say the number and type of particular residue, found in contact with the target residue (the mutated wild type residue). The residues *j* and *i* are said to be in contact if the Euclidean distance between nearest atoms of the residue *i* and residue *j* is less than or equal to the pre assigned value of the threshold (9 Å in our case). The complete spherical 9 Å area around the target residue is known as *'Contact sphere'*. The Euclidean distances between the residues were calculated using the equation 4.2.

$$D= \sqrt{((x_1-x_2)^2+(y_1-y_2)^2+(z_1-z_2)^2)} \qquad 4.2$$

Where x_1, y_1, z_1= x, y, z co-ordinates of target residue respectively
x_2, y_2, z_2= x, y, z co-ordinates of other residue respectively

This information is encoded by vectors of value 20 (for 20 standard amino acid residues). We inserted value '100' in place of the residue for its presence and '0' for rest of the residues.

4.2 Profile information and PSSM generation

The Profile Specific Interactive (PSI) BLAST is a very strong and efficient measure of depicting the residue conservation in an environment. A matrix consisting of such vector representations for all the residues in a given sequence is known as **P**osition **S**pecific **S**coring **M**atrix (PSSM). The sequence of the protein with mutation were scanned against the reference datasets to compile a set of alignment profiles and Position Specific Scoring Matrices (PSSM) using PSI BLAST. Three cycles of PSI BLAST were run for each of the protein sequence and the scores were saved as profile matrices (PSSMs). The inputs consist of the PSSM of the target residue. 20 input vectors represent each residue.

4.2.1 Generation of profile files

We built sequence profiles of each protein having mutations in our 3 D structure dataset via a two-step method.

- We performed BLAST-p for all the chains of our 3D structure dataset against UniProt KB database, which included both reviewed, and non-reviewed entries from Swissprot and TrEMBL respectively. With this step we searched for the similarity of the sequences in our 3D structure dataset and complete database of UniProt KB.

- In the second step we separated the sequences with at least 25% identity and calculated the frequency of each residue at each position.

The conservation value of each residue at each position for all 20 amino acid residues along with gap (so total 21 residues) ranges between 0 and 100, and for each position this conservation value sums up to 100 (including gap).

4.3 Target residue information

For feeding SVM the information regarding the wild type residue and mutant residue we inserted as value ‘-1’ for the target residue (wild type residue to be deleted), ‘1’ for the inserted new residue (mutant residue) and for rest of the residues the value is kept to be ‘0’. Finally we get a vector of 20 values (for 20 standard amino acid residues) with values –1, 1 and 0.

4.4 Relative solvent accessibility

We calculated the accessibility of the target (mutated) residues using DSSP (Define Secondary Structure of Proteins) program (Figure 4.1). Then we normalized the accessibility values with the free residue surface (relative accessibility) of the residue at hand as follows:

Relative accessibility = Accessibility value from DSSP/ Maximal accessibility value 4.3

We set a threshold of 0.16 for distinguishing between core and non-core region. So if the normalized accessibility value of the target residue is less than 0.16 we considered it to be present in the core region else in non-core region. We computed the frequency of residues of both the classes falling in the core (normalized solvent accessibility <= 0.16) and non-core region (normalized solvent accessibility > 0.16) of the protein. For a structure-based prediction relative solvent accessibility can be used as information for feeding the SVM method, giving rise to the increment of one more input vector. The input vector in this case is encoded by a single value representing the relative solvent accessibility of the target residue.

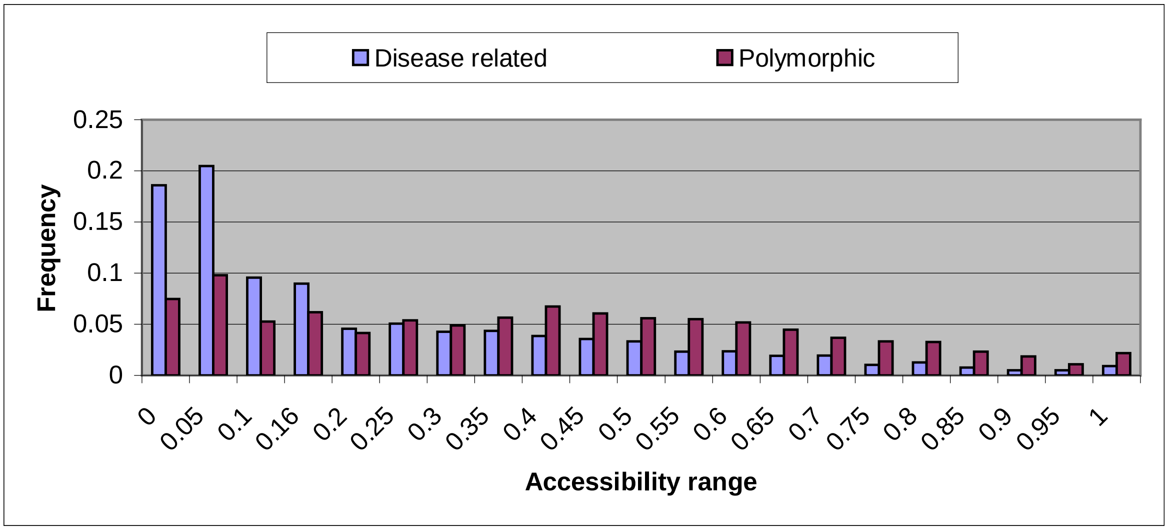


***Figure 4.1 Frequency of mutated residues to lie in the non-core part of the protein** (i.e. ASA > 0.16) and frequency of mutated residues to lie in the core of the protein (ASA <= 0.16).*

4.5 Protein sequence composition

Protein sequence composition can be considered as one of the most basic input feature. As global information, it gives the information about the frequency of different amino acid residues in a protein chain. The distinctive features of these 20 amino acids are size, electrical charge, polarity and shape. This information can be encoded by using a vector of value 20 (as we have 20 standard amino acid residues) each value indicating the contribution of the respective amino acid residue in protein sequence.

4.6 Secondary Structures

The interactions if amino acid residues among each other along with the polypeptide chain and with the solvent surrounding it determines the stability of 3- dimensional structure. Secondary structures like β sheets, α helix and loops etc are the most common secondary structure conformations found in polypeptides. These conformations are found to be responsible for the formation of energetically favourable structures. However in the protein world the DSSP algorithm is considered to be the standard and most popular method for assigning secondary structure to the amino acid residues of a protein, given the atomic resolution of co-ordinates of its atoms [Kabsch and Sander (1983)]. DSSP recognizes eight types of secondary structures depending upon the pattern of hydrogen bonds. DSSP annotation for secondary structure are H (Alpha Helix), B (residues in isolated beta – bridge), E (extended strand, participate in beta ladder), G (3-helix (3/10 helix)), I (5 helix (pi helix)), T (hydrogen bonded turn), S (bend), L (loop). These eight secondary classes are again regrouped into three major classes:

(1) Helix: It includes three classes 3-helix (3/10 helix) (G), Alpha Helix (H) and 5 helix (pi helix) (I)
(2) Strand: It includes two classes extended strand, participate in beta ladder (E), residues in isolated beta – bridge (B)
(3) Loop: Rest other three classes hydrogen bonded turn (T), bend (S) and loop (L) are regarded as the part of loop.

We calculated the frequency of association of different DSSP annotated secondary structures with polymorphic and disease related mutated residues. 3 input vectors encode the secondary structure information in SVM as input.

4.7 Implementing a NEW method

We have used Support Vector Machines in this method. The architecture of the method is similar to that which is used in I Mutant 2.0. In this new method we have exploited the evolutionary information in the form of PSSM matrices on top of all the other information, which improved the performance of the predictor. So in conclusion a sum total of 105 input features are fed to this SVM machine (Table 4.2).

Table 4.2 Input vectors used in the new predictor

Input Vectors Properties	***Number of vectors***
Target (Mutated) residue information	***20***
Residue composition of protein sequence	***20***
Frequency of residues in 9 Å contact sphere *(in the case of structure based prediction)*	***20***
Average Composition of residue in contact sphere	***20***
PSSM values (*Position Specific Scoring Matrix*)	***20***
Relative Solvent Accessibility *(only in the case of structure based prediction)*	***01***
Secondary structure information	***03***
Conservation of residue	***01***
Total vectors	***105***

The NEW predictor fetches different local and global information from protein structure. Figure 4.2 shows the schematic representation of the working of the NEW predictor. Each of the ovals represents either one of the local or global information along with the value of vector in round brackets. All these local and global information is processed in RBF kernel and it gives the value of |ΔΔ G|**.** Table 4.3 shows the performance of NEW predictor in different prospects.

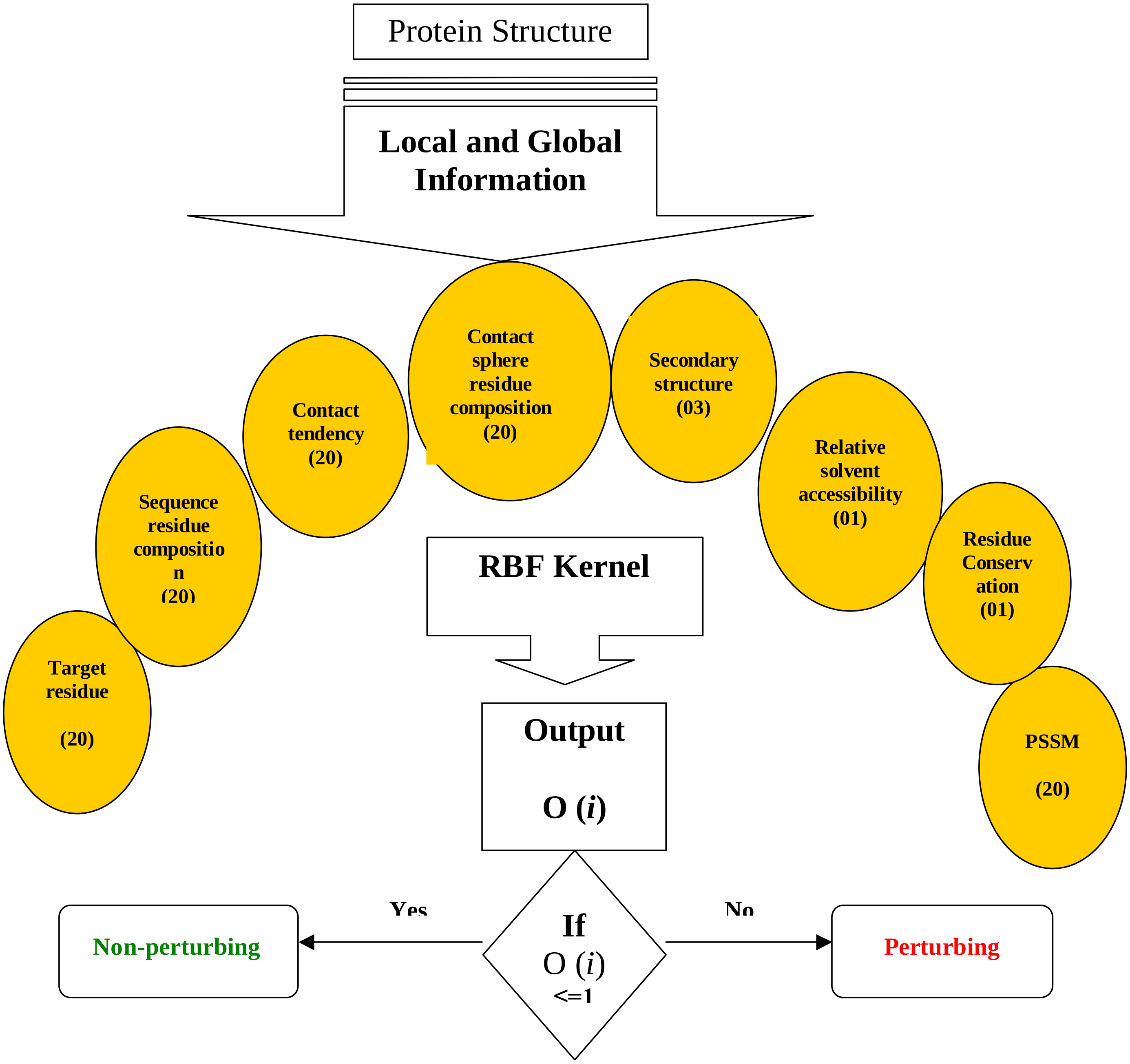

Protein Structure
Local and Global Information
Contact sphere residue composition (20)
Secondary structure (03)
Contact tendency (20)
Relative solvent accessibility (01)
Sequence residue compositio n (20)
Residue Conserv ation (01)
Target residue (20)
PSSM (20)
RBF Kernel
Output
O (i)
Yes
No
Non-perturbing
If O (i) <=1
Perturbing

***Figure 4.2 Schematic representation of 'NEW' predictor**. Local and global properties are represented in each of the ovals along with the value of vectors in round brackets.*

Table 4.3 Performance of NEW predictor

	MCC	Q	Specificity *Perturbing*	Sensitivity *Perturbing*	Specificity *Non-Perturbing*	Sensitivity *Non-Perturbing*
NEW	**0.32**	**0.68**	**0.69**	**0.81**	**0.64**	**0.49**

Chapter 5. Results

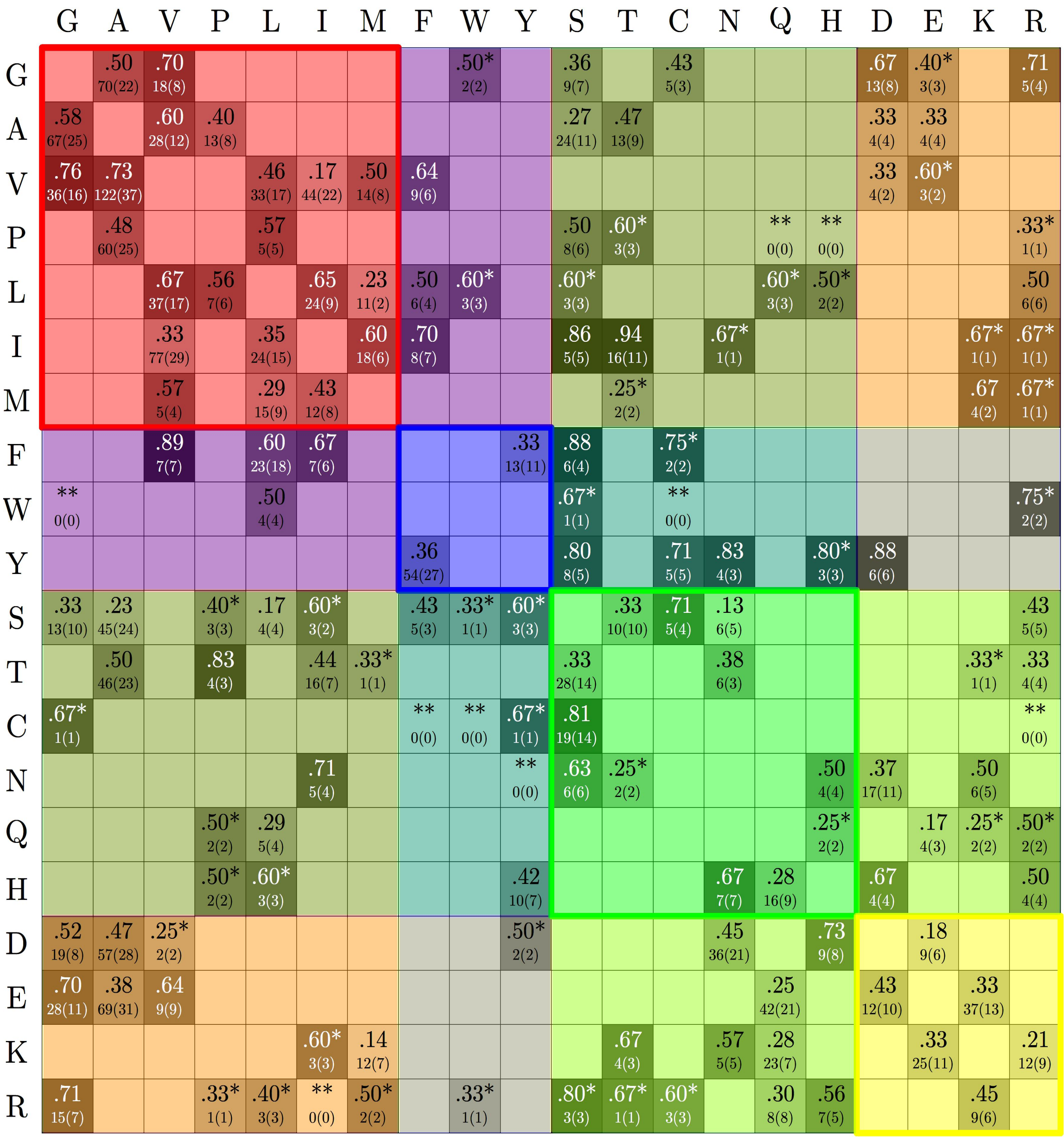
5.1 Calculating the Perturbing index (Pp)

We calculated the Perturbing index (Pp) for all the SNPs available in the latest release of ProTherm database (released on 31 March, 2010). The dataset contains the data related to 141 SNPs out of 150 possible SNPs (Figure 5.1). There are few mutation types which are not covered in this work are P→Q, P→H, W→G, W→C, C→F, C→W, C→R, N→Y, R→I as they were not the part of ProTherm data until 31 March 2010. For 98 and 84 types of mutations Pp values can be derived from at least 4 and 5 different measurements of the effect of the same mutation on the stability of different proteins, respectively. Only 10% of the mutation types are present as a single mutation event in the data base and 11% of the Pp values are endowed with a relative standard error > 60% (Figure 5.1).

In the Figure 5.1 the darker the colours, the higher is the perturbing index value (Pp). In other words the intensity of darkness of each box is relative to the probability of particular SNP to perturb the protein stability. Wild type residues are aligned vertically at the left and the target residues lie horizontally on the top. Each box shows the corresponding Pp value in bold along with the corresponding number of mutations reported in ProTherm and the numbers of proteins associated with these mutations are present among round brackets. Pp values with '*' marks shows that the calculation is done on less than 4 mutation examples. As explained earlier that this work includes 141 out of 150 types of SNPs because of their unavailability in the ProTherm database, the boxes with '**' sign depicts that the Pp value is not calculated for that particular mutation types because the data are not available in ProTherm. Pp values are computed according to equation 3.1.

Residues are grouped as, apolar residues (GAVPLIM, red), aromatic residues (FWY, blue), polar residues (STCNQH, green), charged residues (DEKR, yellow), with residues in the same group listed at increasing order of their size. The boxes comprising mutations within the same residue group are coloured in red, blue, green and yellow, respectively. All the other off diagonal squares comprising mutations included in different groups are shown as a combination of the four colours.

The perturbing index values range from smallest value of 0.13 for S→ N mutation to the maximum value of 0.94 for I→ T mutation. The average Pp value is 0.51 with an average standard error (SE) if 0.16. The distribution of Pp values in the dataset shows maximum types of SNP mutation carries Pp value of 0.5. The perturbing index value for each mutation type indicates the thermodynamic effect of that particular mutation on the stability of protein. The mutation type with a Pp value of 0.70 indicates that this particular mutation type has 70% probability to perturb the protein stability. There are some mutation types like V→ M, L→ H, G→ A, G→ W, P→ S, L→ R, L→ H, L→ F, D→ Y, H→ P, H→ R, N→ H, N→ K, P→ S, Q→ P, Q→ R, R→ M, T→ A and W→ L which have Pp value of 0.50. The perturbing index value of 0.50 is considered to be the neutral discriminative threshold among non-perturbing and perturbing values. The Pp values of 81 SNP mutation types out of total 141 SNP mutation types of our dataset are found to be greater than the Pp value of 0.50, symbolizing that 67% of the mutation types in our dataset are perturbing although to different extent. The remaining 43% of the mutations types are similarly classified as non-perturbing (with a low/modest perturbation probability, according to our definitions of perturbing index).



	G	A	V	P	L	I	M	F	W	Y	S	T	C	N	Q	H	D	E	K	R
G		.50 70(22)	.70 18(8)						.50* 2(2)		.36 9(7)		.43 5(3)				.67 13(8)	.40* 3(3)		.71 5(4)
A	.58 67(25)		.60 28(12)	.40 13(8)							.27 24(11)	.47 13(9)					.33 4(4)	.33 4(4)		
V	.76 36(16)	.73 122(37)			.46 33(17)	.17 44(22)	.50 14(8)	.64 9(6)									.33 4(2)	.60* 3(2)		
P		.48 60(25)			.57 5(5)						.50 8(6)	.60* 3(3)			** 0(0)	** 0(0)				.33* 1(1)
L			.67 37(17)	.56 7(6)		.65 24(9)	.23 11(2)	.50 6(4)	.60* 3(3)		.60* 3(3)				.60* 3(3)	.50* 2(2)				.50 6(6)
I			.33 77(29)		.35 24(15)		.60 18(6)	.70 8(7)			.86 5(5)	.94 16(11)		.67* 1(1)					.67* 1(1)	.67* 1(1)
M			.57 5(4)		.29 15(9)	.43 12(8)						.25* 2(2)							.67 4(2)	.67* 1(1)
F			.89 7(7)		.60 23(18)	.67 7(6)				.33 13(11)	.88 6(4)		.75* 2(2)							
W	** 0(0)				.50 4(4)						.67* 1(1)		** 0(0)							.75* 2(2)
Y								.36 54(27)			.80 8(5)		.71 5(5)	.83 4(3)		.80* 3(3)	.88 6(6)			
S	.33 13(10)	.23 45(24)		.40* 3(3)	.17 4(4)	.60* 3(2)		.43 5(3)	.33* 1(1)	.60* 3(3)		.33 10(10)	.71 5(4)	.13 6(5)						.43 5(5)
T		.50 46(23)		.83 4(3)		.44 16(7)	.33* 1(1)				.33 28(14)			.38 6(3)					.33* 1(1)	.33 4(4)
C	.67* 1(1)							** 0(0)	** 0(0)	.67* 1(1)	.81 19(14)									** 0(0)
N						.71 5(4)				** 0(0)	.63 6(6)	.25* 2(2)				.50 4(4)	.37 17(11)		.50 6(5)	
Q				.50* 2(2)	.29 5(4)											.25* 2(2)		.17 4(3)	.25* 2(2)	.50* 2(2)
H				.50* 2(2)	.60* 3(3)					.42 10(7)				.67 7(7)	.28 16(9)		.67 4(4)			.50 4(4)
D	.52 19(8)	.47 57(28)	.25* 2(2)							.50* 2(2)				.45 36(21)		.73 9(8)		.18 9(6)		
E	.70 28(11)	.38 69(31)	.64 9(9)												.25 42(21)		.43 12(10)		.33 37(13)	
K						.60* 3(3)	.14 12(7)					.67 4(3)		.57 5(5)	.28 23(7)			.33 25(11)		.21 12(9)
R	.71 15(7)			.33* 1(1)	.40* 3(3)	** 0(0)	.50* 2(2)		.33* 1(1)		.80* 3(3)	.67* 1(1)	.60* 3(3)		.30 8(8)	.56 7(5)			.45 9(6)	

***Figure 5.1 Values of the Perturbation Probability index (Pp) for each mutation type**. The darker the colour, the higher the Pp value. Wild type residues are shown on the left, target residues at the top. For each mutation type the corresponding number of mutations reported in ProTherm and, among round brackets, the number of proteins in which mutations were tested are shown (* marks Pp computed from less than 4 mutations). Pp is not computed (**) for a mutation type when data are not available in ProTherm. Pp values are computed according to Equation 1. Residues are grouped as follows: apolar (GAVPLIM, red), aromatic (FWY, blue), polar (STCNQH, green), charged (DEKR, yellow), with residues in the same group listed at increasing order of size. Squares comprising mutations within the same residue group are coloured in red, blue, green and yellow, respectively. All the other off diagonal squares comprising mutations included in different groups are shown as a combination of the four colours.*

Table 5.1 Correlation of Pd and Pp values with different methods

Index§	**Type**	**r**	**r^2**	**P_r**	**r***	**r^{*2}**	**P_{r*}**	**Corr$_s$**	**P_s**	**Corr$_k$**	**P_k**
Pd /Pp	141	0.466	0.217	*5.7E-09*	0.640	0.409	*1.3E-17*	0.463	*7.2E-09*	0.325	*2.4E-08*
Pd° / Pp	141	0.489	0.239	*7.8E-10*	0.669	0.447	*1.3E-19*	0.495	*4.4E-10*	0.352	*1.5E-09*
Pd^/ Pp	141	0.491	0.241	*6.2E-10*	0.672	0.452	*7.4E-20*	0.495	*4.3E-10*	0.351	*1.6E-09*
Pd /Pp*	150	0.524	0.274	*6.1E-12*	0.728	0.530	*1.5E-25*	0.511	*2.3E-11*	0.365	*8.6E-11*
Pd°/Pp*	150	0.544	0.296	*6.2E-13*	0.753	0.568	*3.6E-28*	0.539	*1.2E-12*	0.387	*5.7E-12*
Pd^/Pp*	150	0.547	0.299	*4.7E-13*	0.757	0.572	*1.6E-28*	0.538	*1.3E-12*	0.386	*6.2E-12*

Pp=perturbation index (Equation 3. 1), Pd=disease index (Equation3. 2); Pd° = without somatic cancer mutations in the set (16.5% of the final set). Pd^= without cancer associated mutation in the set (18.6% of the final set). In the second set of rows the reference set is All (Table 3.1) and the number of mutations without somatic cancer and without cancer. In this case Pd is evaluated for the 150 possible mutation type and the missed 9 Pp values (Pp) are computed from the regression line in Figure 5.3. Methods: r=Pearson correlation coefficient (Equation 3.7); r^2= strength of r; P_r= significance (P-value) (Equation3. 8); r^*, r^{*2}, P_{r*} as before when standard errors of the data are considered (Equation 3.9); Corr$_s$, $P_{s:}$ Spearman correlation coefficient and the correspondent P-value; Corr$_k$, P_k: Kendall correlation coefficient and the correspondent P-value. Type: number of SNP mutation types.*

In order to confirm that our Pp index is a general estimator of the mutation effect on protein stability we correlated its values with a set of different matrices computed under different assumptions (Table 5.1). Pp values are correlated with symmetrical and asymmetrical substitution matrices. The correlation is also computed for Pp values obtained with a decreasing number of mutation pairs, depending on how many experiments were in ProTherm. Correlation values are lowering with symmetric than with asymmetric matrices. We obtained a reduced dataset by applying a constriction that at least 5 experiments must be available in ProTherm database for computing its Pp value for each mutation type. When a reduced number of Pp values is considered the correlation increased. In all the cases (Table 5.1) correlation is significant as indicated by the corresponding P-values. Considering the effect of standard errors on the correlation (Equation 3.9) regularly increases the correspondent value (4th column in Table 5.1).

Historical matrices such as the McLachlan's and the Grantham's ones are derived from amino acid physico-chemical properties, being the latter based on a mean chemical distance for each amino acid pair of three basic properties, such as overall chain composition, polarity and molecular volume [McLachlan (1971); Grantham (1974)]. For this reason, and only in this case we observe a positive correlation value, in all other cases, correlation is negative indicating that the higher the Pp value is the lower the probability of observing the targeted substitution as scored under different assumptions.

The importance of the physicochemical properties of the protein was described and exploited time to time in development of different prediction tools [Tophem CL et al (1997), Dimitri G. (2000), Guerois R (2002)]. In line with previous observations, the conservation of the physico-chemical properties is not the only important variable when considering mutation effects on protein stability. The relevance of polar solvent exposure and of the protein structure was also described. As shown in table 5.1 it was observed that when matrices based on properties directly derived from protein structures, such as the extent of amino acid exchangeability, and from the weight of the genetic code redundancies were considered [Johnson and Overington, 1993; Müller et al, 2002; Crooks and

Brenner, 2005], the correlation of the Pp index with the corresponding matrices increases with respect to that based on physico-chemical properties and chemical distances. Furthermore a good correlation is also observed when the weight of evolution is taken into account (BLOSUM 62).

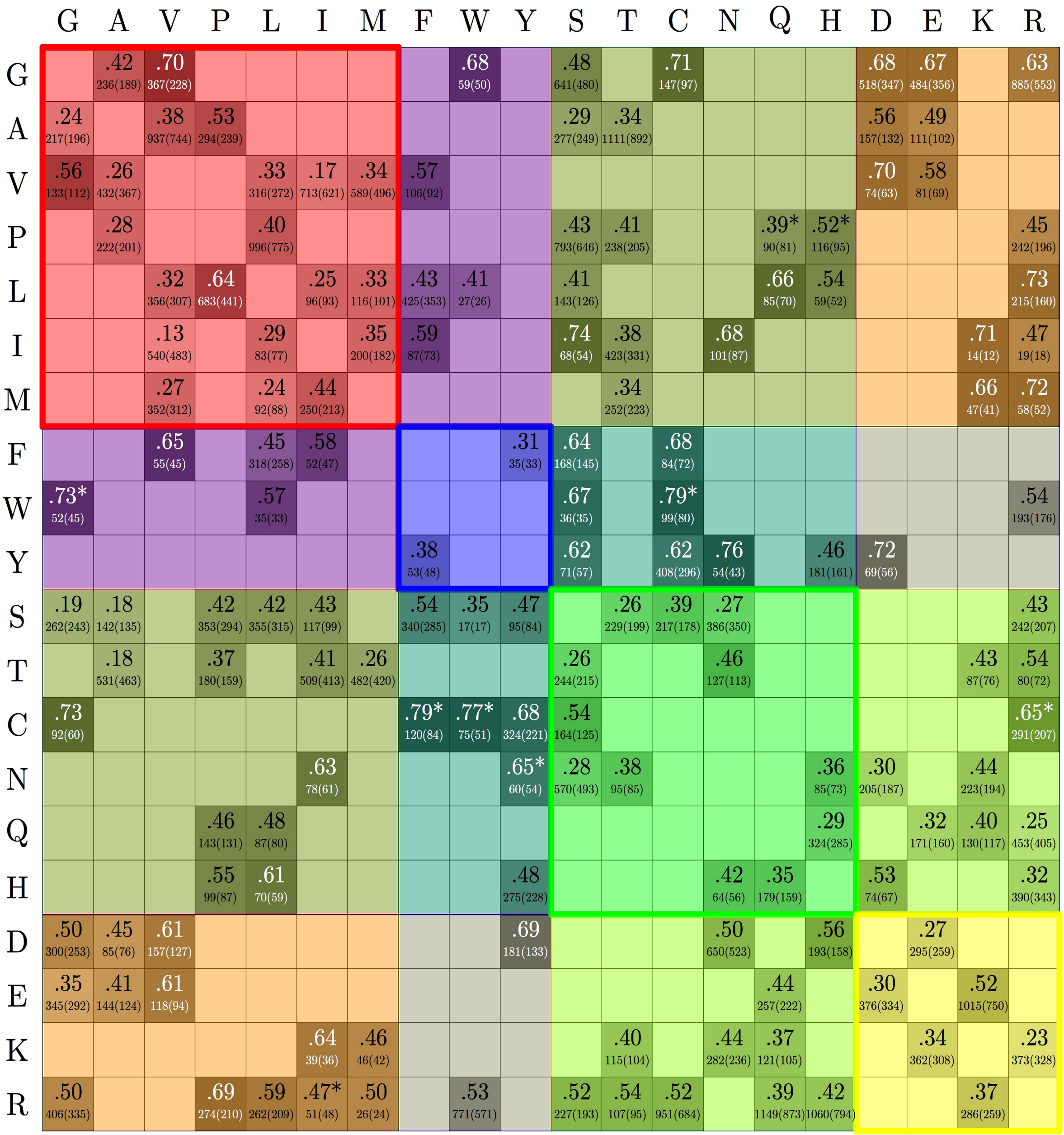
Interestingly enough and again in line with what previously discussed, mutations may also be asymmetrical in relation to their effect on protein stability. Pp correlation values increase when asymmetric substitution matrices are considered [Overington et al., 1992; Koshi and Golstein 1995; Yampolski and Stoltfus, 2005] and the best correlation is obtained with the environment-specific substitution matrix for accessible residues was taken into account [Overington et al., 1992]. Indeed some 32 mutations types (23% of the total) corresponding to 16 mutation pairs have a Pp difference $(|P(X \rightarrow Y) - P(Y \rightarrow X)|) > 0.3$, which is sufficient to include one of the two mutation types below or above the 0.5 perturbing threshold value (Figure 5.1). As explained earlier the residue mutation is considered to affect the protein stability mainly because of its effect on polarity and steric effect of the substituted residue. Considering Pp values and mutation types in Figure 5.1, we noticed that on average changing steric effect when a given side chain is mutated from small to large perturbs the protein stability more than changing its polarity.

5.2 Calculating the disease index (Pd)

For each mutation type of any residue type 'X' to residue type 'Y' (X→ Y) we computed associated probability of this mutation to actually cause a disease, and we called it as disease index (equation 3.2). SNPs without any annotation of a disease were considered to be non-damaging in our database (SNP-UniProtKB+COSMIC, Table 3.1). The disease index values for each mutation type are listed in Figure 5.2 with the corresponding number of mutations and number of proteins (in round brackets).

All the 150 possible SNP mutations types are present in the set. Wild type residues are aligned vertically on the left and mutant residues horizontally at the top. *. Residues are grouped as in Figure 5.1 apolar (GAVPLIM, red), aromatic (FWY, blue), polar (STCNQH, green), charged (DEKR, yellow). The residues falling in the same group are listed in the increasing order of their size. The index value ranges from the lowest value of 0.13 for I→V SNP up to the highest value of 0.79 for C→F and W→C SNP. The average Pd value is 0.47 with an average standard error of 0.04. Stars label the 9 Pd indexes without Pp counterparts that indicate a high/medium probability of disease association with the exception of P→Q. A Pd value of 0.5 can be considered a natural threshold to classify mutation types as endowed with low/medium and high probability of being disease associated.

There are 57% of the SNP types, which have Pd values less than 0.5. High Pd values (<0.5) are characteristics of mutation type where either the polarity is not conserved or the steric effect decreases upon residue substitution. The same characteristics were noted while analyzing the spectrum of Pp values. We computed the Pd correlation with the substitution matrices (Table 5.1, Pd values). Correlation values are routinely higher than those obtained with Pp. The values are higher even when the average standard errors for Pp and Pd were considered Correlation of Pd values with substitution matrices is high both for symmetric and asymmetric substitution scoring values. The highest correlation value is obtained with the Blosum62 substitution matrix, indicating that the more a mutation type is allowed through evolution the less it is conducive to disease. Summing up, for the Pd index different physico-chemical and protein structural features are conducive to Pd values as indicated by the significant correlation observed with scoring matrices derived under different assumptions (Table 5.1, Pd values).



	G	A	V	P	L	I	M	F	W	Y	S	T	C	N	Q	H	D	E	K	R
G		.42 236(189)	.70 367(228)						.68 59(50)		.48 641(480)		.71 147(97)				.68 518(347)	.67 484(356)		.63 885(553)
A	.24 217(196)		.38 937(744)	.53 294(239)							.29 277(249)	.34 1111(892)					.56 157(132)	.49 111(102)		
V	.56 133(112)	.26 432(367)			.33 316(272)	.17 713(621)	.34 589(496)	.57 106(92)									.70 74(63)	.58 81(69)		
P		.28 222(201)			.40 996(775)						.43 793(646)	.41 238(205)			.39* 90(81)	.52* 116(95)				.45 242(196)
L			.32 356(307)	.64 683(441)		.25 96(93)	.33 116(101)	.43 425(353)	.41 27(26)		.41 143(126)				.66 85(70)	.54 59(52)				.73 215(160)
I			.13 540(483)		.29 83(77)		.35 200(182)	.59 87(73)			.74 68(54)	.38 423(331)		.68 101(87)					.71 14(12)	.47 19(18)
M			.27 352(312)		.24 92(88)	.44 250(213)						.34 252(223)							.66 47(41)	.72 58(52)
F			.65 55(45)		.45 318(258)	.58 52(47)				.31 35(33)	.64 168(145)		.68 84(72)							
W	.73* 52(45)				.57 35(33)						.67 36(35)		.79* 99(80)							.54 193(176)
Y								.38 53(48)			.62 71(57)		.62 408(296)	.76 54(43)		.46 181(161)	.72 69(56)			
S	.19 262(243)	.18 142(135)		.42 353(294)	.42 355(315)	.43 117(99)		.54 340(285)	.35 17(17)	.47 95(84)		.26 229(199)	.39 217(178)	.27 386(350)						.43 242(207)
T		.18 531(463)		.37 180(159)		.41 509(413)	.26 482(420)				.26 244(215)			.46 127(113)					.43 87(76)	.54 80(72)
C	.73 92(60)							.79* 120(84)	.77* 75(51)	.68 324(221)	.54 164(125)									.65* 291(207)
N						.63 78(61)				.65* 60(54)	.28 570(493)	.38 95(85)				.36 85(73)	.30 205(187)		.44 223(194)	
Q				.46 143(131)	.48 87(80)											.29 324(285)		.32 171(160)	.40 130(117)	.25 453(405)
H				.55 99(87)	.61 70(59)					.48 275(228)				.42 64(56)	.35 179(159)		.53 74(67)			.32 390(343)
D	.50 300(253)	.45 85(76)	.61 157(127)							.69 181(133)				.50 650(523)		.56 193(158)		.27 295(259)		
E	.35 345(292)	.41 144(124)	.61 118(94)												.44 257(222)		.30 376(334)		.52 1015(750)	
K						.64 39(36)	.46 46(42)					.40 115(104)		.44 282(236)	.37 121(105)			.34 362(308)		.23 373(328)
R	.50 406(335)			.69 274(210)	.59 262(209)	.47* 51(48)	.50 26(24)		.53 771(571)		.52 227(193)	.54 107(95)	.52 951(684)		.39 1149(873)	.42 1060(794)			.37 286(259)	

***Figure 5.2 The disease index values (Pd) for each mutation type**. The darker the colour, the higher the Pd value. Wild type residues are shown on the left and target residues at the top. For each mutation type the corresponding number of mutations and, between round brackets, the number of proteins in which mutations are annotated, are shown. Pd is computed for the 150 possible SNPs; when for a given SNP type no thermodynamic data are available in Protherm, the corresponding Pd is marked with *. Residues are grouped as in Figure 5.1 apolar (GAVPLIM, red), aromatic (FWY, blue), polar (STCNQH, green), charged (DEKR, yellow), with residues in the same group listed at increasing order of size.*

5.3 Correlating Pp and Pd

The properties of the mutational spectrum are encoded into a pair of indices for each mutation type. The mutational spectrum can also be shown for 141 SNP types out of the 150 possible SNP mutations due to the lack of thermodynamic data for the rest of the 9 mutation types. We are now in the position of testing whether the two indexes (Pp and Pd) correlate. We find a linear correlation between Pp and Pd values for each mutation type (141 mutation types) (Figure 5.3), where all the data are plotted with the corresponding standard error as evaluated with bootstrapping method. The regression line is computed along with parallel lines drawn at a y distance equal to one and two fold/s the root mean square of the residuals.

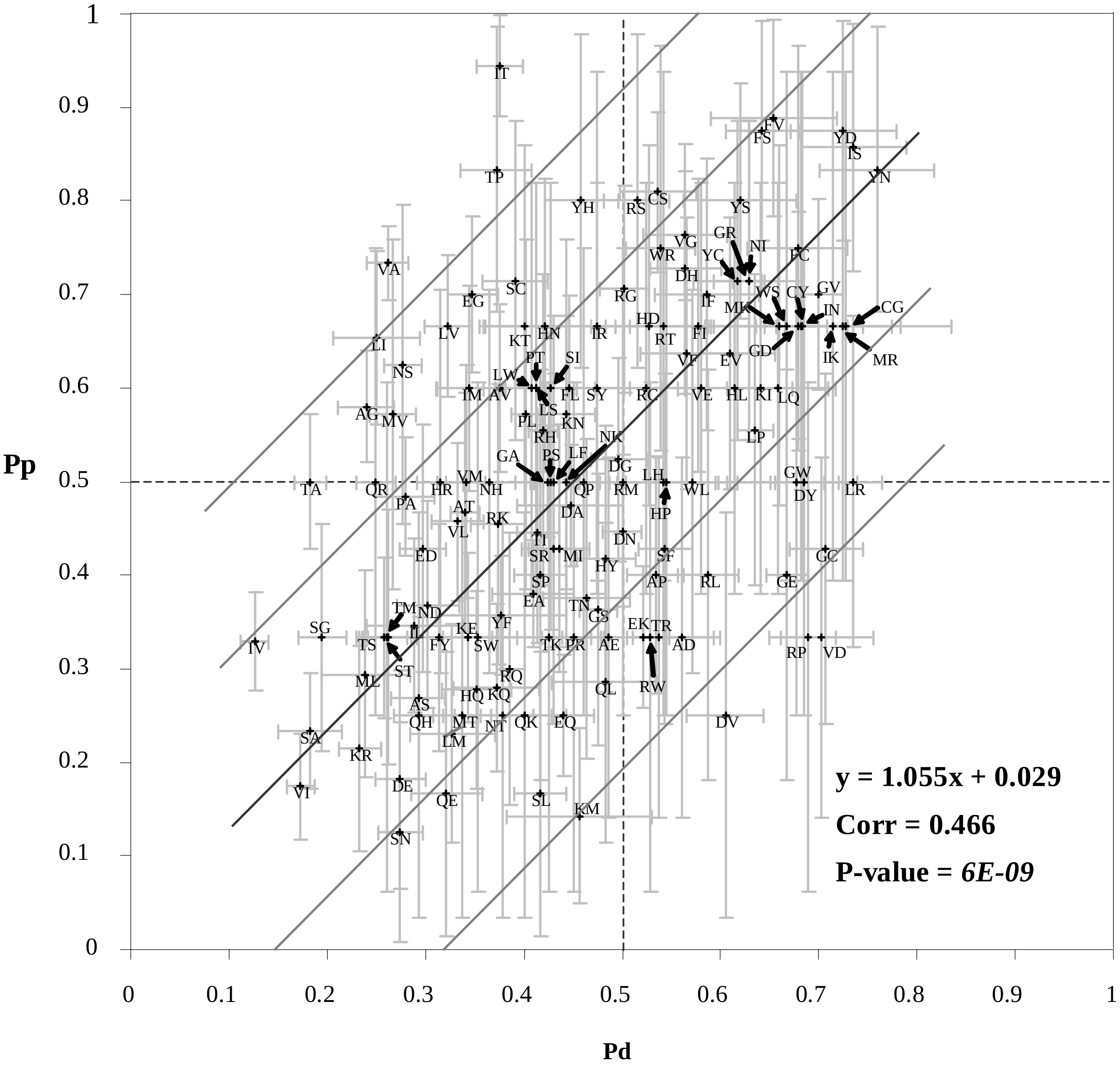


Figure 5.3 Correlation between Pd and Pp *Correlation among the disease (Pd) and the perturbing (Pp) indexes for 141 SNP mutation types. The regression line is computed with BLMS regression,(Materials and Methods). Pearson correlation (Corr) and P-value are computed with equation 3.9 and 3.8, respectively. Lines parallel to the fitting one are drawn at y distance equal to one/two fold the root mean square of the residuals (root mean square error (Equation 3.6), RMSE=0.181).*

Our results indicate that Pd and Pp significantly (with low P-values) correlate with different combinations as shown in Table 5.1. The strength of the correlation (r^2) is about 20% and it doubles when the standard errors of the data are considered. This value indicates that the correlation is moderate/good. The correlation increases when a large fraction of the data (about 18%, cancer related mutations) is removed from the data set (Figure 5.4).

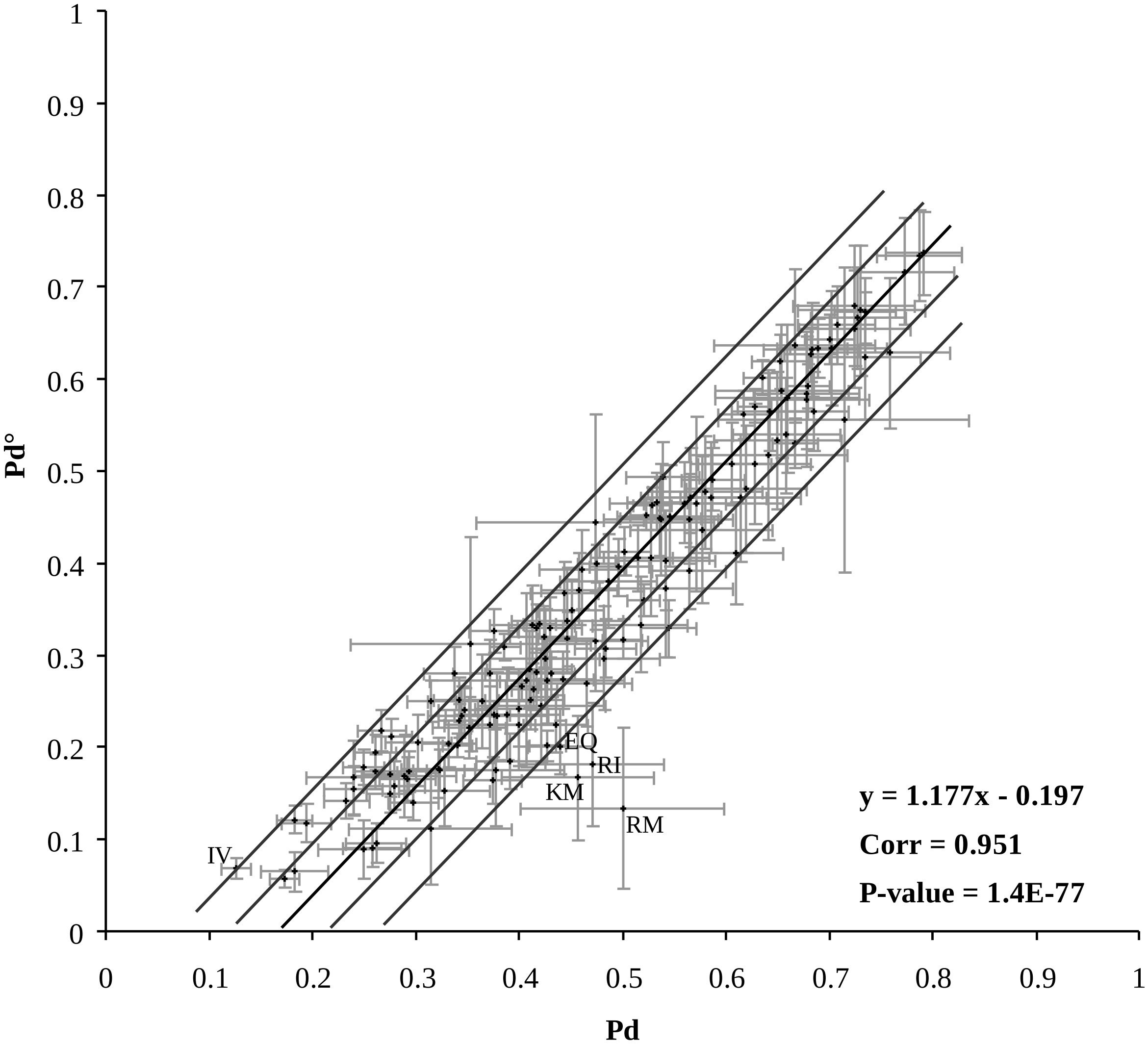


Figure 5.4 Correlation between Pd and Pd° values. *Pd values are computed for the 150 SNP types considering all the mutation data set (SNP-UniProtKB+COSMIC, Table 3.1); Pd°: mutations annotated as somatic cancer-related are removed (16.5% of the total mutations in SNP-UniProtKB+COSMIC; remaining mutations: 32,714) and Pd° is computed. The regression line is computed with a BMLS method (Materials and Methods). Lines parallel to the fitting one are drawn at a y distance equal to one and two fold the root mean square value of the residuals (root mean square error (Equation 3.6), RMSE=0.056).*

In Figure 5.5 A and B residuals (the differences between Pp(X→Y) and the corresponding fitting value on the fitting line BLMS(X→Y)), weighted with respect to the variances of the mutation type X→Y on both axes are plotted as a function of Pd and Pd° (values computed without somatic cancer mutations). Noticeably most of the mutation types have a weighted residual that clusters within one standard deviation from the average of the distribution value for both sets of values. In both cases two outliers are detected: V→A and I→T. We can conclude that for these mutation types the Pp value is much higher than that expected considering the Pd associated value. The correspondent Pd indexes (0.26 and 0.38 for V→A and I→T, respectively) indicate a moderate tendency to be disease related.

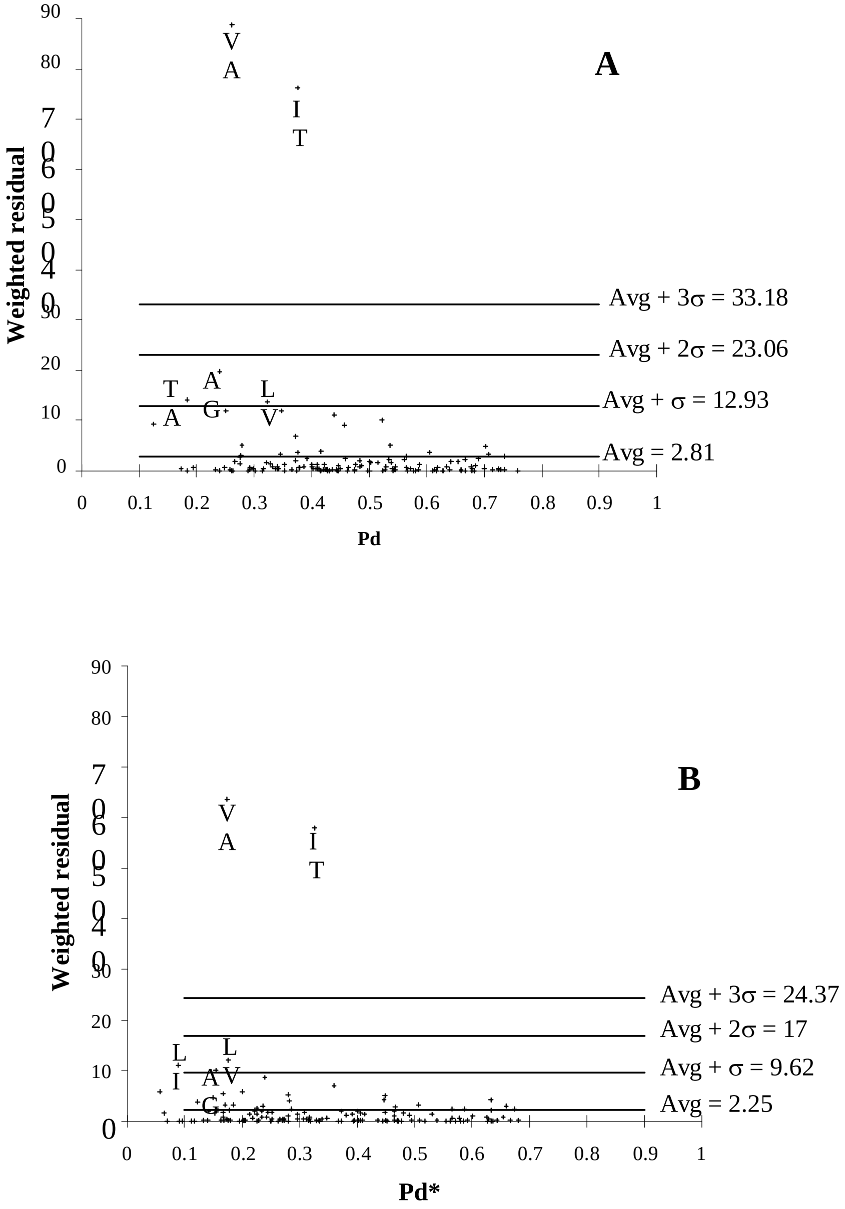


Figure 5.5 Weighted residual analysis of the Pd/Pp regression *Differences between Pp(X→Y) and the corresponding fitting value on the fitting line BLMS(X→Y), weighted with respect to the variance of the mutation type X→Y on both axes (Weighted residual) with respect to Pd values (Pd). (A) all the (SNP-UniProt+COSMIC) data base.(B) The analysis is restricted to diseases of genetic origin, with the exclusion of SNPs annotated as somatic cancer (Pd*). The mean value of the residuals (Avg) and the corresponding standard deviation (σ) are shown for reference. The mutation type (X→Y) is only shown for residual values >1σ.*

5.4 The NEW Predictor at work

The second part of this work consists of development of a predictor, to classify destabilizing and stabilizing mutations with respect to protein structure and its testing with our 1st result. We developed a SNP predictor (discussed in chapter 4). Our predictor is competent to discriminate between perturbing and non-perturbing mutation with an overall accuracy of 68%, which is highest among all the predictors involved in this study when tested starting from the protein structures. When predicting ΔΔG values associated with mutations, the Matthews correlation coefficient value is 0.32, which is best among all the other predictors (Table 5.2).

Table 5.2 Comparison between the performances of different predictors.

Threshold |ΔΔG|= 1.0 kcal/mol	**MCC**	**Q**	**Specificity_** *perturbing*	**Sensitivity** *perturbing*	**Specificity** *non-perturbing*	**Sensitivity** *non-perturbing*
I Mutant 2.0	0.17	0.57	0.55	0.76	0.63	0.39
I Mutant 3.0	0.17	0.54	0.46	0.78	0.73	0.38
Fold-X	0.24	0.58	0.51	0.81	0.75	0.41
Popmusic 2.0	0.28	0.66	0.69	0.79	0.61	0.48
NEW	**0.32**	**0.68**	**0.69**	**0.81**	**0.64**	**0.49**

** For scoring index definition see above.*

Strengths of NEW predictor:

(1) It can be concluded from Table 5.2 that our NEW predictor is best among all other state of the art predictors, whether they are energy function based predictors like PoPMuSiC and Fold-X or Machine Learning based predictors like I Mutant 2.0 and I Mutant 3.0.

(2) Secondly in our later sections it has been proved that the predictions of our NEW predictor is in line with the probability indices calculated as the first part of the work.

5.4.1 Computing Pp and Pd for the 3D Structure dataset

We calculate the perturbing index values (equation 3.1) and disease index values (equation 3.2) over the 3D structure dataset used to train/test the NEW predictor. This corroborates the view that the predictive method is capable of correctly annotating the mutation as neutral and disease related. Indeed the results obtained by computing the probability indexes and their correlation prove that protein perturbation in the data bases correlates with probability of being disease-related of a given mutation. With the NEW predictor we can now compute the extent of perturbation of a given mutation starting from the protein structure. Also the mutation is or is not disease related. By correlating perturbing index from the predicted data and the probability disease index over the 3D structure data set we are now in the position of validating our predictor performances.

Table 5.3 The 3-D structure-SNP dataset

	Polymorphic	**Common in polymorphic and disease related mutations**	**Disease related Mutations**
No. of mutations (Total =7300)	2318 (32%)	-	4982 (68%)
No. of PDB chains (Total= 1185)	803 (68%)	207 (17%)	175 (15%)
No. of residues according to PDB (Total=327035)	218578 (67%)	67881 (21%)	40576 (12%)

The 3-D structure-SNP dataset was created after excluding all the non-SNP data from the 3-D Structure dataset used for the predictor (table 3.3)

The 3D Structure-SNP dataset contains only those SNP related mutation examples for which the 3-D structures are known and comprises 7300 mutations in 1185 proteins. 2318 (32%) of the total SNP mutations were polymorphic and rest 68% (4982) SNP mutations were disease related. 68% of the protein structures were having only polymorphic mutations and 15% of protein structures were having only disease related mutations, whereas 17% protein structures were endowed with both disease related as well as polymorphic mutations (Table 5.3).

5.4.2 Correlation between calculated perturbing index (Pp) and predicted perturbing index (Pp-3D)

We calculated perturbing index values (Pp) (equation 3.1) for our 3 D structure–SNP dataset, based on the predicted ΔΔG values by our NEW predictor (Pp-3D). The Pp-3D values are well correlating (Figure 5.6) with the previously calculated Pp values for the whole mutation data set (ProTherm database (released on 31 March, 2010), Table 3.1) (in Figure 5.1). The regression line is calculated using the BMLS method The two perturbing values are significantly correlating with a Pearson correlation coefficient of 0.49 and a low P-value of 6E-10 which indicates the significance of this correlation.

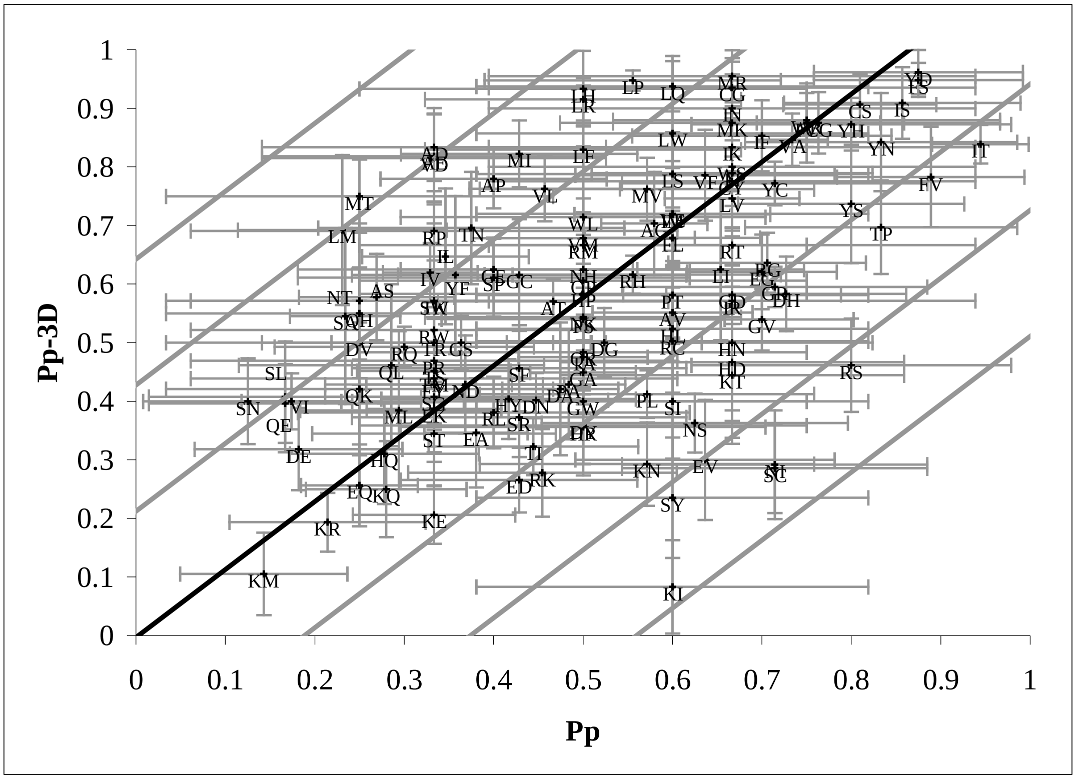


__Figure 5.6 Correlation between calculated perturbing index value (Pp)__ (equation 3.1) and perturbing index values calculated with predicted ΔΔG by new predictor (Pp-3D). The regression line is computed with a BMLS method (Materials and Methods). Pearson correlation coefficient = 0.49, P-value= 6E-10

5.4.3 Correlation between disease index (Pd) values for the whole mutation dataset and disease index values for the 3D Structure-SNP dataset (Pd-3D)

The disease index values for the 3D structure dataset (Pd-3D) are well correlating with the disease index values calculated for the whole mutation dataset (SNP-UniProtKB+COSMIC, table 3.1) (Figure 5.7). The disease index values for two datasets are correlating with Pearson correlation coefficient of 0.75 and a P-value of 9E-28. The regression line is computed with a BMLS method (see Materials and Methods). Interestingly enough it turns out that the reduced set including proteins endowed with a 3D structures collects a relative abundance of disease related SNPs higher than that in the whole data base of mutations. This is noticeable from the intercept of the regression line in Figure 5.7 (different from zero).

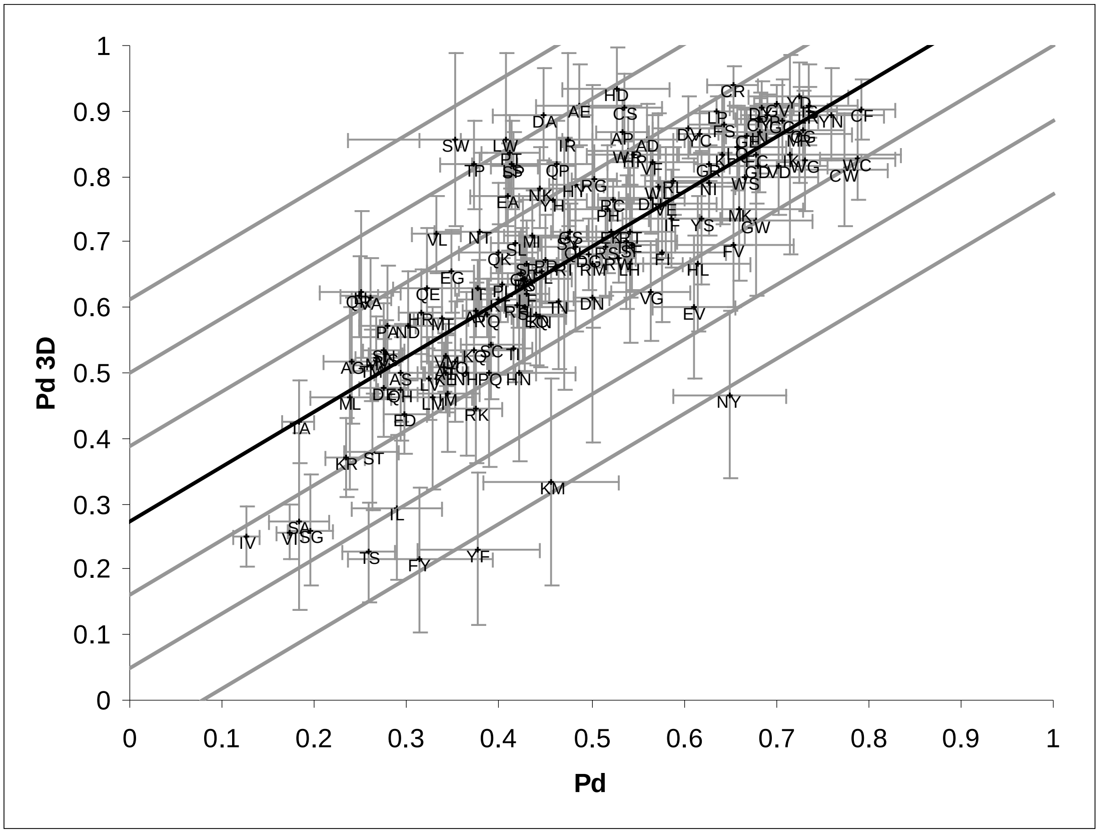


Figure 5.7 Correlation among Pd and Pd-3D values Pd values are computed for the 150 SNP types considering all the mutation data set (SNP-UniProtKB+COSMIC (table 3.1)); Pd-3D values computed for our 3D-structure dataset (used for our predictor) mutations annotated as somatic cancer-related. The regression line is computed with a BMLS method. Pearson correlation coefficient= 0.75, P-value = 9E-28.

5.4.4 Correlation between Pp-3D and Pd-3D

Finally the two indices (Pp-3D and Pd-3D) are well correlated (low P-value (4E-08)) with a Pearson correlation coefficient of 0.44. This finding supports the notion that indeed a correlation exists among Pd and Pp that is independent from the set of the data set (Figure 5.8) and more importantly that our NEW predictor gives predictions that are on line with what present in the mutation data set.

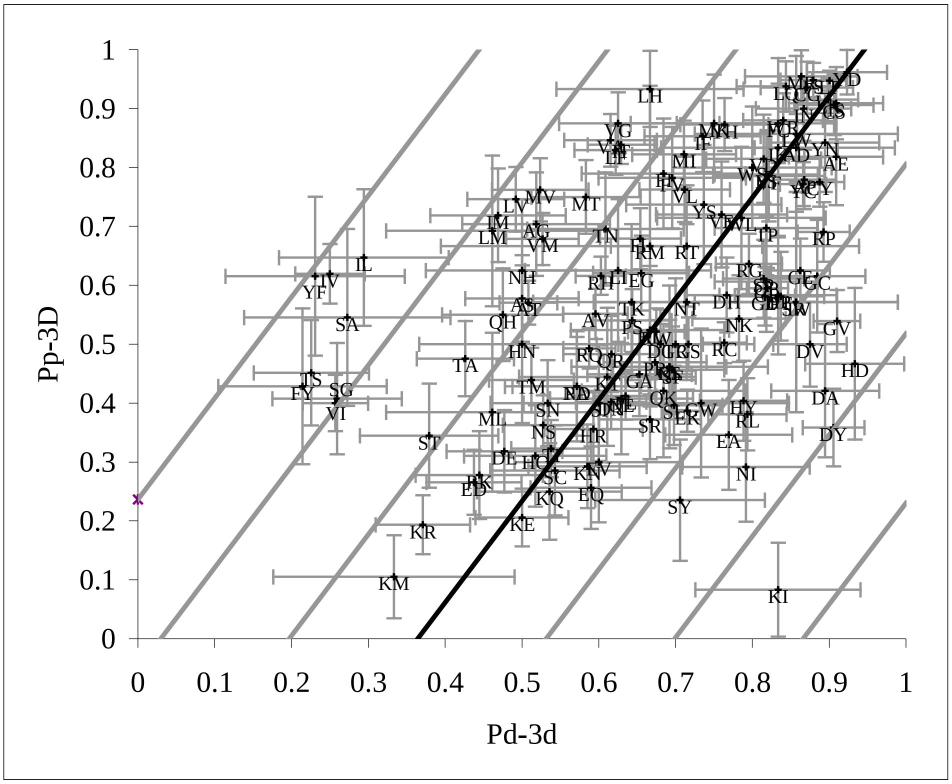


*Figure 5.8 **Correlation between Pp-3D and Pd-3D values** Pp-3D values are computed for the 141 SNP types considering only 3D-Structure-SNP dataset (table 5.3); Pd-3D values are disease index computed for the same dataset. Pearson correlation coefficient= 0.44, P-value= 4E-08, RMSE = 0.29.*

Chapter 6. Conclusions

SNP annotation is an issue in many deep sequencing experiments aiming at clinical association studies. In this work and for the first time we tested the hypothesis that mutations conducive to disease are also perturbing to some extent the stability of the corresponding protein. In this work we have elaborately proved the direct correlation of mutations related to disease and destabilization of the related protein. The problem is addressed separately in the pertinent literature. A number of predictors are available to predict if a particular mutation is conducive to a disease [Tatvigian et al., (2008)] or a particular mutation is perturbing the protein stability [Thusberg and Vihinen, (2010)]. Our present effort aims at casting information of the effect of mutation types into features derived from the available databases of thermodynamic data (from ProTherm) and SNP description data (from UniProtKB) at the protein level. For each mutation type we computed two probability indices, perturbing index (Pp) and disease index (Pd) that represent the probability of promoting perturbation of the protein stability and disease, respectively.

At the very first step of our analysis we compute Pp and Pd maps for the data derived from the Universal Protein Resource (UniProt KB) and ProTherm. By this each mutation type is endowed with Pd and Pp indexes. The probability indices are computed over a large number of mutations for all the mutation types both in relation to mutational data and also to thermodynamic data, when possible. The standard error associated to each Pd and Pp values estimates the variability due to sampling fluctuation. Given the different volume of available data, Pp has variability (Average Standard Error) fourfold larger than Pd (0.16 vs. 0.04).

In the second step of our analysis we proved that the two indices are linearly correlated with only two outliers namely V→ A and I→ T. This is obtained with a robust procedure that takes standard errors on both indices into account [Del Rio et al., (2001)]. This ensures that the hypothesis of the direct relation between a disease causing mutation with the protein stability perturbation cannot be rejected (with a strength of about 25%) and that it is likely to be one of the possible molecular mechanisms in the whole diseasome [Goh et al., 2007]. Further the correlation computation considering only germ line mutations (Pd°) is still significant, with the same outliers (V→ A and I→ T). This proves that data are not biased by somatic cancer mutations, including some 18% of the total mutation data set.

When disattenuation is evaluated [Spearman 1904/1987] and errors are taken into account, results indicate that data are compatible with an increased correlation coefficient value (from 0.466 to 0.640 (please refer table 5.1). Also it is important to point out here that, our error estimator indicates, probabilistic indices and their correlation are enough tolerant to increment in database volumes that is most likely to occur in the near future giving the great shift of interest in SNP detection.

Two outliers V→ A and I→ T were also detected. For these mutation types the Pp values are 0.73 and 0.94 for V→ A and I→ T respectively are much higher than that expected considering the associated Pd values 0.26 and 0.38 for V→A and I→T respectively. These values indicate a moderate tendency of these mutation types to be disease related. In ProTherm, mutational data (122 V→A mutations in 37 proteins, and 16 I→T mutations in 11 proteins, respectively) are derived from proteins whose structure is known with atomic resolution. In both cases about 90% of the mutations are located in the protein core region (buried) and rather perturbing as indicated by the Pp values. This sampling is different from average in our data where only 40% of the mutations are buried and it can bias this association with low Pd values.

With this statistical analysis we can conclude that all together our results support the notion that the perturbing Pp index, as derived from thermodynamic data of mutation effects on proteins, is indicative of the effect that a mutation type can have on protein stability. The Pp value sums up physico-chemical properties such as solvent exposure, structure conservation and also conservation through evolution. Finally the Pp mutation index is enough general and representative of the effect of a mutation type on protein stability in spite of the few thermodynamic data presently available in the ProTherm database as compared to the amount of data describing the effect of mutations on human health and generated as a result of high throughput sequencing methods.

All these observations are crucial in suggesting which types of features are relevant for predicting whether a mutation in a protein is or is not disease related. Our NEW predictor, based on support vector machines, outperforms existing methods by considering 105 input features. It computes whether a mutation is or is not destabilizing the protein structure with an overall accuracy of 68% and Matthew correlation coefficient of 0.33. From the predicted values we are also able to derive a perturbing index that we call the predicted perturbing index. The “predicted” perturbing index well correlates with the one computed over the entire set of thermodynamic data and with the disease related index computed over the set of 3D structures presently available. This result support therefore the idea that predictors capture much information available in the data sets and can be adopted as reliable methods in annotating disease-related mutations.

With the advent of next generation sequencing machines our computational methods will be necessary to complement experimental results with efficient and robust systems to endow data with labels useful to perform diagnosis and further translate to the clinics. Therefore our effort contributes to this line of research and provides support to the notion that machine learning based approaches can help in generating valuable tools for biomedical data analysis. Also in the field of protein engineering our NEW predictor can help in designing new proteins with higher stability a resting to temperature changes. This is particularly useful in the field of bioreactor production for generating cell factories competing for the productions of molecules of interest such as drugs, in the field of biosensors, and nanotechnology for the development of new materials.

References.

Altmann KH, Wojcik J, Vasquez M and Scheraga HA (1990). Helix-coil stability constants for the naturally occurring amino acids in water. XXIII. Proline parameters from random poly (hydroxybutylglutamine-co-L-proline). *Biopolymers,* **1**, 107-120

Baldi P, Brunak S(2001) The Machine Learning Approach. *Massachusetts Institute of Technology press.*

Bartoli L, Montanucci L, Fronza R, Martelli PL, Fariselli P, Carota L, Donvito G, Maggi G, Casadio R. (2009). The Bologna Annotation Resource: a non-hierarchical method for the functional and structural annotation of protein sequences relying on a comparative large-scale genome analysis. *J Proteome Res*, **8**, 4362- 4371.

Capriotti E, Fariselli P, Calabrese R, Casadio R. (2005) I-Mutant2.0. predicting protein stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Res*. 33(Web Server issue): W306-W310.

Capriotti E, Fariselli P, Rossi I, Casadio R. (2008). A three-state prediction of single point mutations on protein stability changes. *BMC Bioinformatics* **9** (Suppl 2): S6

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**, 2729-2734.

Catherine L. Worth, G. Richard, J. Bickerton, Adrian Schreyer, Julia R. Forman, Tammy M. K. Cheng, Semin Lee, Sungsam Gong, David F. Bruke and Tom L. Blundell (2007). A structural bioinformatics approach to the analysis of non synonymous single nucleotide polymorphisms (nsSNPs) and their relation to disease. *J Bioinform Comput Bio*, **5**, 1297-1318

Chakrabartty A, Schellman JA and Baldwin RL (1991). Large differences in the helix propensities of alanine and glycine. *Nature,* **351**, 586 –588.

Chen J and Stites WE (2001). Packing is a key selection factor in the evolution of protein hydrophobic cores. *Biochemistry*, **40**, 15280- 15289.

Cho JH and Raleigh DP (2006). Electrostatic interactions in the denatured state and in the transition state for protein folding: effects of denatured state interactions on the analysis of transition state structure. *J Mol Biol*, **359** (1437-1446).

Chou PY and Fasman GD (1974). Prediction of protein conformation. *Biochemistry* **13**: 222-245

Cristianini N, Shawe-Taylor J (2007) Introduction to Support Vector Machine. *Cambridge University Press*.

Crooks GE, Brenner SE. (2005). An alternative model of amino acid replacement. *Bioinformatics*, **21**, 975-980.

Del Río FJ, Riu J, Rius FX. (2001).Linear regression taking into account errors in both axes in the presence of outliers. *Anal Lett* **, 34: 14,** 2547-2561

Gilis D. and Rooman. M (2000). PoPMuSiC, an algorithm for predicting protein mutation stability changes. Application to prion proteins. *Protein Eng*. **13**, 849-856

Dosztanyi Z, Fiser A and Simon I (1997). Stabilization centers in proteins: Identification, characterization and predictions. *J Mol Biol.***, 272,** 597-612.

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

Durbin R, Eddy SR , Krogh A, Mitchison G. (1998). Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. *Cambridge University Press*

Feyereisen C, Morcellet M and Loucheux C (1977). Preferential and absolute adsorption to poly[N5-(3-hydroxypropyl)-L-glutamine] in water/2-chloroethanol solvent mixtures. *Macromolecules,* **10,** 485-488.

Fischer S, Verma CV (1999) Binding of buried structural water increases the flexibility of proteins. *PNAS*, **17**, 9613 –9615.

Forbes SA, Tang G, Bindal N, Bamford S, Dawson E, Cole C, Kok CY, Jia M, Ewing R, Menzies A, Teague JW, Stratton MR, Futreal PA. (2010). COSMIC (the Catalogue of Somatic Mutations in Cancer): a resource to investigate acquired mutations in human cancer. *Nucleic Acids Res,* **38** D652-657

Giver L, Gershenson A, Freskgard PO and Arnold FH (1998). Directed evolution of a thermostable esterase. *Proc Natl Acad Sci U S A,* **95,** 12809-12813.

Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabási A L. (2007). The human disease network. *Proc Natl acad Sci USA*, **104**, 8685-8690.

Grantham, R. (1974). Amino acid difference formula to help explain protein evolution. *Science* **185** 862-864

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

Hess ST, Blake JD, Blake RD. (1994). Wide variations in neighbor-dependent substitution rates. *J. Mol Biol,* **236**, 1022–1033.

Hill T, Lewicki P. (2007) STATISTICS Methods and Applications. StatSoft, Tulsa, USA. http://www.statsoft.com/textbook/.

Jimin Pei and Nick V. Grishin (2004). Combining Evolutionary and Structural Information for Local Protein Structure Prediction. *Proteins*, **4**, 782-794.

Johnson MS, Overington JP. (1993). A structural basis for sequence comparisons. An evaluation of scoring methodologies. *J Mol Biol* **, 233,** 716-738**.**

Kabsch W, Sander C (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers,* **22,** 2577-2637.

Kawashima S, Pokarowski P, Pokarowska M, Kolinski A, Katayama T, Kanehisa M. (2008). AAindex: amino acid index database progress report 2008. *Nucleic Acids Res,* **36** D202-205

Kilosanidze GT, Kutsenko AS, Esipova NG and Tumanyan VG (2004).Analysis of forces that determine helix formation in proteins. *Protein Sci* **, 13,** 351-357.

Koshi JM, Goldstein RA. (1995). Context-dependent optimal substitution matrices. *Protein Eng*, **8,** 641-645

Kossiakoff AA, Sintchak MD, Shpungin J, Presta LG. (1992). Analysis of solvent structure in proteins using D2O-H2O solvent maps: Pattern of primary and secondary hydration of trypsin. *Proteins*, **12**, 223-236.

Kotelchuck D and Scheraga HA (1969). The influence of short-range interactions on protein onformation. II. A model for predicting the helical regions of proteins. *Proc Natl Acad Sci U S A,* **62,** 14-21.

Kumar MD, Bava KA, Gromiha MM, Parabakaran P, Kitajima K, Uedaira H, Sarai A. (2006) ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions. *Nuleic Acids Res,* **34,** D204-6.

Lee J, Dubey VK, Longo LM and Blaber M (2008). A logical OR redundancy within the Asx-Pro-Asx-Gly type I .-turn motif. *J Mol Biol,* **377,** 1251-1264.

Lesk AM and Chothia C (1980). How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. *JMol Biol* ,**136,** 225-270.

Lesk AM, (2008) Introduction to Bioinformatics (3rd edition), *Oxford University Press.*

Levitt M (1978) Conformational preferences of amino acids in globular proteins. *Biochemistry,* **17,** 4277-4285.

Madan B, Lee B. (1994) Role of hydrogen bonds in hydrophobicity: the free energy of cavity formation in water models with and without the hydrogen bonds. *Biophysical Chem.*, **8**, 279-289.

Malkov SN, Zivkovic MV, Beljanski MV, Hall MB and Zaric SD (2008) A reexamination of the propensities of amino acids towards a particular secondary structure: classification of amino acids based on their chemical structure. *J Mol Model*, **14,** 769-775

Mark Lorch, Jody M. Mason, Richard B. Sessions, Anthony R. Clarke (2000) Effects of mutations on the thermodynamics of a protein folding reaction: Implications for the mechanism of formation of the intermediate and transition states. *Biochemistry*, **39**, 3480-3485.

Mclachlan A. D. (1971) Tests for comparing related amino-acid sequences. Cytochrome c and cytochrome c 551 . *J Mol Biol,* **61,** 409-424.

Minor DL, Jr. and Kim PS (1996) Context-dependent secondary structure formation of a designed protein sequence. *Nature* **, 380,** 730-734**.**

Mirsky AE and Pauling L (1936) On the Structure of Native, Denatured, and Coagulated Proteins. *Proc Natl Acad Sci U S A,* **22,** 439-447.

Moore DS, McCabe, GP. (2006) Introduction to the practice of statistics (5th ed.) *W.H. Freeman and Company. USA*

Müller T, Spang R, Vingron M. (2002) Estimating amino acid substitution models: a comparison of Dayhoff's estimator, the resolvent approach and a maximum likelihood method. *Mol Biol Evol* **19,** 8-13**.**

Myers JL, Well AD (2003) Research Design and Statistical Analysis (2nd edition), *Routledge, Taylor and Francis Group, USA.*

Overington J, Donnelly D, Johnson MS, Sali A, Blundell TL. (1992) Environment-specific amino acid substitution tables: tertiary templates and prediction of protein folds., *Protein Sci*, **1**, 216-226.

Pace CN (1975) The stability of globular proteins. *CRC Crit Rev Biochem***, 3,** 1-43.

Pauling L, Corey RB and Branson HR (1951) The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. *Proc Natl Acad Sci U S A,* **37**, 205-211.

Privalov PL and Khechinashvili NN (1974) A thermodynamic approach to the problem of stabilization of globular protein structure: a calorimetric study. *J Mol Biol,* **86,** 665-684.

Rashin A, Rashin BH, Rashin A, Abagyan R (1997). Evaluating the energetics of empty cavities and internal mutations in proteins. *Protein Sci,* **6**, 2143 – 2158.

Ruvinov S, Wang L, Ruan B, Almog O, Gilliland GL, Eisenstein E and Bryan PN (1997) Engineering the independent folding of the subtilisin BPN' prodomain: analysis of two-state folding versus protein stability. *Biochemistry***, 36,** 10414-10421.

Schiffer M and Edmundson AB (1967) Use of helical wheels to represent the structures of proteins and to identify segments with helical potential. *Biophys J,* **7**, 121-135.

Smith CK and Regan L (1997) Construction and design of b-sheets. *Acc Chem Res,* **30**, 153–161

Spearman C. (1910). Correlation calculated from faulty data. *Br J Psychol* **3**, 217-295

Strop P, Marinescu AM and Mayo SL (2000) Structure of a protein G helix variant suggests the importance of helix propensity and helix dipole interactions in protein design. *Protein Sci,* **9,** 1391-1394.

Taddei N, Chiti F, Fiaschi T, Bucciantini M, Capanni C, Stefani M, Serrano L, Dobson CM and Ramponi G (2000) Stabilisation of .-helices by site-directed mutagenesis reveals the importance of secondary structure in the transition state for acylphosphatase folding. *J Mol Biol* **, 300,** 633-647.

Takano K, Yamagata Y and Yutani1 K (2003). Buried water molecules contribute to the conformational stability of a protein. *Protein Eng,* **16**, 5-9.

Tavtigian SV, Greenblatt MS, Lesueur F, Byrnes GB (2008) IARC Unclassified Genetic Variants Working Group. In silico analysis of missense substitutions using sequence-alignment based methods. *Hum Mutat* **29,** 1327-36.

Thornton JM (1981) Disulphide bridges in globular proteins. *J Mol Biol,* **151**, 261-287

Thusberg J, Vihinen M. (2009) Pathogenic or not? And if so,then how? Studying the effects of missense mutations using bioinformatics methods. *Hum Mutat* **30,** 703-14.

Topham CM, Srinivasan N, Blundell TL, (1997) Prediction of the stability of protein mutants based on structural environment-dependent amino acid substitution and propensity tables. *Protein Eng*., **10**, 7-21.

Vihinen M (1987) Relationship of protein flexibility to thermostability. *Protein Eng,***1,** 477-480.

Villegas V, Viguera AR, Aviles FX and Serrano L (1996) Stabilization of proteins by rational design of .-helix stability using helix/coil transition theory. *Fold Des,* **1**, 29-34.

Vitkup D, Sander C, Church GM. (2003) The amino-acid mutational spectrum of human genetic diseases. *Genome Biol,* **4,** R72-72.10

Vogl T, Jatzke C, Hinz HJ, Benz J and Huber R (1997) Thermodynamic stability of annexin V E17G: equilibrium parameters from an irreversible unfolding reaction. *Biochemistry* **36,** 1657-1668.

Wade, R.C., M.H. Mazor, J.A. McCammon, F.A. Quiocho. (1991) A Molecular Dynamics Study of Thermodynamic and Structural Aspects of the Hydration of Cavities in Proteins. *Biopolymers*, **8**, 919-931.

Yampolsky LY, Stoltzfus A. (2005) The exchangeability of amino acids in proteins. *Genetics,* **170**, 1459-1472

Yang AS and Honig B (1995) Free energy determinants of secondary structure formation: I. -Helices. *J Mol Biol,* **252,** 351-365.

Yip YL, Famiglietti M, Gos A, Duek PD, David FP, Gateau A, Bairoch A. (2008) Annotating single amino acid polymorphisms in the UniProt/Swiss-Prot knowledgebase. *Hum Mutat,* **29**,361-6.