FEBS Letters 580 (2006) 2071–2080

Understanding mutations and protein stability through tripeptides

Sharmila Anishetty[a], Ramesh Anishetty[b,*], Gautam Pennathur[a,c,*]

[a] *Centre for Biotechnology, Anna University, Chennai 600 025, India*
[b] *The Institute of Mathematical Sciences, Taramani, Chennai, Tamil nadu 600 113, India*
[c] *AU-KBC Research Centre, Anna University, Chennai 600 044, India*

Received 3 January 2006; revised 13 February 2006; accepted 27 February 2006

Available online 10 March 2006

Edited by Miguel De la Rosa

Abstract A novel methodology to predict the local conformational changes in a protein as a consequence of missense mutations is proposed. A pentapeptide at the locus of mutation plays the dominant role and it is analyzed in terms of tripeptides. A measure for spatial and temporal fluctuations in a pentapeptide is devised and validated. The method does not involve any prior knowledge of structural templates from sequence homology studies. Structural deformations can be predicted with about 70–80% reliability in any protein. Disease causing mutations and benign mutations have been addressed. In particular, p53, retinoblastoma protein and lipoprotein lipase are studied in detail.

Keywords: SNPs; Missense mutation; Tripeptide; Protein structure; Disease; Structural biology

1. Introduction

Inherited differences in DNA sequence contribute to phenotypic variation, influencing an individual's anthropometric characteristics, disease susceptibility and response to the environment [1]. About 90% of sequence variants in humans are due to differences in single bases of DNA called single nucleotide polymorphisms (SNPs) [2]. SNPs are DNA allelic variants, which arise out of single nucleotide substitutions and have an appreciable allele frequency in a population [3]. Non-synonymous SNPs in the coding regions of genes (nsSNPs) or in regulatory regions are more likely to cause functional differences than SNPs elsewhere [4]. The availability of SNP data in public databases like dbSNP and locus centric mutation databases has made a systematic analysis of SNP data possible. nsSNPs, which occur near functionally important sites of a protein, are most likely to cause structural deformation of the protein leading to disease. However, not all nsSNPs are associated with disease. In instances where a mutation is found at an important site of a protein and yet does not result in a disease phenotype, it is realized that there are alternate mechanisms/proteins [5] in the organism masking the deleterious effect of the mutation. To understand the relationship between genetic and phenotypic variation, it is essential to assess the structural consequences of the nsSNPs in proteins.

Several studies have addressed the effect of nsSNPs on protein structure and function by mapping missense mutations to three dimensional structures and homology models. There are numerous reports of site directed mutagenesis experiments in the literature, which shed light on the structural and possible functional consequences of non-synonymous cSNPs. In one comprehensive structure based study, known disease mutations were mapped onto three dimensional structures of proteins in order to qualify how a disease phenotype can be explained by a destructive effect on protein structure or function [6]. 70% of the disease causing mutations mapped to structurally and functionally important sites of proteins. In another study [7], the structural features of a number of non-synonymous cSNPs were studied and a model for assigning a mechanism of action of each mutation in the protein was developed. Allelic variants were classified as neutral or deleterious. The variants were further classified as affecting protein stability, binding, catalysis, allosteric response and post translational modifications. Most of the structure related studies rely on the availability of 3D structures or templates for homology based predictions. Typically proteins, which are atleast 40% homologous to the one being modeled, are used as structural templates. This is a severe restriction to model builders in understanding observed mutations. Structural neighbourhood models and phylogenetic information was also used to derive features, which can serve as indicators of nsSNPs effect on function [8]. Purely sequence based approaches have also been employed in assessing the effect of nsSNPs. One such approach SIFT uses sequence homology to predict whether an amino acid substitution will affect protein stability and function and hence potentially alter the phenotype [9,10]. The relative strengths and complementarity of the evolutionary and structural models was also investigated [11]. The performance of classifiers was characterized as a function of the number of homologs available for the calculation of evolutionary features. It was observed that when fewer homologous sequences were available, structural information substantially improved the prediction accuracy of deleterious mutations. With the advent of the genomic era, understanding mutations at an automated pace has become a necessity [12].

When a protein undergoes a single amino acid change, it is expected that its shape will be deformed at the same locus. However this small perturbation does affect the neighbouring residues as well. The overall shape of the protein may remain unchanged, but a short stretch of about 10 residues may undergo a small deformation. In vivo, when the protein interacts with other molecules the local fluctuations of the conformation can also dominate the interaction. Therefore a mutation can

*Corresponding authors. Fax: 91 44 22541586.
E-mail addresses: ramesha@imsc.res.in (R. Anishetty), pgautam@annauniv.edu (G. Pennathur).

0014-5793/$32.00
doi:10.1016/j.febslet.2006.02.079

alter its wild type characteristics in terms of reaction rates also. In this study we attempt to predict the deformation and fluctuation (temporal deformation) of the conformation from the knowledge of short peptides in a protein such as tripeptides.

Tripeptide is the smallest unit, which captures the bending of the main chain of a protein. By concatenating overlapping tripeptides we can build the entire protein chain. At the locus of mutation, three tripeptides are affected. Consequently, if any of the three disturb the wild type behaviour, structural and temporal deformations can be seen. This picture shows that in any protein, a pentapeptide unit, which has three different tripeptides, plays the dominant role, which we purport to understand.

A tripeptide for example "ANR" might occur in various proteins and each of these creates a different local molecular environment, consequently it takes various conformations. We analyze these shapes statistically to infer mean and standard deviations (S.D.) in various neighbour distances. These S.D. can be taken as a measure of fluctuations of the tripeptide "ANR" in vivo. Tripeptide data analyzed from the PDB databank show that corresponding mean distances between various C_α and C_β atoms are about the same across all tripeptides making them indistinguishable. The S.D.s in various distances however, are varied across tripeptides implying that each of these fluctuates differently. Indeed about 18% of tripeptides have very little (<0.4 Å), while 4% have large (>0.7 Å) fluctuations [13]. We find that typically, a mutation affects the fluctuating properties of the local pentapeptide, which in turn causes a local conformation change along the peptide chain. We define a measure, which captures fluctuations of pentapeptides. We validate this measure against crystallographic variants, disease causing SNPs, non-disease causing or benign SNPs, and short functional motifs. Our prediction of the effects of disease causing and the benign SNPset was also cross validated with the results of homology modeling based studies [7]. We apply our strategy for the prediction of effects of disease causing mutations of certain key proteins like p53 [14], retinoblastoma protein (RB1) [15] and lipoprotein lipase [16].

2. Materials and methods

In an earlier study, we had analyzed 0.27 million tripeptides from 1220 good resolution protein structures from the PDB [13]. Interatomic distances between the C_α and C_β atoms of each of these tripeptides was computed (Fig. 1) and the results were statistically analyzed. The distances between residues 1 and 2 and residues 2 and 3 of a tripeptide, respectively, were termed as nearest neighbour distances and the distances between residues 1 and 3 of a tripeptide were called as next to nearest neighbour distances. The tripeptide knowledge base consists of statistics of all distances between nearest neighbour and next to nearest neighbour positions of C_α and C_β atoms. The tripeptides in this knowledge base can be considered as sequence structure correlates. This data set typically has about 0.2 Å S.D. in the co-ordinate positions of any atom. It was found that the nearest neighbour C_α and C_β atomic distances have S.D.s ranging from 0 to 0.4 Å suggesting that they are all within allowed (experimental) S.D.s. However, in order to capture the local bending information we need at least one next to nearest neighbour distance data. There are four next to nearest neighbour mean distances which have a S.D. ranging typically between 0.4 and 0.7 Å. Next to nearest neighbour mean distances when compared across all tripeptides show that they are all within allowed S.D.s. A measure of the fluctuating characteristic of a tripeptide can be understood after we present how this data can be used to construct the three dimensional structure of a protein. We begin with a solid structure in three dimensions given by mutual distances between $C_{\alpha1}$, $C_{\beta1}$, $C_{\alpha2}$, $C_{\beta2}$ points of the first two residues. To fix $C_{\alpha3}$ we need three distances from the previous points. Noting that the nearest neighbour distances have the standard deviations in the range 0–0.4 Å and are the least fluctuating, these have to be $C_{\alpha2}$ $C_{\alpha3}$, $C_{\beta2}$ $C_{\alpha3}$ (nearest neighbour distances), the third can be either $C_{\alpha1}$ $C_{\alpha3}$ or $C_{\beta1}$ $C_{\alpha3}$, whichever has lower S.D. Then to determine $C_{\beta3}$ we need to use only $C_{\alpha2}$ $C_{\beta3}$, $C_{\beta2}$ $C_{\beta3}$, $C_{\alpha3}$ $C_{\beta3}$. Alternatively, we may fix $C_{\beta3}$ first and $C_{\alpha3}$ subsequently, in which case the other two nearest neighbour distances become relevant. This procedure can be iterated to fix the co-ordinates down the polypeptide chain. Hence, for any particular tripeptide we need only one next to nearest neighbour distance of the four possible next to nearest neighbour distances αα, αβ, βα, ββ from the data. We pick the one with the lowest S.D. amongst the next to nearest neighbour distances (1,3 distances) and call it the characteristic of the tripeptide.

In mutation analysis we use this characteristic of the tripeptide as a measure of fluctuations. We compare the fluctuating characteristic of two different tripeptides. If the first tripeptide has the lowest S.D. in αβ ($\sigma_1\alpha\beta$) while the second has in αα ($\sigma_2\alpha\alpha$), we look at the ratios $R_1 = (\sigma_1\alpha\beta)/(\sigma_2\alpha\beta)$ and $R_2 = (\sigma_2\alpha\alpha)/(\sigma_1\alpha\alpha)$. If either R_1 or R_2 is outside the range of 0.8–1.2, i.e., 20% deviation, we conclude that the tripeptides fluctuate differently.

In a missense mutation, we compare the corresponding tripeptides of the wild type and mutant at the locus of the affected pentapeptide as described above. After computing the six ratios: three for the wild type tripeptides going into the corresponding mutant tripeptides and vice versa, we define the largest deviant of these from unity as R. If R is within the range 0.8–1.2, we conclude that the fluctuations of the wild type and mutant pentapeptides are similar otherwise they are dissimilar. R is the measure of fluctuations due to the mutation. This is illustrated with an example: a disease causing SNP resulting in R273Q in p53, the wild type pentapeptide at the site of the mutation is EV*R*VC; and the mutant EV*Q*VC (Fig. 2). The three tripeptide pairs to be compared are (EVR, EVQ); (VRV, VQV); (RVC, QVC). The corresponding standard deviations of the next to nearest neighbour residue distances of these tripeptides are obtained from the tripeptide knowledge base [http://www.au-kbc.org/research_areas/bio/projects/protein/tri.html] and the characteristic of each tripeptide is determined. The six ratios are computed. The largest fluctuation R was found to be in the last pair, namely in RVC, the characteristic S.D. is σ αα

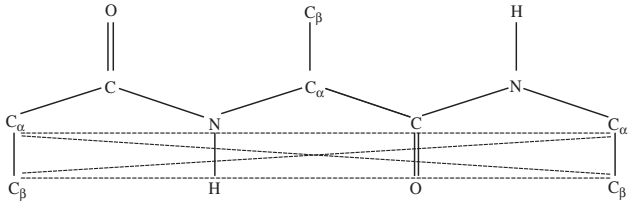

Fig. 1. Tripeptide distances representation of a tripeptide. The dotted lines show the next to nearest neighbour distances (αα, αβ, βα, ββ).

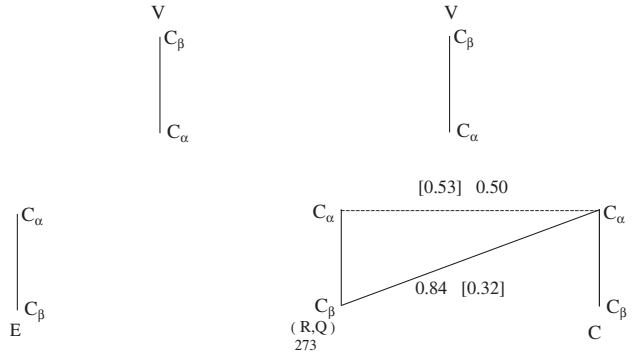

Fig. 2. Missense mutation "R273Q" in p53 protein. The wild type ("EVRVC") and the mutant ("EVQVC") pentapeptides are shown. The first number in each pair corresponds to wild type's S.D. or the characteristic if enclosed in [] and the second to that of the mutant.

[0.53], while σ $\alpha\alpha$ for QVC is 0.5; the characteristic S.D. of QVC is $\sigma\beta\alpha$ [0.32] while σ $\beta\alpha$ of RVC is 0.81. The most deviant ratio for this pair is 0.32/0.81 = 0.39 = *R*.

3. Datasets

Four different datasets were used in this study; crystallographic variants, functional motifs, disease causing SNPs and non-disease causing or benign SNPs. For the crystallographic data involving site-directed mutagenesis experiments a keyword search of Protein Data Bank [17] was used to identify PDB entries, which are single residue variants. The PDB entries were then retrieved from the Protein Data Bank. The corresponding primary citation for the PDB entry was retrieved from PubMed literature database and the experimental analysis of the effect of the mutation recorded. Disease causing and benign missense mutation data was retrieved from Swiss-Prot [18] and dbSNP [19] database. A subset of this data corresponds to the disease and the non-disease set used in homology model assisted mutation studies [7]. Locus specific mutation databases IARC p53 database Release 8.0 [14], RB1 gene mutation database [15] were used to retrieve the missense mutations in the case of p53 and retinoblastoma proteins respectively. Since, a mutation may be recorded more than once in the p53 database, only distinct mutations were extracted as a dataset for further analysis.

In the case of functional motifs a slightly different strategy was adopted. Motifs are short conserved subsequences present within all members of a protein family. These motifs arise because of particular requirements in the structure of specific region(s) of a protein, which may be important for example for their binding properties or enzymatic activity. Since they are all functionally identical, it may be assumed that the local structure adopted by these regions specified by the motifs is essentially the same. The changes in the actual subsequence of the motif in different members of the family can be likened to benign variations leading to similar local structures.

Short functional motifs with lengths ranging from 4 to 10 were retrieved from the Prosite [20] motif database. The exact peptide sequences (as seen across species including *Homo sapiens*) that conform to each of these motifs were also retrieved from the multiple sequence alignment given alongside the Prosite entry. The peptide, which occurs the maximum number of times, is taken as equivalent to wild type and the others as benign variants. Note that, depending on the length of the motif and the number of positions that the peptides differ from the wild type, the number of pairs of tripeptides evaluated and cross ratios computed in the case of each motif differs. 0.6–1.4 is set as the allowed range of fluctuations for those variants, which differ at more than one position from the wild type peptide.

4. Results

The format used for presenting the effect of a mutation is explained with an example entry R58H−/− 0.65. This notation

Table 1
Co-relation with crystallographic variants header line shows the PDB ID along with the protein name

PDB ID: 1kb3 human pancreatic α amylase				
R195A+/− 1.12	R195Q−/− 1.25	N298S−/− 1.68	R337A−/− 0.63	R337Q−/− 0.65
PDB ID: 1dlr/1dlt human dihydrofolate reductase				
L22Y+/+ 1.20	L22F−/+ 0.63	L22W−/+ 2.22	L22R−/− 1.40	
PDB ID: 133l/134l human lysozyme				
R115H+/+ 0.95	R115E−/− 0.45			
PDB ID: 1h4a human γ-D crystallin				
R36S−/− 1.40	R58H−/− 0.65			
PDB ID: 1egd/1ege human medium chain acyl-coA dehydrogenase				
T255E−/− 1.62	E376G−/− 1.40			
PDB ID: 1b5z human lysozyme				
S82A−/− 0.56				
PDB ID: 1f8u human acetylcholinesterase complex				
E202Q+/+ 0.82				
PDB ID: 1n5o breast cancer type 1 susceptibility protein Brct domain 1646–1859				
M1775R−/− 1.41				
PDB ID: 1t2u breast cancer type 1 susceptibility protein Brct domain 1646–1859				
V1809F−/− 0.74				
PDB ID: 1fkc human prion protein (Mutant E200K) fragment 90–231				
E200K−/+ 2.08				
PDB ID: 1hik human interleukin-4				
R88Q−/− 0.48				

Each entry represents the mutation in the crystallographic variant followed by the symbol "+" (no structural change) or "−" (structural change). The first "+" or "−" corresponds to the tripeptide based prediction, and the second "+" or "−" corresponds to the literature report. This is followed by the most deviant ratio *R*.

depicts that Arg at position 58 mutates to His; this is followed by a symbol "+", denoting that it is within the fluctuating range and therefore structurally similar to the wild type or a "−" denoting that it has caused a deformation significantly different from the wild type. Similarly the second "+" or "−" denotes whether a change has occurred or not, according to another study such as homology modeling or crystallographic variant studies. Note that for functional motifs, the second symbol is always a "+" as they are all benign variants. The number that follows the symbols denotes the most deviant tripeptide S.D. ratio R as explained in Section 2. The results obtained in various datasets are discussed below.

5. Crystallographic variants

Many mutant protein crystallographic structures (crystallographic variants) have been analyzed in the literature. We analyzed 21 mutations across 14 crystal structures. Our analysis agrees with 17 out of the 21 giving a prediction accuracy of 81%. The results of these variants are presented in Table 1. We discuss the results of two key proteins BRCA1 and γDCrystallin in detail.

Many forms of genetic cataracts are associated with mutations in the Arginine residues of the γDcrystallin protein. Two forms of genetic cataracts are caused by mutations in γDcrystallin gene: the aculeiform cataract associated with the R58H mutation [21] and the crystal cataract associated with the R36S mutation [22]. Arginine residues play an important role in maintaining the solubility of γ Dcrystallins. Both these mutant proteins have lower solubility and crystallize more readily than the wild type, leading to lens opacity due to the formation of crystal structures. The R58H mutant protein loses the direct ion–pair interaction present in the wild type [23]. In our study, we find that both these mutations cause altered local structural conformation as is apparent from our most deviant ratio being 1.40 for R36S (R36S−/− 1.40) and 0.65 for R58H (R58H−/− 0.65).

Another interesting case study is presented in the mutant studies of BRCA1 protein. BRCA1 is one of the breast cancer susceptibility genes involved in DNA repair and tumor suppression [24]. The carboxy terminal BRCT repeats in BRCA1 protein is essential for its tumor suppressor activity. A well characterized cancer associated mutation M1775R in the BRCT tandem repeat domain of BRCA1 is shown to cause charge–charge repulsion, rearrangement of the hydrophobic core and disruption of hydrogen bonding network at the interface between the two BRCT repeats leading to a conformationally unstable mutant [25]. Our analysis indicates a local structural deformation for this mutant: M1775R−/− 1.41. The BRCT repeats interact with phosphorylated protein targets containing the sequence "pSer-X-X-Phe". The diminished peptide binding capacity observed for cancer associated BRCA1–BRCT variants provide an explanation for increased

Table 2
Functional motifs

Motif Id: PS00014 WT: HDEL ln = 4 ER_target sequence signature			
KDEL+/+ 1.07	H**N**EL−/+ 0.49	**RE**EL−/+ 0.59	**RN**EL+/+ 1.30
RDEL+/+ 1.0	**KE**EL−/+ 0.59	H**E**EL−/+ 0.59	**KN**EL+/+ 1.30
ADEL+/+ 1.28	**KQ**EL+/+ 0.66	**Q**DEL+/+ 1.11	
Motif Id: PS00032 WT: IYPWMR ln = 6 homeobox antennapedia-type protein signature			
LYPWMR+/+ 1.20	**E**YPWM**K**−/+ 1.43	**F**YPWM**A**+/+ 1.19	**M**YPWMR+/+ 1.19
VYPWMR+/+ 1.22	I**F**PWM**K**+/+ 0.97	**V**YPWM**K**+/+ 1.22	**EF**PWM**K**+/+ 0.93
I**F**PWMR+/+ 0.97	**LF**PWMR+/+ 1.0	**MF**PWMR−/+ 0.52	**V**YPWM**T**+/+ 1.22
Motif Id: PS00161 WT: KKCGHM ln = 6 isocitrate lyase signature			
KKCGH**L**+/+ 0.92	KKCGH**Q**+/+ 1.13	K**R**CGH**L**+/+ 0.64	K**R**CGH**R**+/+ 0.64
Motif Id: PS00199 WT: CTHLGCV ln = 7 Rieske iron-sulfur protein signature1			
C**K**HLGC**T**+/+ 1.14	CTHLGC**I**+/+ 0.92	CTHLGC**T**+/+ 0.96	CTHLGC**L**+/+ 1.17
CTHLGC**S**+/+ 1.13			
Motif Id: PS00064 WT: LGEHGDS ln = 7 L-lactate dehydrogenase active ite.			
LGEHG**N**S−/+ 0.50	**M**GEHGDS+/+ 0.94	**I**GEHGD**T**−/+ 0.50	**A**GEHGDS+/+ 0.67
IGEHGDS−/+ 1.29	**V**GEHGDS+/+ 0.98	LGEHGD**T**−/+ 0.50	**M**GEHGD**T**−/+ 0.50
Motif Id: PS00120 WT: VHLLGYSLGA ln = 10 lipases, serine active site			
IWVTGHSLGG−/+ 0.69	**LAISGHSRGG**+/+ 0.71	V**AVMGHSRGG**+/+ 0.72	**LTVTGHS**LGA+/+ 1.29
V**FLIGHSV**GC−/+ 1.53	V**FLIGHSLGC**−/+ 1.53	**VYYVGHSQGT**−/+ 0.56	**IYYVGHSQGC**+/+ 1.34
V**QLIGHS**LGA+/+ 0.72	VH**LIGHS**LGA+/+ 0.75	VH**VIGHS**LGA+/+ 0.75	V**QYVGHSQGT**−/+ 0.54
V**LVSGHS**LGG+/+ 1.34	**VVVSGHS**LGG+/+ 0.76	VH**FLGHS**LGA+/+ 1.30	VHLIGYSLGA+/+ 0.81
LHYVGHSQGT−/+ 0.59	**IHYVGHSQGT**−/+ 0.59	VH**IIGHS**LGS+/+ 0.75	VH**VIGHS**LGS+/+ 0.75
IHVIGHSLGA+/+ 0.75	V**HLIGHS**LGS+/+ 0.75	V**NLVGHSQGG**+/+ 1.34	V**NLIGHSHGG**−/+ 2.44
V**NLIGHSQG**A+/+ 1.34	V**NLIGHSQGG**+/+ 1.34	V**AVTGHS**LGG−/+ 1.54	**VIVTGHS**LGG+/+ 1.29
VHLVG**HSMGG**+/+ 0.70	**IHLVGHSMGG**+/+ 0.70	VH**FIGHSMGG**+/+ 1.30	**VVFTGHS**LGG+/+ 1.29
LVVVGHSLGA+/+ 0.83	**IRLVGHS**LGA+/+ 0.72	**IRLIGHS**LGA+/+ 0.72	V**AVMGHSFGG**+/+ 0.60
IAIIGHSFGG+/+ 0.60	**IAVIGHSFGG**+/+ 0.60	**IAVMGHSFGG**+/+ 0.60	V**NVIGVSWGG**+/+ 1.38
IVLVGHSMGC−/+ 0.45	**IVVTGHS**LGA+/+ 1.29	V**CIVGHSMGG**+/+ 0.70	**IALMGHSFGG**+/+ 0.60
V**NVIGVSWGG**+/+ 1.38			

The header line of each motif has the Prosite Motif ID, the most frequent peptide Wild type, length of the motif, and a short description of the motif. Each entry has the benign variants with the positions at which they differ from the wild type shown in bold. The first "+" (no structural change) or "−" (structural change) corresponds to the tripeptide based prediction. The second symbol is always a "+".

Table 3
Disease causing mutations

Gene: ARSA arylsulfatase A precursor					
G32S+/– 0.91	L68P–/– 0.59	P82L–/– 0.54	R84W–/– 1.26	G86D–/– 0.22	P94A–/– 0.66
S95N–/ 0.55	S96F–/– 0.52	G99D–/– 1.32	G99V–/– 1.27	G122S–/– 0.62	L135P–/– 0.52
P136L–/– 0.7	P148L–/+ 1.65	D152Y–/– 1.6	G154D–/– 2.18	P155R–/+ 1.82	P167R–/+ 1.60
D169N–/– 1.21	C172Y+/– 1.11	I179S–/– 0.48	Q190H–/– 1.47	P191T+/– 1.14	Y201C–/– 2.02
A212V–/– 0.64	A224V–/– 1.44	H227Y–/– 1.53	P231T–/– 0.58	R244H–/+ 1.36	R244C–/– 1.32
G245R–/– 1.9	D255H–/– 1.54	T274M–/– 0.48	S295Y–/– 2.58	L298S–/+ 0.62	C300F–/– 1.41
R311Q+/– 1.17	A314T–/– 0.59	T327I–/+ 0.55	D335V–/– 0.61	N350S–/– 1.35	K367N–/– 1.28
R370Q–/– 0.43	R370W–/– 2.8	P377L–/– 0.61	E382K–/– 1.32	R384C–/– 0.69	R390Q–/– 1.23
R390W–/– 1.51	T391S–/– 0.56	H397Y–/– 0.47	T409I–/+ 0.53	P425T–/– 0.27	P426L–/– 1.81
L428P–/– 1.79	A464V–/– 1.93				
Gene: SOD1 superoxide dismutase [Cu–Zn]					
A4V–/– 0.74	A4T–/– 2.23	V7E–/– 1.43	L8Q–/– 0.35	G12A–/– 1.48	V14M–/– 1.55
G16S–/– 0.52	E21G–/– 0.4	E21K–/– 0.59	G37R–/– 1.93	L38V–/– 1.23	G41D–/– 1.52
G41S–/– 0.65	H43R+/– 0.86	H46R–/– 1.4	H48Q–/– 1.37	G72S–/– 0.36	L84F–/– 1.31
L84V–/– 2.38	G85R–/– 1.24	N86S–/+ 1.35	D90A–/+ 1.22	G93D–/ 0.5	G93S–/ 0.49
G93R–/ 0.65	E100K–/– 1.48	D101G–/– 1.25	D101N–/– 1.25	I104F+/– 1.19	L106V+/– 1.13
G108V+/– 0.9	I112T+/– 0.83	I113T+/– 0.85	D124V–/ 0.5	D125H–/– 0.75	S134N–/– 1.51
N139K–/– 0.66	L144S–/– 0.62	L144F–/– 0.54	A145T–/– 0.64	C146R–/– 1.87	V148G–/– 1.35
V148I+/– 1.12	I149T–/– 1.31	I151T–/– 0.44			
Gene: GUSB β-glucuronidase					
C38G–/– 2.66	S52F–/– 1.33	G136R+/– 1.2	P148S–/+ 1.34	E150K–/– 0.73	D152N+/– 1.2
L176F–/– 0.74	R216W–/– 0.84	Y320S–/– 0.86	Y320C–/– 0.53	H351Y–/– 0.71	A354V–/+ 1.24
R374C–/– 0.43	R382H–/– 1.25	R382C+/– 1.14	P408S–/– 2.45	P415L–/+ 1.69	R435P–/– 1.49
R477P–/– 1.84	Y495C–/– 1.81	Y508C+/– 0.96	G572D–/– 1.34	K606N–/– 1.93	R611W–/– 2.08
A619V–/– 0.68	Y626H+/– 0.9	W627C–/– 2.04			
Gene: ALDH10 aldehyde dehydrogenase					
I45F+/ 1.09	V64D–/ 1.69	L106R–/– 0.67	P114L–/ 0.58	P121L–/ 0.67	T184M–/ 0.51
T184R–/ 0.51	G185A–/ 0.79	C214Y–/– 0.29	C226W+/– 1.16	C237Y–/ 1.9	D245N–/2.22
Y279N–/ 0.67	M328I–/3.17	S365L–/+ 0.47	N386S–/ 1.21	G406R+/ 1.09	H411Y–/ 1.35
S415N–/ 0.45	F419S–/ 1.33	R423H–/2.1	K447E–/ 1.34		
Gene: ACADM acyl-CoA dehydrogenase					
R28C–/– 0.41	Y42H–/ 0.65	I53T–/ 0.61	C91Y–/– 1.29	T96I-/1.63	G112R–/ 0.59
M124I–/+ 1.65	T168A–/– 0.28	G170R–/– 1.37	R181L–/ 1.54	C219R–/– 0.68	S220L–/ 0.74
R256T+/ 1.13	M301T–/– 1.61	K304E–/– 0.66	S311R–/– 0.48	Y327T–/– 0.59	Y327C–/ 1.39
I350T–/+ 0.7					
Gene: FGG fibrinogen					
G268E–/– 1.68	R275H–/– 2.67	R275C–/– 0.52	N308I+/ 1.11	N308K+/– 1.09	G309D–/ 1.82
M310T+/ 0.96	Q329R–/3.96	D330Y–/– 5.16	D330V–/3.76	N337K–/ 1.21	S358C–/– 1.84
D364H–/– 1.51	K380N–/– 0.55	M384V–/ 0.72			
Gene: GNAS1 guanine nucleotide-binding protein G(s), α-subunit					
L99P–/– 1.74	P115S–/– 0.45	P115L–/ 0.54	R165C–/– 1.45	R201S–/ 0.41	R201L–/ 1.81
R201H–/– 0.41	R201C–/– 4.12	Q227H–/ 1.46	Q227R–/ 1.43	R231H–/ 0.7	S250R–/ 1.50
R258W–/ 0.61	E259C+/– 0.96	A366S–/ 0.58	R385H–/+ 0.33		
Gene: ARSB arylsulfatase B					
T92M–/ 1.20	R95Q–/ 1.29	C117R–/– 0.56	G137V–/– 1.51	G144R–/– 2.70	R160Q–/– 0.70
C192R–/– 1.27	Y210C–/ 0.79	L236P–/– 0.72	L321P–/ 0.37	V358M–/ 1.80	H393P–/ 0.49
C405Y–/– 0.56	L498P–/ 1.74	C521Y–/– 1.68			
Gene: F13A1 coagulation factor XIII A chain					
N60K–/– 1.31	M242T–/– 0.33	R252I–/– 1.4	R260H+/– 1.18	R326Q–/– 0.32	A394V–/– 0.62
R408Q–/– 0.59	V414F–/– 0.73	L498P–/– 1.29	N541K–/– 2.36	G562R–/– 1.38	L660P–/– 0.49
L667P–/– 1.45					
Gene: ACADS acyl-CoA dehydrogenase					
R46W–/+ 0.69	G90S+/ 1.06	G92C–/+ 0.28	R107C–/– 1.75	R171W–/+ 1.43	W177R–/+ 0.75
A192V–/ 0.61	G209S–/– 1.69	R352W–/ 1.98	S353L–/ 1.63	R380W–/ 0.76	R383C–/– 0.41
Gene: PPGB protective protein for galactosidases					
Q21R–/– 0.78	S23Y–/– 1.46	W37R+/– 1.1	S62L–/– 1.25	V104M+/– 0.80	L208P–/– 0.61
Y221N–/+ 0.55	Y367C–/– 1.41	M378T–/– 0.58	G411S–/– 0.76	F412V–/– 0.73	
Gene: IVD isovaleryl-CoA dehydrogenase					
L13P–/+ 1.62	R21P–/– 1.57	D40N–/– 1.59	G170V–/– 1.25	A282V–/+ 0.68	C328R–/– 2.86
V342A–/– 0.62	R363C–/– 0.48	R382L–/– 1.49			

(*continued on next page*)

Table 3 (*continued*)

Gene: ALDOB fructose-bisphosphate aldolase B					
C134R−/− 0.54	W147R−/− 0.37	A149P−/− 0.46	A174D−/− 0.69	L256P−/− 1.29	R303W−/− 0.72
N334K−/− 0.37	A337V−/ 1.35				
Gene: NP purine nucleoside phosphorylase					
S51G−/ 1.4	E89K−/ 0.39	D128G+/− 1.15	A174P−/− 0.60	Y192C−/− 0.44	R234P−/ 1.59
Gene: CA2 carbonic anhydrase					
K17E−/ 1.54	Q91P−/− 0.56	H106Y−/− 0.74	P235H−/2.20	N251D−/ 1.24	
Gene: PCBD pterin-4-alpha-carbinolamine dehydratase					
T78I−/− 0.55	C81R−/+ 2.86	R87Q−/− 0.67	E96K−/− 0.69		
Gene: ETFA electron transfer flavoprotein α-subunit					
G116R−/− 0.27	V157G+/− 1.18	T171I−/+ 1.52	T266M−/− 1.43		
Gene: LYZ human lysozyme					
I56T−/− 0.23	D67H−/− 1.86				
Gene: PYGL glycogen phosphorylase					
N339S−/− 1.33	N377K−/− 2.01				
Gene: RBP4 retinol binding protein					
I41N−/− 0.70	G75D−/− 0.73				
Gene: ALDOA fructose-bisphosphate aldolase A					
D128G−/− 1.25	E206K−/− 1.29				
Gene: TSHB thyroid-stimulating hormone β subunit					
C105V−/− 1.48					
Gene: GNAT1 guanine nucleotide-binding protein G(t), α-1 subunit					
G37D−/− 1.5					

The header line shows the gene name along with the protein name. Each entry represents the mutation followed by the symbol "+" (no structural change) or "−" (structural change). The first "+" or "−" corresponds to the tripeptide based prediction, and the second "+" or "−" when present corresponds to the homology model based prediction. This is followed by the most deviant ratio *R*.

cancer risks associated with these mutations [26] The BRCT variant V1809F in this instance has also been predicted to be deleterious by our analysis: V1809F−/− 0.74. It shows that our approach can be used to evaluate the molecular effects of naturally occurring mutants or site directed mutagenesis experiments, albeit at a gross level.

6. Functional motifs

In the functional motif category, 106 peptides across 11 functional motifs were analyzed. Results show that prediction accuracy in this category is around 75%. A representative dataset is presented in Table 2. Functional motifs are short stretches of subsequences, which are important for the functionality of a protein. These motifs are more like regular expressions rather than exact peptides when considered across species. The peptide that occurs with greatest frequency in a particular motif is taken as the wild type peptide. The other peptide sequences that a particular motif spans across species can be considered as benign variants of this wild type peptide. Benign variants in this dataset may differ from the wild type peptide motif at more than one position simultaneously. These are comparable to double mutants, triple mutants and so on. In such cases, the allowed range of fluctuations is set to 0.6–1.4. We present a motif "PS00120" as a case study.

Triacylglycerol lipases [27] are lipolytic enzymes that hydrolyze the ester bond of triglycerides. "PS00120" is a Prosite motif designed around the active site residue Serine of triacylglycerol lipases. There are 45 occurrences of this motif across species including *H. sapiens*. Most of the peptides differ at more than three positions from the wild type. 36 out of 45 peptides fall within the allowed fluctuating range. Moreover, a good majority 23 out of 36 are within 0.7–1.3 range. This goes to illustrate the fact that a well designed motif around a conserved site of a protein should yield good results with our approach. However, one should note that as the regular expression describing the motif gets more lengthy and generic, the ambiguity of residues at each position increases and hence the reliability decreases. We therefore suggest that the usage of our approach to evaluation of motifs be restricted to motifs, which are less generic and of length below 10.

7. Disease causing mutations

A total of 1478 disease causing mutations from 26 proteins are analyzed. This includes 1147 mutants from p53, 28 from lipoprotein lipase and 20 from retinoblastoma protein. 1300 mutants in this set have been predicted to cause local structural changes giving a prediction accuracy of 88%. The fraction of mutants not showing structural deformation, i.e., the ratio *R* in the range 0.8–1.2, is 12%. Since, the bulk of the mutations

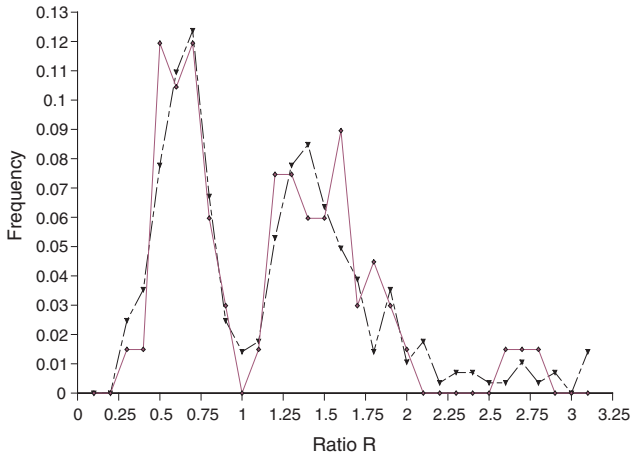

Fig. 3. Ratio R vs. normalized frequency of disease (dash and dot line) and non-disease sample set (solid line).

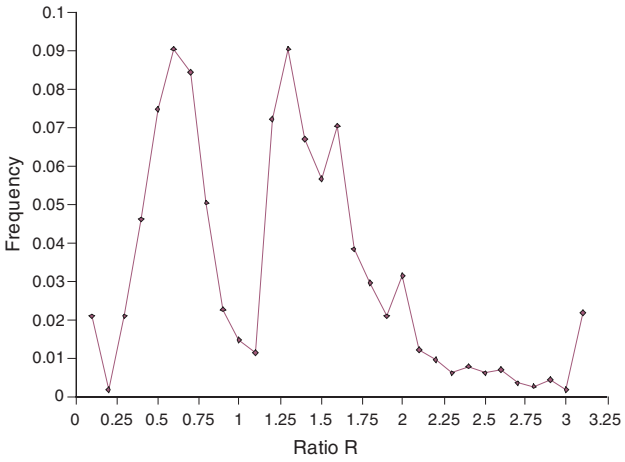

Fig. 4. Ratio R vs. normalized frequency of p53 disease sample set.

are from p53, the p53 dataset was also analyzed separately. The prediction accuracy remains the same. Further, 228 mutants from 23 proteins in this disease dataset have results from homology modeled studies [7]. The agreement with the homology studies is 78%. The results for these mutants are presented in Table 3. It should be noted that the false negatives in this set

Table 4
Tripeptide based prediction: disease causing mutations

p53 protein					
D7H− 1.52	L35F− 1.35	L43S− 0.60	W53C− 0.25	P60S− 2.45	P87Q− 1.81
S94T− 1.53	R110C+ 1.20	R110L− 1.22	R110P− 1.71	F113C− 0.43	T125M− 2.06
Y126D− 0.09	Y126N− 6.0	S127F− 1.48	P128S− 1.52	A129D− 0.53	L130R− 0.58
N131S− 1.24	N131K− 1.53	K132M− 0.66	K132Q− 0.11	M133T− 1.27	C135S− 0.29
C135F− 0.51	Q136E− 2.91	Q136K− 0.32	L137Q− 1.44	A138P− 0.62	K139N− 1.43
C141G− 1.76	C141F− 1.43	C141Y− 1.50	V143A− 1.93	Q144P− 1.64	L145P− 1.52
L145Q− 0.64	V147D− 1.28	V147G− 0.53	S149P− 1.76	P151A− 1.59	P151S− 1.79
P151T− 1.82	P152L− 0.44	P152S− 1.96	P153T− 2.0	T155A+ 1.16	R156P− 1.61
V157D− 0.54	V157I− 1.28	R158C− 0.38	R158H− 0.50	M160I− 1.60	A161S− 1.31
I162S− 0.45	I162V− 1.21	K164N− 0.66	K164Q− 0.50	Q165L+ 0.80	Q165R− 1.29
S166L− 2.36	H168R− 0.24	M169I− 1.48	M169T− 0.33	T170M− 1.74	T170S− 1.48
V172A− 0.67	V173E− 1.27	V173L− 1.26	V173M− 1.87	R175C− 1.70	R175G− 4.23
R175L− 1.52	R175P− 2.52	R175H− 0.21	C176F− 2.76	C176W− 0.28	P177L− 0.53
R181L− 0.34	C182S− 0.61	D184Y− 0.61	D186Y− 1.58	G187S− 1.46	A189P− 1.75
P190L− 1.65	P191T− 2.26	Q192R− 0.68	H193D− 1.30	H193R− 0.70	L194P− 1.54
L194R− 0.25	I195T− 2.82	Y205C− 1.51	R213Q− 0.20	Y220C− 0.51	Y220H− 0.54
D228E− 0.47	T230I− 1.78	I232T− 2.0	Y234H− 2.56	M237I− 0.21	C238F− 3.35
C238Y− 3.18	S240I− 2.33	S241F− 0.17	C242F− 0.60	G245A+ 0.94	G245C− 1.51
G245D+ 1.08	M246R− 0.47	M246T− 1.42	M246V− 0.57	N247I− 0.50	R248G− 1.62
R248L− 1.32	R248Q− 1.57	R248W− 0.41	R249S− 2.0	L252P− 2.03	I254N− 0.55
I254T− 0.54	E258D− 0.75	E258K+ 0.85	V272L+ 1.15	R273C− 0.12	R273H− 0.27
R273Q− 0.39	V274F− 1.55	C275Y− 2.41	C275W− 0.58	C277G− 5.13	P278A− 3.04
P278H− 0.44	P278L− 0.32	P278S− 2.84	G279E− 1.21	R280K− 0.71	R280I− 1.29
R280T+ 1.15	D281A− 1.46	D281E− 0.72	D281G+ 1.17	D281V− 0.61	R282L− 1.40
R282W− 3.31	R283C− 1.53	R283G− 0.49	R283H− 0.72	R283P− 1.71	T284A− 1.46
T284P− 1.87	E285K− 0.63	E285Q− 0.75	E285V+ 1.20	E286A+ 0.80	E286D− 1.49
E286G− 0.63	E286K− 0.71	K292I− 1.32	P300R− 1.46	P301L− 1.93	R306Q− 1.25
A307T− 1.42	P309S− 0.67	R337C− 2.38			
Lipoprotein lipase					
W113G+ 1.16	W113R− 0.29	H163R− 0.59	G169E− 1.43	G181S+ 0.87	D183G− 1.89
D183N− 1.51	P184R− 2.0	A185T− 0.51	A203T− 1.39	D207E− 0.53	H210Q− 1.40
G215E− 0.74	I221T− 1.31	D231E− 0.65	I232S− 0.94	C243S− 1.43	R270H− 0.39
S271T− 0.38	D277N− 0.6	S278C− 1.90	S286G− 0.77	S286R+ 1.15	M328T− 0.48
L330P− 0.61	A361T− 1.57	E437K− 1.22	E437V− 1.19		
Retinoblastoma protein RB1					
E72Q− 1.79	E137D+ 0.86	I185T− 0.61	R358Q− 1.40	K447Q− 1.25	M457R− 1.48
R500G+ 1.08	K530R− 0.49	H549Y− 0.56	S567L− 0.51	K616E− 0.74	A635P− 1.40
V654E− 1.54	R661W− 2.27	L662P− 1.35	H673P− 0.36	Q685P− 0.50	C712R− 1.30
N803K− 0.72					

The mutation, along with the first "+"(no structural change) or "−"(structural change) and the most deviant ratio R is shown for each entry.

Table 5
Non-disease causing mutations

Gene: AVP vasopressin-neurophysin 2-copeptin					
P82L−/+ 0.59	G119V−/ 0.63				
Gene: GH1 growth hormone					
V136I−/+ 0.78					
Gene: ELAM E-selectin					
C130W−/− 1.69	S149R−/ 1.80	E295K−/ 1.22	E421Q−/ 1.77	H468Y−/ 0.75	L575F+/ 1.08
Gene: ALDR aldehyde reductase					
I14F−/− 1.9	H41L−/+ 1.21	L72V−/ 0.52	G203S−/ 1.36	T287I−/ 1.41	
Gene: ICAM1 intercellular adhesion molecule-1					
K155N−/ 0.56	G241R−/ 1.4	V315M−/ 1.54	P352L−/ 0.58	R397Q−/ 1.41	E469K−/ 0.62
Gene: COX1/PTGS1 cyclooxygenase 1					
R8W−/ 0.70	P17L−/ 0.70	R53H−/ 1.35	R149L−/ 0.61	L237M−/ 0.59	K359R−/2.6
I443V−/ 0.60					
Gene: COX2/PTGS2 cycloxygenase 2					
R228H−/ 0.37	P428A−/ 0.48	E488G−/− 0.50	V511A+/+ 0.80	G587R−/ 1.24	
Gene: KLK1 kallikrein					
R77H−/ 1.42	Q145E−/+ 1.43	E186K−/ 1.53	V193E+/ 1.19		
GENE: LEP leptin					
I45V−/+ 0.71	V94M−/2.65	V110M−/ 0.48			
Gene: PAI2 plasminogen activator inhibitor-2					
N120D+/ 0.83	R229H−/ 1.6	N404K−/+ 1.58	S413C−/− 1.96		
Gene: ANX3 annexin A3					
S19N+/+ 1.19	I219N−/− 0.66	P251L−/− 1.72	F291S−/ 0.49		
Gene: APOD apolipoprotein D					
F15S+/ 1.18	S115L−/2.78	T178K−/+ 0.50			
Gene: F3 thromboplastin					
T36A−/ 1.34	I145V−/ 0.5	R163W−/ 0.70			
Gene: CYH chymase					
G46R+/+ 1.17	H66R−/− 1.61				
Gene: CYP11A cytochrome P450 11A1					
E314K+/+ 0.86					
Gene: GNB3 guanine nucleotide-binding protein β-subunit 3					
D76N−/ 1.57	G272S−/+ 0.60				
Gene: ICAM2 intercellular adhesion molecule-2					
A37T−/− 0.43	R199H−/ 0.22				
Gene: PLA2 phospholipase A2					
D16A−/ 0.43	N89T−/+ 1.60	N89K−/ 1.88			
Gene: CNTF ciliary neurotrophic factor					
H182R−/+ 1.30					
Gene: GH2 growth hormone					
R90W−/− 1.22					
Gene: HMGCR 3-hydroxy-3-methylglutaryl-coenzyme A reductase					
I638V+/+ 1.19					

The header line shows the Gene name along with the protein name. Each entry represents the mutation followed by the symbol "+"(no structural change) or "−" (structural change). The first "+" or "−" corresponds to the tripeptide based prediction, and the second "+" or "−" when present corresponds to the homology model based prediction. This is followed by the most deviant ratio *R*.

do not have an overlap with the false negatives of the homology modeling studies. The statistical profile of these mutants is presented in Fig. 3.

We discuss three proteins p53, RB1 and lipoprotein lipase in detail. p53 and RB1 are tumor suppressor proteins. Mutations in p53 have been linked to many forms of cancer. Mutations in

both the alleles of RB1 gene lead to the development of retinoblastoma, a childhood tumor of the eye [15]. Out of the 1147 disease causing mutations from p53 protein, 1002 mutants were predicted to cause structural deformations, whereas 145 were predicted not to alter the local structure. An overview of the fluctuating nature of all the p53 mutations is shown in Fig. 4. Out of the 20 pathological mutations analyzed in RB1, 19 of them are predicted to change the local structure.

Lipoprotein lipase is a key enzyme in lipid metabolism. Many diseases including atherosclerosis, coronary heart disease and chylomicronemia appear to be directly or indirectly associated with abnormalities in lipoprotein lipase function [16]. Out of the 28 mutations analyzed in lipoprotein lipase, we confirm structural deformations in 25 instances. Only the mutations presented in the corresponding SwissProt entries of these three proteins are shown in Table 4.

8. Non-disease mutations

67 missense mutations from the non-disease category have been evaluated and results presented in Table 5. Bulk of the mutations in this set, fall outside of the allowed fluctuating range as shown in Fig. 3. 13% of the mutants have the ratio R in the 0.8–1.2 range, while 22% of the mutants fall within 0.75–1.25 fluctuation range. 35 of these have homology modeled studies. The agreement of our method with the homology models is 43%. Our methodology shows that in this dataset, 87% of the mutants do cause local structural deformations, yet in the system do not manifest in any deleterious effects due to reasons detailed in Section 9. It is further seen that all the profiles in Figs. 3 and 4 are alike suggesting that structural variations due to mutations are always the same.

9. Discussion

Tripeptide analysis shows single point mutations invariably cause protein instability. The precise cause such as loss of H-bonds or salt bridges, backbone strain, change in catalytic activity or ligand binding cannot be ascertained by our method. But in each instance, with further inputs into the analysis, some of them may be inferred. We concentrated on a methodology to assess whether there is a structural change or not. The prediction accuracy in various categories is summarized in Table 6. The main advantage of our methodology being, the analysis does not require prior structural templates as in homology modeling studies and therefore has more utility in the post genomic era. The method can be utilized in designing structural analogues for de novo proteins as well.

We have studied disease causing mutations and non-disease or benign mutations using our methodology. The disease sample set confirms structural deformations and a statistical profile of these changes across many proteins is shown in Fig. 3. We find that in the non-disease dataset also there are structural changes, 43% of the time in agreement with other modeling studies, yet there are no deleterious effects. Furthermore, the statistical profile of these changes is about the same as in the disease set as shown in Fig. 3. The structural changes are perhaps irrelevant in these cases, because the locus of mutation is at a functionally redundant domain or there are overlapping protein functions [5], alternative pathways or the protein itself is not crucial to the organism. The issue of disease vs. non-disease mutation is a multi-parameter decision. Our methodology can only ascertain the structural aspects, there should be other corroborative knowledge on the particular domain of the protein, to conclude about the deleterious effects of the mutation.

Table 6
Summary of positive confirmations of ratio R against various categories

Category	Sample size	Positive confirmation in %
Crystallographic variants	21	81
Functional motifs	106	75
Disease	1478	88
Non-disease	67	13
Homology models [disease]	228	78
Homology models [non-disease]	35	43

10. Conclusions

Local pentapeptide structure is important to understand the conformation and fluctuations of a protein, which in turn dictates the protein stability. This aspect can be successfully understood in terms of tripeptides. Disease and non-disease causing mutations are indistinguishable from the structural deformation statistics alone. Functional motifs are normally categorized based on sequence homology alone, we suggest that structural similarity can also be imposed for further refinement.

Acknowledgements: G.P. is thankful to DBT India, for support through Department of Biotechnology Grant No. BT/PR/1737/BID/18/015/099. The authors thank BTIS DBT Programme and ILGTI Project of Institute of Mathematical Sciences for computational facilities.

References

[1] Sachidanandam, R., Weissman, D., Schmidt, S.C., Kakol, J.M., Stein, L.D., Marth, G., Sherry, S., Mullikin, J.C., Mortimore, B.J., Willey, D.L., Hunt, S.E., Cole, C.G. and Coggill, P.C., et al.International SNP Map Working Group (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409 (6822), 928–933.
[2] Collins, F.S., Brooks, L.D. and Chakravarthi, A. (1998) A DNA polymorphism discovery resource for research on human genetic variation. Genome Research 8, 1229–1231.
[3] Brookes, A.J. (1999) The essence of SNPs. Gene 234, 177–186.
[4] Collins, F.S., Guyer, M.S. and Chakravarthi, A. (1997) Variations on a theme: cataloging human DNA sequence variation. Science 278, 1580–1581.
[5] Labow, M.A., Norton, C.R., Rumberger, J.M., Lombard Gilloly, K.M., Shuster, D.J., Bertko, R., Knaack, P.A., Terry, R.W. and Harbison, M.L. (1994) Characterization of E-selectin deficient mice: demonstration of overlapping function of the endothelial selectins. Immunity 1, 709–720.
[6] Sunyaev, S., Ramensky, V. and Bork, P. (2000) Towards a structural basis of human non-synonymous single nucleotide polymorphisms. Trends in Genetics 16 (5), 198–200.
[7] Wang, Zhen and Moult, John (2001) SNPs, protein structure and disease. Human Mutation 17, 263–270.
[8] Chasman, Daniel and Mark, Adams R. (2001) Predicting the functional consequences of non-synonymous single nucleotide

polymorphisms: structure based assessment of amino acid variation. Journal of Molecular Biology 307, 683–706.
[9] Ng, C. Pauline and Henikoff, S. (2002) Accounting for human polymorphisms predicted to affect protein function. Genome Research 12, 436–446.
[10] Ng, C. Pauline and Henikoff, S. (2003) SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Research 31, 3812–3814.
[11] Saunders, Christopher T. and Baker, D. (2002) Evaluation of structural and evolutionary contributions to deleterious mutation prediction. Journal of Molecular Biology 222, 891–901.
[12] Sunyaev, S., Lathe, W. and Bork, P. (2001) Integration of genome data and protein structures: prediction of protein folds, protein interactions and molecular phenotypes of single nucleotide polymorphisms. Current Opinion in Structural Biology 11, 125–130.
[13] Anishetty, S., Pennathur, G. and Anishetty, R. (2002) Tripeptide analysis of protein structures. BMC Structural Biology 2, 9.
[14] Olivier, M., Eeles, R., Hollstein, M., Khan, M.A., Harris, C.C. and Hainaut, P. (2002) The IARC TP53 Database: new online mutation analysis and recommendations to users. Human Mutation 19 (6), 607–614.
[15] Lohmann, D.R. (1999) RB1 mutations in retinoblastoma. Human Mutation 14 (4), 283–288.
[16] Murthy, V., Julien, P. and Gagne, C. (1996) Molecular pathobiology of the human lipoprotein lipase gene. Pharmacology Therapy 70 (2), 101–135.
[17] Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000) The protein data bank. Nucleic Acids Research 28, 235–242.
[18] Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.C., Estreicher, A., Gasteiger, E., Martin, M.J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S. and Schneider, M. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Research 31, 365–370.
[19] Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M. and Sirotkin, K. (2001) dbSNP: the NCBI database of genetic variation. Nucleic Acids Research 29 (1), 308–311, Jan 1.
[20] Hulo, N., Sigrist, C.J.A., Le Saux, V., Langendijk-Genevaux, P.S., Bordoli, L., Gattiker, A., De Castro, E., Bucher, P. and Bairoch, A. (2004) Recent improvements to the PROSITE database. Nucleic Acids Research 32, D134–D137.
[21] Heon, E., Priston, M., Schorderet, D.F., Billingsley, G.D., Girard, P.O., Lubsen, N. and Munier, F.L. (1999) The γcrystallins and human cataracts: a puzzle made clearer. American Journal of Human Genetics 65, 1261–1267.
[22] Kmoch, S., Brynda, J., Asfaw, B., Bezouska, K., Novak, P. and Rezacova, P., et al. (2000) Link between a novel human gD-crystallin allele and a unique cataract phenotype explained by protein crystallography. Human Molecular Genetics 9, 1779–1786.
[23] Basak, A., Bateman, O., Slingsby, C., Pande, A., Asherie, N., Ogun, O., Benedek, G. and Pande, J. (2003) High-resolution X-ray crystal structures of human γD crystallin (1.25 Å) and the R58H mutant (1.15 Å) associated with aculeiform cataract. Journal of Molecular Biology 328, 1137–1147.
[24] Kerr, Peter and Ashworth, Alan (2001) New complexities for BRCA1 and BRCA2. Current Biology 11 (16), R668–R676.
[25] Williams, R.S. and Glover, J.N.M. (2003) Structural consequences of a cancer-causing Brca1–Brct missense mutation. Journal of Biological Chemistry 278 (4), 2630–2635.
[26] Williams, R.S., Lee, M.S., Duong, D.D. and Glover, J.N.M. (2004) Structural basis of phosphopeptide recognition by the Brct domain of Brca1. Natural Structural of Molecular Biology 11 (6), 519–525.
[27] Villeneuve, P., Muderhwa, J.M., Graille, J. and Haas, M.J. (2000) Customizing lipases for biocatalysis: a survey of chemical, physical and molecular biological approaches. Journal of Molecular Catalysis B: Enzymatic 9, 113–148.