Genomics 99 (2012) 330–339

Contents lists available at SciVerse ScienceDirect

Genomics

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

Functional genomics based prioritization of potential nsSNPs in EPHX1, GSTT1, GSTM1 and GSTP1 genes for breast cancer susceptibility studies

Tariq Ahmad Masoodi [a,b], Venkateswar Rao Talluri [a], Noor Ahmad Shaik [b,c,*], Jumana Yousuf Al-Aama [c], Qurratulain Hasan [b]

[a] Department of Biotechnology, K L University, Andhra Pradesh, India
[b] Department of Genetics and Molecular Medicine, Vasavi Medical and Research Centre, Khairatabad, Hyderabad-500004, India
[c] Princess Al-Jawhara Center of Excellence in Research of Hereditary Disorders, Department of Genetic Medicine, Faculty of Medicine, King Abdulaziz University, Jeddah-21589, Saudi Arabia

ARTICLE INFO

Article history:
Received 3 February 2012
Accepted 23 April 2012
Available online 1 May 2012

Keywords:
Computational analysis
Single nucleotide polymorphism
Predisposition genes
Breast cancer

ABSTRACT

In the present study, nsSNPs in EPHX1, GSTT1, GSTM1 and GSTP1 genes were screened for their functional impact on concerned proteins and their plausible role in breast cancer susceptibility. Initially, SNPs were retrieved from dbSNP, followed by identification of potentially deleterious nsSNPs using PolyPhen and SIFT. Functional analysis was done with SNPs3D, SNPs&GO and MutPred methods. Prediction and evaluation of the functional impact on the 3D structure of proteins were performed with Swiss PDB viewer and NOMAD-Ref servers. On analysis, 13 nsSNPs were found to be highly deleterious and damaging to the protein structure, of which 6 nsSNPs, rs45549733, rs45506591 and rs4986949 of GSTP1, rs72549341 and rs148240980 of EPHX1 and rs17856199 of GSTT1 were predicted to be potentially polymorphic. It is therefore hypothesized that the 6 identified nsSNPs may alter the detoxification process and elevate carcinogenic metabolite accumulation thus modifies the risk of breast cancer susceptibility in a group of women.

1. Introduction

Breast cancer is the leading malignancy responsible for the increased number of worldwide deaths among pre and postmenopausal women. Most breast cancer cases occur in a small percentage of the population that is at increased risk due to genetic susceptibility factors [1]. High-penetrance genes account for only 5% of the cases, whereas polymorphic low-penetrance genes acting in concert with lifestyle/environmental risk factors are likely to account for a much higher proportion [1]. Among the various genetic factors, polymorphic epoxide hydrolase gene (EPHX1) and the glutathione S-transferase genes (GSTT1, GSTM1, and GSTP1) are known for their involvement in breast cancer predisposition [2]. EPHX1 gene encodes the epoxide hydrolase protein mEH, which catalyzes the hydrolysis of arene and aliphatic epoxides to trans-dihydrodiols and typically results in detoxification and preparation for phase II conjugation reactions. GSTT1, GSTM1, and GSTP1 genes encode enzymes that are among the superfamily of GST enzymes involved in the detoxification of reactive metabolites of carcinogens such as polycyclic aromatic hydrocarbons. Polymorphisms in EPHX1, GSTT1, GSTM1, and GSTP1 genes have been examined as risk factors for breast cancer in a number of studies [3]. Two meta-analysis studies observed a significant association of breast cancer with the GSTM1 and GSTT1 polymorphisms. In another study, an increase in the risk of breast cancer with a growing number of G alleles of the GSTP1 has been found [4,5]. Cases carrying reduced EPHX1 activity were found to be at a higher risk of breast cancer in comparison with carriers of high EPHX1 activity [6].

Single nucleotide polymorphisms (SNPs) are the most commonly known genetic variations used in mapping the complex human genetic traits. Approximately 544,000 SNPs are believed to be localized in the coding region of the human genome, among them the nsSNPs are the most interesting ones owing to their direct effect on protein structure and function. These are also considered to be important factors contributing to functional diversity of the encoded proteins in human populations [7]. A number of SNP databases are available. Important among them are the human genome variation database, HGVBase [8] and the National Center for Biotechnology Information database, dbSNP [9]. Since the effect of non-coding SNPs on gene regulation is difficult to understand, attention is being focused toward non-synonymous coding SNPs. These types of mutations are believed to cause a change more likely in the structure and as such alter the function of a protein molecule. These nsSNPs affect gene expression by modifying DNA and transcription factor binding [6], inactivate active sites of enzymes or change splice sites, thereby producing defective gene products [10].

* Corresponding author at: Department of Genetics and Molecular Medicine, Vasavi Medical and Research Centre, Khairatabad, Hyderabad-500004, India. Fax: +91 40 24022277.
E-mail address: noorahmadh@gmail.com (N.A. Shaik).

0888-7543/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2012.04.006

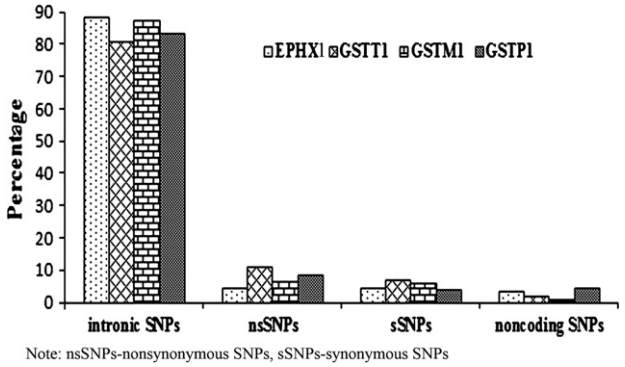

Fig. 1. Distribution of SNPs in predisposition genes. Note: nsSNPs—nonsynonymous SNPs, sSNPs—synonymous SNPs.

In the past decade, many of the epidemiologic studies have focused on identifying the nsSNPs in EPHX1, GSTT1, GSTM1, and GSTP1 genes examining their association with breast cancer risk [11]. Although the specific association of EPHX1 and GST gene polymorphisms with breast cancer risk is not clear, it is has been postulated that a decrease in EPHX1 and GST enzyme activities could result in a higher frequency and a specific pattern of mutations leading to cancer predisposition. Many studies focused on studying SNPs in coding regions, in the hope of finding a significant association between SNPs and breast cancer susceptibility, but often they find little or no association [12]. With the availability of high-throughput SNP detection techniques, the population of nsSNPs is increasing rapidly, providing a platform for studying the relationship between genotypes and disease phenotype. Our ability to better select a nsSNP for an association study can be enhanced by first examining the possible potential impact of an amino acid variant on the encoded protein functions using different genomics programs such as Sort Intolerant from Tolerant (SIFT) Polymorphism Phenotype (PolyPhen), Single Nucleotide Polymorphism Database (SNPs) & Gene Ontology (GO) and MutPred etc. Discovering the deleterious nsSNPs out of a reported huge pool of SNPs could prove to be very useful in epidemiological studies. Hence, the current study was undertaken to identify the potential deleterious nsSNPs in EPHX1, GSTT1, GSTM1, and GSTP1genes, to predict their plausible role in breast cancer susceptibility.

2. Results

2.1. SNP distribution

The EPHX1 gene investigated in this work was reported to have a total of 639 SNPs, of which 27 were nsSNPs2, 29 were synonymous SNPs, and 20 were in the non-coding regions, which comprise 8 SNPs in the 5′ UTR and 12 SNPs in the 3′ UTR while the rest were in the intronic regions. GSTT1 gene is reported to have a total of 103 SNPs, of which 11 were nsSNPs, 7 were synonymous SNPs, 2 were in noncoding regions both of which are present in the 3′ UTR and the rest were in the intronic regions. GSTM1 was identified to have 154 SNPs, of which 10 were nsSNPs, 9 were synonymous SNPs, 1 in a non-coding region which is present in the 5′ UTR region and the rest were in the intronic regions. GSTP1 gene had a total of 155 SNPs, of which 13 were nsSNPs, 6 were synonymous SNPs, which comprise 7 SNP in the 5′ UTR and 4 SNPs in the 3′ UTR and the rest were in the intronic regions. Thus a very small percentage of the SNPs were present either in coding or non-coding regions whereas the majority was present in the intronic region (Fig. 1). We selected only non-synonymous coding SNPs for our investigation, which totaled 61 nsSNPs in the four genes evaluated (Table 1).

2.2. Prediction of deleterious nsSNPs by SIFT

All nsSNPs retrieved from these genes, were submitted independently to the SIFT server. The result showed a total of 19 nsSNPs to be deleterious with the score of ≤ 0.05. Seven each in EPHX1 and GSTP1 genes, 4 in GSTT1 and one in the GSTM1 gene were predicted to be deleterious (Table 2).

2.3. Prediction of functional modification of coding nsSNPs

To identify the nsSNPs that affect protein structures, they were analyzed for predicting the possible impact of altered amino acids on the structure and function of the proteins using PolyPhen server. The fasta protein sequence of each gene with nsSNP position and their 2 amino acid variants were submitted as input for analyzing the structural change of proteins. Our result showed a total of 24 nsSNPs that were damaging with the PSIC score of ≥ 1.5. Out of the 24 nsSNPs predicted, 9 were in EPHX1, 6 in GSTP1, 7 in GSTT1 and 2 in GSTM1 genes (Table 2). Fifteen nsSNPs which were observed to be deleterious by SIFT server were also predicted to be damaging by PolyPhen server.

2.4. SNPs3D analysis

We have identified a total of 25 deleterious nsSNPs by SNPs3D analysis, detailed results of which are shown in Table 2. Out of the said nsSNPs, 10 were predicted for EPHX1, 6 for GSTT1, 2 for GSTM1and 7 for GSTP1 genes. Thirteen nsSNPs which were found to be deleterious and damaging by SIFT and PolyPhen servers were also predicted to be deleterious by SNPs3D analysis.

2.5. SNPs&GO and MutPred analyses

A total of eight nsSNPs were found to be deleterious by SNPs&GO (Table 2) which correspond to 1 nsSNP in EPHX1, 2 in GSTT1 and 5 in GSTP1. From these eight, 7 nsSNPs were also predicted to be polymorphic by SIFT, PolyPhen and SNPs3D (Tables 3 and 4). MutPred

Table 1
Prediction results of nsSNPs of breast cancer genes under study.

Prediction result	SIFT algorithm[a]		PolyPhen algorithm[b]		SNPs3D algorithm[c]		SNPs&GO algorithm[d]		MutPred algorithm[e]	
	No. of nsSNPs	%	No. of nsSNPs	%	No. of nsSNPs	%	No. of nsSNPs	%	No. of nsSNPs	%
Deleterious	19	31.14	24	39.34	25	40.98	8	13.12	28	45.90
Tolerated	42	68.85	37	60.65	36	59.02	53	86.88	33	54.10
Total	61	100	61	100	61	100	61	100	61	100

[a] See web site: SIFT (http://blocks.fhcrc.org/sift/SIFT.html). Positions with normalized probabilities <0.05 are predicted to be damaging, and those ≥0.20 are predicted to be tolerated.
[b] See web site: PolyPhen (http://genetics.bwh.harvard.edu/pph/). Positions with normalized probabilities <1.5 are predicted to be tolerated, and those ≥1.5 are predicted to be damaging.
[c] See web site: SNPs3D (http://www.snps3d.org/). Positions with normalized probabilities <0 are predicted to be deleterious, and those >0 are predicted to be tolerated.
[d] See web site: SNPs&GO (http://snps-and-go.biocomp.unibo.it/snps-and-go//). Result is interpreted on the basis of the reliability index score.
[e] See web site: MutPred (http://mutpred.mutdb.org//). Scores with g>0.5 and p<0.05 are predicted to be deleterious and those g<0.5 and p>0.05 are tolerated.

Table 2
Distribution of deleterious nsSNPs in predisposition genes predicted by SIFT, PolyPhen, SNPs3D, SNPs&GO and MutPred algorithms.

Gene	No. of deleterious nsSNPs predicted by SIFT	No. of deleterious nsSNPs predicted by PolyPhen	No. of deleterious nsSNPs predicted by SNPs3D	No. of deleterious nsSNPs predicted by SNPs&GO	No. of deleterious nsSNPs predicted by MutPred	No. of deleterious nsSNPs predicted by either SIFT, PolyPhen, SNPs3D, SNPs&GO or MutPred
EPHX1	7	9	10	1	8	4
GSTT1	4	6	6	2	6	2
GSTM1	1	2	2	0	5	1
GSTP1	7	7	7	5	9	6
Total	19	24	25	08	28	13

predicted 28 nsSNPs to be polymorphic (Table 2) corresponding to 8 in EPHX1, 6 in GSTT1, 5 in GSTM1 and 9 in GSTP1 (Table 2). Thirteen nsSNPs were predicted to be polymorphic by all the programs except SNPs&GO which predicted only seven polymorphic nsSNPs. These correspond to 4 nsSNPs of EPHX1, 2 of GSTT1, 1 of GSTM1 and 6 of GSTP1 (Tables 3 and 4). Since these thirteen nsSNPs were predicted to be deleterious and damaging with higher scores by all the methods, these nsSNPs were found to be more harmful to the normal function of the proteins concerned.

2.6. Structural modeling of mutant protein

Single amino acid mutations can significantly change the stability of a protein structure. So, the knowledge of a protein's 3D structure is essential for a full understanding of its functionality. Information regarding how to map the deleterious nsSNPs into protein structure was taken from the Protein Data Bank. The PDB IDs of the available structures are given in Table 5. The mutations for the given structures were performed by Swiss-PDB viewer to get modeled structures. Then, energy minimizations were performed by NOMAD-Ref server [13] for the native structures (including homology modeled structure) and their mutant models. It can be seen from Table 5 that the total energy for the native structures and the mutant models shows a great variation with high RMSD values. The native and mutant models were superimposed by the Swiss PDB viewer to get their RMSD values (Figs. 2–4). It can be seen that the total energy value of the GSTP1 native structure (−10448.928 kcal/mol) and its mutant modeled structures S150W, E31V and D147Y were found to be −9713.515 kcal/mol, −10127.256 and −9892.780 kcal/mol respectively while the remaining three models showed almost no difference or less negative energies. The RMSD values between its native and the S150W, E31V and D147Y models were 2.67 Å, 2.08 Å and 2.45 Å while for the other three models the values were lower. The total energy value of the GSTT1 native structure (−12983.094 kcal/mol) and its mutant modeled structure F45C was found to be −11260.579 kcal/mol while the other mutant model (E173K) has an energy of −12819.685 which is almost similar to the native structure. The RMSD values between its native and the F45C and E173K models were 2.33 Å and 0.99 Å respectively. The total energy values of the EPHX1 native structure (−7891.090 kcal/mol) and its mutant modeled structures Y374S, R71H, R49C and R71C were found to be −6289.072 kcal/mol, −7832.321 kcal/mol, −7240.202 kcal/mol and −6542.200 kcal/mol respectively with RMSD values of 2.74 Å, 0.39 Å, 1.35 Å and 2.60 Å respectively. "The Ramachandran plot of the EPHX1 modeled structure revealed amino acid residues to be 87.9%, 9.1%, 1.5% and 1.5% in the most favored, additional allowed, generously allowed and disallowed regions respectively (Fig. 6)". The high quality of the model is also confirmed from the VERIFY 3D server as 72.88% of the residues of the modeled protein showed a score higher than 0.2 (Fig. 7). The total energy value of GSTM1 and its mutant model does not show any major difference. Since the RMSD values and the total energy after energy minimization are very high for the 13 mutant models as compared to the native structures we may presume that these mutations cause a significant change in the mutant structure of the protein with respect to the native structure. These mutations were also predicted to be functionally significant based on SIFT, PolyPhen, SNPs3D and MutPred results. Among these 13 mutations, 6 mutations showed high energy differences with higher RMSD values (Table 5). The detailed account of all the nsSNPs for the genes under study predicted by the algorithms used is shown in supplementary information S1–S4.

Table 3
Deleterious nsSNPs predicted by SIFT, PolyPhen, SNPs3D, SNPs&GO and MutPred algorithms.

Gene symbol	SNP ID	Amino acid change	Gene symbol	SNP ID	Amino acid change
EPHX1	rs72549341	Y374S	EPHX	rs58623835	R71H
	rs2234697	R49C	EPHX	rs148240980	R71C
GSTT1	rs17856199	F45C	GSTT1	rs2234953	E173K
GSTM1	rs142484086	R145W			
GSTP1	rs45549733	S150W	GSTP1	rs45543438	D58N
	rs45506591	E31V	GSTP1	rs4986949	D147Y
	rs1804666	G78E	GSTP1	rs71534294	D158H

3. Discussion

SNPs are the common form of genetic variations among individuals that are accountable for the majority of inherited traits, including a significant portion of breast cancer cases. Nonetheless, the exact mechanisms by which a SNP may result in a phenotypic change are for the most part unknown. About 2% of all the known single nucleotide variants associated with polygenic disease are non-synonymous SNPs in protein-coding regions (i.e., SNPs that alter a single amino acid in a protein molecule). As a result, it is anticipated that this class of SNPs is related to complex inherited disease traits. Therefore, to identify nsSNPs that affect protein functions and are related to breast cancer is an important task. The effect of many nsSNPs will probably be neutral as natural selection removes mutations on essential positions. Assessment of non-neutral SNPs is mainly based on phylogenetic information (i.e. correlation with residue conservation) extended to a certain degree with structural approaches. Much attention has been focused on modeling by different methods to determine the possible phenotypic effect of nsSNPs, and only recently interest is focused on functional SNPs which affect the regulatory regions or the splicing process. Moreover, because of their widespread distribution in the genome of a species, SNPs have become important target genetic makers in breast cancer diagnosis and treatment. Besides the numerous ongoing efforts to identify millions of these SNPs by high-throughput methods now, there is also a focus on studying associations between these genetic variations and breast cancer risk by using a molecular epidemiological approach. This plethora of SNPs poses a challenge to scientists in planning expensive population-based genotyping [14,15].

Currently, most molecular studies are focusing on SNPs located in coding and regulatory regions, yet many of these studies have been

Table 4
Score of nsSNPs and their impact predicted by all the algorithms.

Gene	SNP ID	SIFT score	Predicted impact	PolyPhen score	Predicted impact	SNPs3D score	Predicted impact	SNPs&GO score	Predicted impact	MutPred score	Predicted impact
EPHX1	rs72549341	0.00	Damaging	3.429	PRD	−3.81	Deleterious	4	Pathogenic	0.798	Supp. data
	rs58623835	0.00	Damaging	2.676	PRD	−2.37	Deleterious	6	Neutral	0.797	Supp. data
	rs2234697	0.00	Damaging	2.596	PRD	−0.45	Deleterious	8	Neutral	0.406	Supp. data
	rs148240980	0.02	Damaging	3.217	PRD	−3.40	Deleterious	5	Neutral	0.844	Supp. data
GSTT1	rs17856199	0.00	Damaging	2.993	PRD	−3.43	Deleterious	1	Pathogenic	0.657	Supp. data
	rs2234953	0.03	Damaging	2.240	PRD	−2.72	Deleterious	3	Neutral	0.242	Supp. data
GSTM1	rs142484086	0.01	Damaging	2.390	PRD	−0.44	Deleterious	3	Neutral	0.468	Supp. data
GSTP1	rs45549733	0.00	Damaging	2.509	PRD	−3.26	Deleterious	7	Pathogenic	0.740	Supp. data
	rs45543438	0.03	Damaging	1.178	PD	−2.22	Deleterious	2	Neutral	0.724	Supp. data
	rs45506591	0.02	Damaging	2.639	PRD	−2.18	Deleterious	3	Pathogenic	0.745	Supp. data
	rs4986949	0.00	Damaging	2.439	PRD	−1.94	Deleterious	5	Pathogenic	0.827	Supp. data
	rs1804666	0.01	Damaging	1.794	PD	−1.13	Deleterious	2	Pathogenic	0.622	Supp. data
	rs71534294	0.02	Damaging	2.234	PRD	−0.62	Deleterious	0	Pathogenic	0.710	Supp. data

Note: PRD—probably damaging; PD—possibly damaging, Supp. data—predicted impact is shown in supplementary data.

unable to detect their significant associations with disease susceptibility. To develop a coherent approach for prioritizing SNP selection for genotyping in molecular studies, an evolutionary perspective to SNP screening is applied. It was suggested that nsSNPs corresponding to conserved amino acids are more likely to be functionally significant to disease susceptibility [14,15]. It is becoming clear that the application of the molecular evolutionary approach may be a powerful tool for prioritizing SNPs to be genotyped in future molecular epidemiological studies. A computational approach was exploited to study the systematic analysis of SNPs by means of PolyPhen, SIFT, SNP3D, SNPs&GO and MutPred programs. Therefore, an effort was made to identify SNPs that can modify the function and expression of the genes which predispose to breast cancer.

SIFT is a program designed on the levels of conservation among the species. Information regarding the common position of the amino acid substitution relative to critical, structural and functional characteristics provides further understanding of evolutionary conservation. A low score indicates that the position is either severely gapped or unalignable and we also expect poor prediction at this position which is based on an already established classification [16]. Among the total 61 nsSNPs, 19 nsSNPs were predicted to be deleterious by SIFT with the score of ≤ 0.05. Several groups have tried to evaluate the ability of SIFT to distinguish between neutral and deleterious substitutions [17,18]. The performance of SIFT was also analyzed in healthy individuals by Cargill et al. [19]. In another study, Palmer et al. tried to validate the SIFT in MSHR gene, and found that predicted, tolerated substitutions L60V and R163Q by SIFT were in concordance with the experimental results [20]. So far, data on the validity of these algorithms have come from benchmarking studies based on the analysis of "known" deleterious substitutions annotated in databases, such as SwissProt [16,18]. Experimental studies of individual proteins have also confirmed the accuracy of SIFT [21,22].

The PolyPhen web server developed by Sunyaev and coworkers [23] uses both sequence and structural information to predict whether mutations are deleterious. BLAST is used to make a multiple sequence alignment from homologous proteins (30–94% sequence identity with the query), and PSIC [24] is used to construct a position-specific scoring matrix (PSSM) from this alignment. PolyPhen uses BLAST to identify proteins of known structure homologous to the query, but restricts its results to sequences in the PDB that are 50% identical to the query, and cases for which the amino acid in the structure is the same as the wild-type amino acid under study. PolyPhen uses the sequence alignment and the known structure to determine residue accessibility and proximity to ligands and interfaces with other subunits in the structure. A total of 24 nsSNPs of EPHX1, GSTP1, GSTT1 and GSTM1 genes were predicted to be damaging by PolyPhen.

Using SNPs3D, a total of 25 nsSNPs were found to be deleterious in all the four genes under study. For the accuracy of this program, Yue and Moult [25] developed a stability base model in SNPs3D to identify which non-synonymous single mutations have a deleterious effect on protein function in vivo using support vector machines (SVM). This model uses experimental or predicted protein structures to estimate the deleterious effects of mutations on the stability of folded proteins [26]. Using the stability model, 10,263 disease-causing mutations in 731 proteins were retrieved from HGMD [27] and the same 731 monogenic disease proteins were used to obtain 16,682 mutations not significant to the disease in the control set. Appropriate structural information was revealed for only 37% of mutations in the disease set and 14% of mutations in the control set, which were used in training and testing the data set. A set of factors was used to approximate the stability effect of a mutation on protein structure, including electrostatic interactions, over-packing and cavities, hydrophobic burial, surface accessibility, structural rigidity from crystallographic B-factors, backbone strain, buried charged or polar residues, and breakage of disulfide bonds. Ten-fold cross validation was used to train the SVM. The final true positive rate (TPR) and true negative rate (TNR) for the stability model were equal to 74% and 85%, respectively.

SNPs&GO, a new SVM based method uses different pieces of information derived from the Gene Ontology (GO) annotation to predict disease-related mutations. SNPs&GO was trained on a set of more than 33,000 mutations and tested with cross validation procedure over sets in which similar proteins were kept in the same dataset also for the calculation of the LGO score, derived from the GO data

Table 5
RMSD and total energy of native-structures and their mutant models.

	Total energy (native-structure) (kJ/mol)	Total energy (mutant models) (kJ/mol)	RMSD (Å)
GSTM1 (PDB ID 2f3m)			
rs142484086	−10405.183	−10060.114	1.87
GSTP1 (PDB ID 3pgt)			
rs45549733	−10448.928	−9713.515	2.67
rs45543438	−10448.928	−10629.228	1.86
rs45506591	−10448.928	−10127.256	2.08
rs4986949	−10448.928	−9892.780	2.45
rs1804666	−10448.928	−10484.553	1.81
rs71534294	−10448.928	−10555.339	1.97
GSTT1 (PDB ID 2c3n)			
rs17856199	−12983.094	−11260.579	2.33
rs2234953	−12983.094	−12819.685	0.99
EPHX1 (homology model)			
rs72549341	−7891.090	−6289.072	2.74
rs58623835	−7891.090	−7832.321	0.39
rs2234697	−7891.090	−7240.202	1.35
rs148240980	−7891.090	−6542.200	2.60

Note: Highly polymorphic nsSNPs are highlighted as bold.

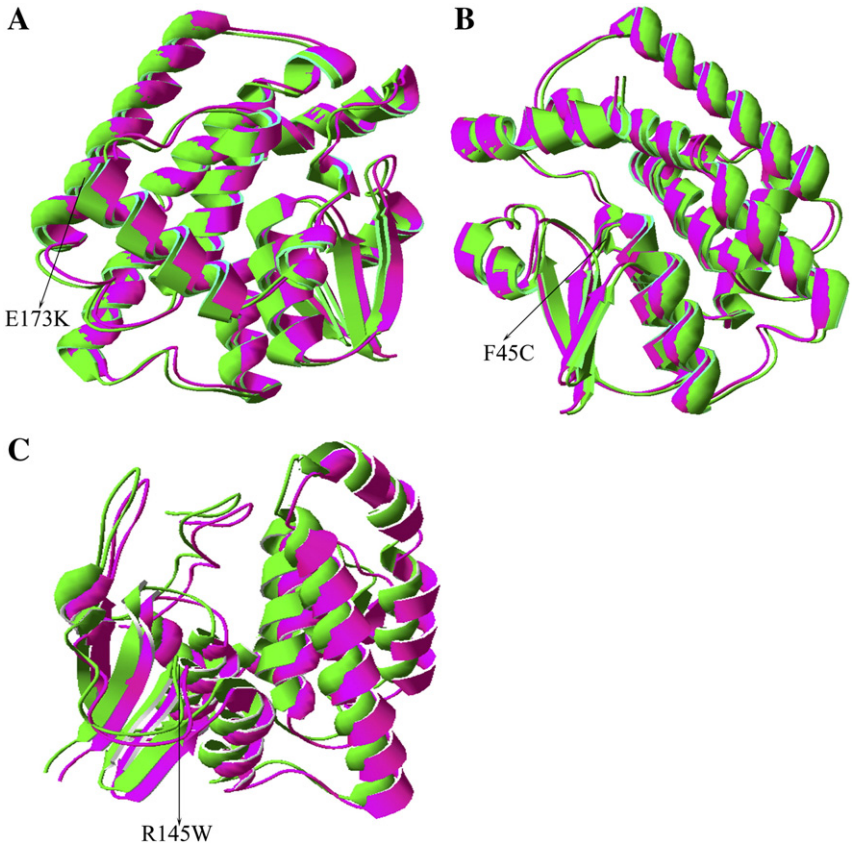

Fig. 2. {A} Superimposed 3D structure of GSTT1 (pink) with mutant E173K (green), {B} superimposed 3D structure of GSTT1 (pink) with mutant F45C (green), and {C} superimposed 3D structure of GSTMl (pink) with mutant R145 (green).

base. At an increasing input level of complexity, the performance is also increased, suggesting that on top of the sequence profile, LGO, derived from the protein GO annotation, is also a crucial added value to discriminate disease related polymorphisms from neutral ones. Finding the increased level of performance upon the addition of information to the input data corroborates the notion that support vector machines can capture all the correlations existing in complementary knowledge. Calabrese R et al., [28] in their analysis proved that presently SNPs&GO is one of the best scoring classifiers available for predicting whether a mutation at the protein level is disease-related or not.

MutPred employs the SIFT method for unfolding the evolutionary elements, together with PSI-BLAST, transition frequencies and Pfam results [29]. The structural descriptors in MutPred include the estimation of solvent accessibility and secondary structure by the PHD method [30], transmembrane helix estimation by TMHMM [31], coiled-coil structure estimation by MARCOIL [32], stability exploration by I-Mutant 2.0 and disorder exploration by DisProt [33]. Functional characteristics comprise exploration of DNA-binding residues, catalytic residues and posttranslational modification sites. The MutPred method estimates effects of an amino acid substitution on the set of defined properties of a protein and based on those estimates, predicts whether an amino acid substitution is likely to have phenotypic effects. In a recent study Thusberg J et al. [34], suggested that SNPs&GO and MutPred are the best methods to identify deleterious SNPs with accuracies of 0.82 and 0.81, respectively.

Protein structural analysis was carried out based on the screened results obtained from SIFT, PolyPhen, SNP3D, SNPs&GO and MutPred. Protein 3D structural information is an important feature for predicting the effects of deleterious nsSNPs. Moreover, protein structure analysis provides key information about the environment of a mutation. However, not all protein variants have 3D structures that are analyzed and deposited in PDB. Therefore, it is necessary to construct 3D models using molecular modeling protocols. This will help in understanding the adverse effects, a mutation can implicate on the concerned protein. To predict the structural deviations caused by single amino acid substitutions between mutant and native forms, their RMSD values were computed and examined using Swiss-PDB viewer. Computing the energy gives the information about the protein structure stability. Hence, total energy values (kcal/mol) of native and mutated modeled structures were also compared. Mutant structures of GSTP1 with PDB ID 3pgt at positions S150W, E31V and D147Y, mutant structures of GSTT1 with PDB ID 2c3n at position F45C and mutant structures of homology modeled EPHX1 at positions Y374S and R71C showed an increase in total energy level (less favorable change) and increase in RMSD value deviation in comparison with the native structure. To improve the strength of our analysis, the data was evaluated by the combination of all the tools used. Significant concordance was observed between the functional consequences of nsSNPs. By comparing the scores of the different methods used in this analysis, 6 nsSNPs with IDs rs45549733, rs45506591 and rs4986949 of GSTP1, rs72549341 and rs148240980 of EPHX1 and rs17856199 of GSTM1 were predicted to be functionally significant. Evidence suggests that polymorphisms in activating and detoxifying enzymes may interact to affect the level of DNA damage sustained by a specific tissue and ultimately influence disease risk [35]. Although some population-based studies have examined the association of polymorphic EPHX1, GSTT1, GSTM1 and GSTP1 genes with breast cancer [36,37], none of the 6 nsSNPs identified in this study, has so far been screened. The 6 identified nsSNPs could elevate the accumulation of carcinogenic metabolites by altering the activation and detoxification processes, thus leading to a modified risk of

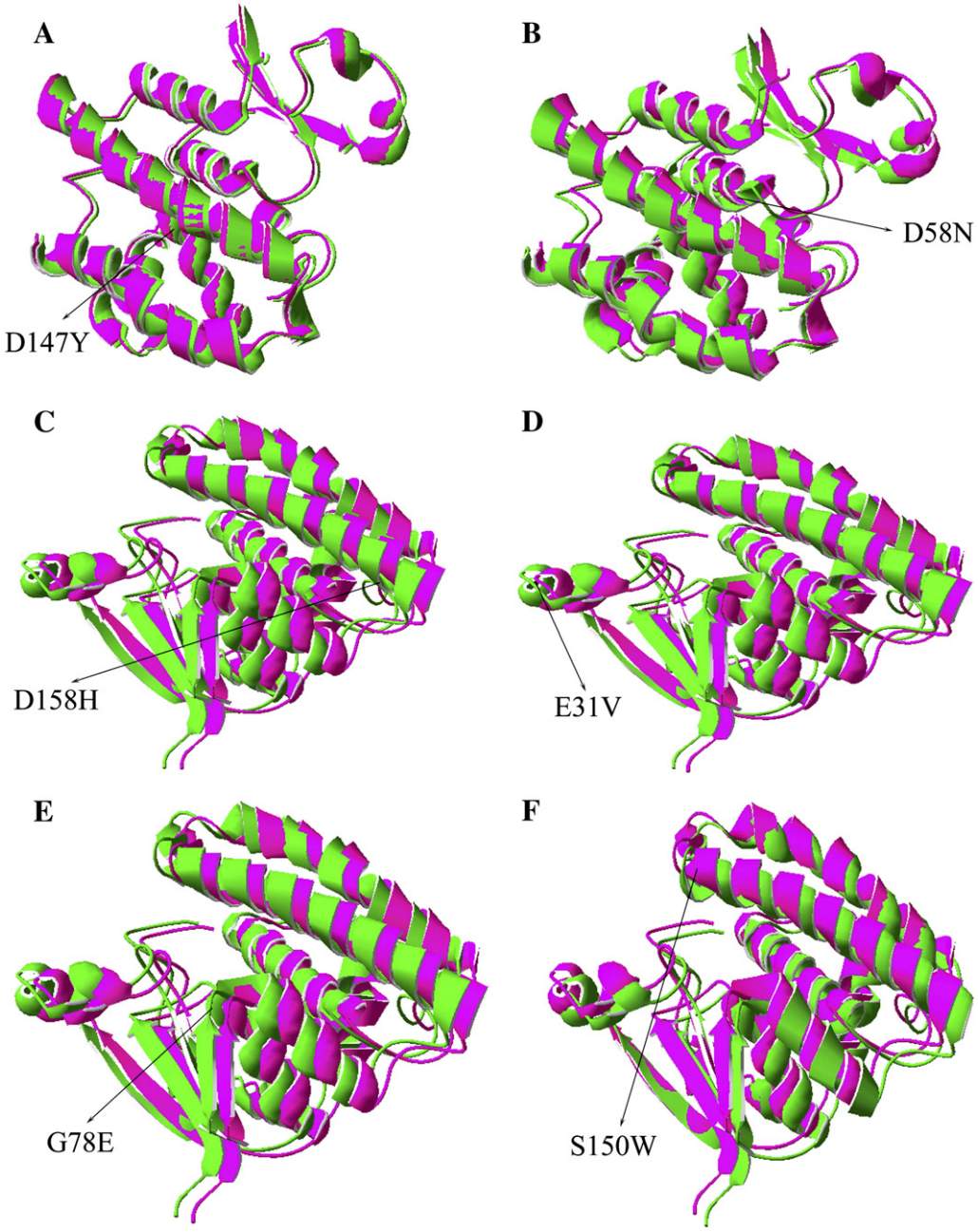

Fig. 3. Superimposed 3D structure of GSTP1 (pink) with its mutant models (green) {A} D147Y, {B} D58, {C} D158H, {D}, E31V, {E} G78E and {F} S150W.

breast cancer susceptibility in a group of women. Hence, they constitute a unique resource of SNPs that may considerably increase the power of breast cancer epidemiological studies.

The current study shows that utilization of SIFT, PolyPhen, SNPs3D, SNPs&GO and MutPred servers has facilitated to filter the number of reported SNPs in target genes. Additionally, this study has identified that rs45549733, rs45506591 and rs4986949 of GSTP1, rs17856199 of GSTM1 and rs72549341 and rs148240980 of EPHX1 are potentially polymorphic nsSNPs, which can be further hypothesized to have a plausible role in BC susceptibility. Animal models or breast cancer cell line based testing of these nsSNPs can help to determine if the functions of these proteins are indeed altered. Prioritization of nsSNPs before attempting to conduct large-scale population-based epidemiologic studies may not only potentiate the disease risk assessment but also curtail study costs.

4. Methodology

The methods followed in this study were the same as described previously [7,38]. Briefly, well-known and widely accessible computational techniques such as SIFT, PolyPhen, SNPs3D, SNPs&GO and MutPred are used to determine whether an nsSNP is neutral or deleterious. These methods have been trained on existing sets of mutation association data and use the sequence and structural information as an input to machine learning methods for phenotype extrapolation.

4.1. Screening of SNPs

The data on human EPHX1, GSTT1, GSTM1 and GSTP1genes was collected from the Online Mendelian Inheritance in Man (OMIM) and Entrez Gene on National Center for Biological Information

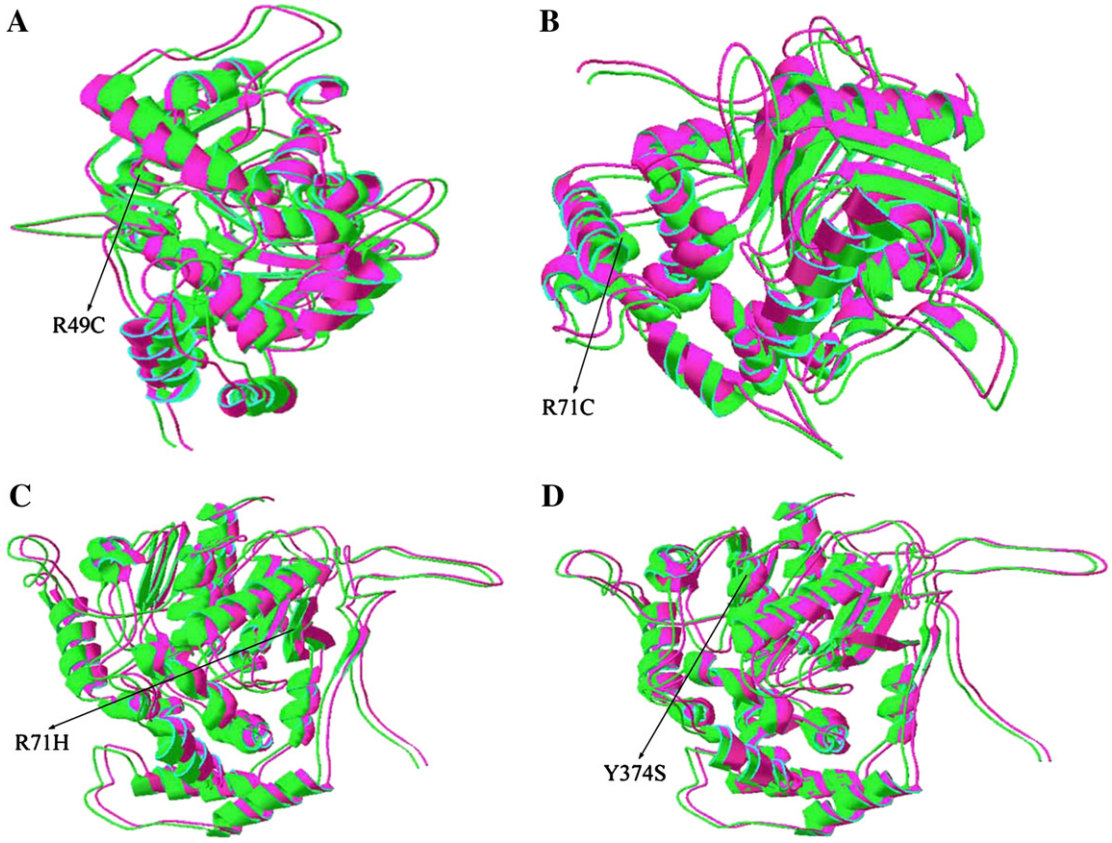

Fig. 4. Superimposed homology modeled structure of EPHXl (pink) with its mutant models (green) {A} R49C, {B} R71C, {C} R71H and {D} Y374S.

(NCBI) web sites. The information about SNPs such as the protein accession number (NP), mRNA accession number (NM) and SNP ID of all the genes was retrieved from the NCBI dbSNP (http://www.ncbi.nlm.nih.gov/snp/) and SwissProt databases (http://expasy.org/).

4.2. Prediction of tolerant and deleterious SNPs using SIFT

Sorting the intolerant from the tolerant available from http://sift.jcvi.org/ is a suite that can distinguish between functionally neutral and deleterious amino acid changes; hence it is widely used to detect if an amino acid substitution affects the function and phenotypic expression of a protein [39]. SIFT algorithm uses a modified version of PSIBLAST [40] and Dirichlet mixture regularization [41] and the underlying principle of this program is that it generates alignments with a large number of homologous sequences and assigns scores to each residue, ranging from zero to one. SIFT scores were classified as damaging (0.00–0.05), potentially damaging (0.051–0.10), borderline (0.101–0.20), or tolerant (0.201–1.00). The SIFT result is in the

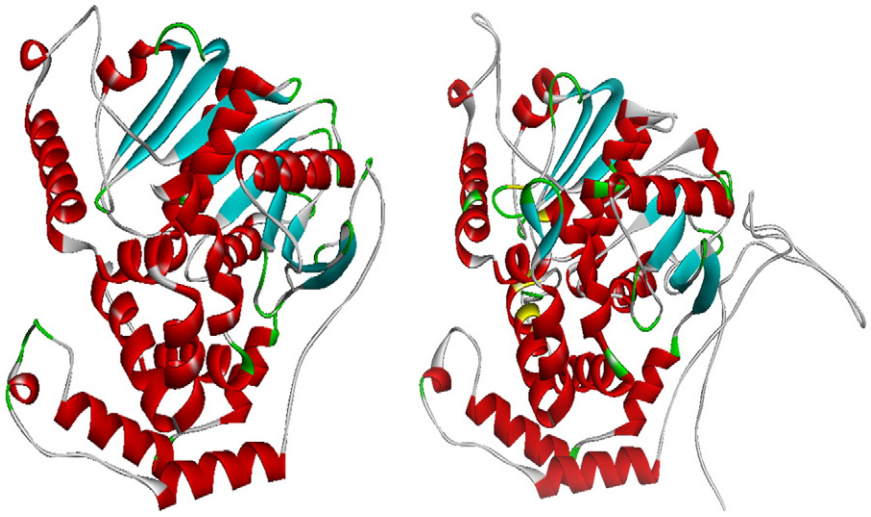

Fig. 5. {A} 3D structure of template (PDB ID 1QO7), {B} homology modeled structure of EPHXl protein predicted using MODELLER v9.10.

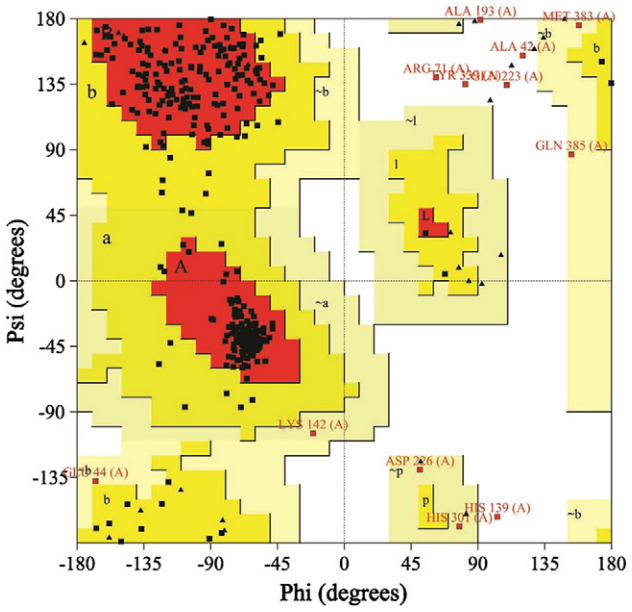

Fig. 6. Ramachandran plot of the EPHXl model. The most favored regions are colored red, additional allowed, generously allowed and disallowed regions are indicated as yellow, light yellow and white fields, respectively.

form of fraction of sequences with the representation of amino acids in color code: black (nonpolar); green (uncharged polar); red (basic); blue (acidic). A low fraction indicates that the position is a severely gapped position that has very little information.

4.3. Simulation for functional change in coding nsSNPs by PolyPhen

PolyPhen which is available from Harvard School of Medicine (http://genetics.bwh.harvard.edu/pph/) is a software tool which predicts the possible impact of amino acid substitutions on the structure and function of human proteins based on a combination of phylogenetic, structural and sequence annotation information characterizing a substitution and its position in the protein. The input of PolyPhen is an amino acid sequence or corresponding ID, the position of the amino acid variants [42]. PolyPhen searches for the three-dimensional protein structures, multiple alignments of homologous sequences and amino acid contact information in several protein structure databases. Then, it calculates position-specific independent count (PSIC) scores for each variant, and computes the difference of the PSIC scores between the two variants. The higher a PSIC score difference, the higher functional impact a particular amino acid substitution is likely to have. The PolyPhen scores can be classified as probably damaging (>2.00), possibly damaging (1.50–1.99), potentially damaging (1.25–1.49), or benign (0.00–0.99).

4.4. Protein stability prediction by SNPs3D

In SNPs3D (http://www.SNPs3D.org), the likely functional impact of non-synonymous SNPs is assessed by two methods. One method makes use of protein structure to identify which amino acid substitution significantly destabilizes the folded state and the second method identifies deleterious substitutions through analysis of the extent and nature of amino acid conservation at the affected sequence position [43]. The protein functional change was predicted from the accession numbers (NP_000111, NP_000552, NP_000843 and NP_000844 corresponding to EPHX1, GSTM1, GSTP1 and GSTT1 respectively). The algorithm makes use of a machine learning technique, the support vector machine (SVM), to assign each SNP as deleterious or non-deleterious to protein function. A negative value indicates that the mutated protein is deleterious and vice versa.

4.5. Prediction by SNPs&GO and MutPred

SNPs&GO, developed in the Laboratory of Biocomputing at the University of Bologna available from http://snps.uib.es/snps-and-go/ is an accurate method based on the support vector machine for predicting disease related mutations from the protein sequences which uses scoring with an accuracy of 82% and a Matthews correlation coefficient of 0.63. SNPs&GO collects unique framework information derived from protein sequence, protein sequence profile, and protein function. The output binary prediction (pathogenic/neutral) is taken into consideration [28]. MutPred is a Random Forest-based classification method that utilizes numerous attributes related to protein structure, function, and evolution. MutPred, developed by the Buck Institute for Age Research and Indiana University available at http://mutpred.mutdb.org/ is based upon SIFT and a gain/loss of 14 different structural and functional properties. It has been trained using the deleterious mutations from the Human Gene Mutation Database and neutral polymorphisms from Swiss-Prot [44]. The training data set contains 39,218 disease-related mutations from HGMD and 26,439 putatively neutral substitutions from Swiss-Prot. The output of MutPred contains a general score (g), i.e., the probability that the amino acid substitution is deleterious/disease-associated, and top 5 property scores (p), where p is the p-value that certain structural and functional properties are impacted [45].

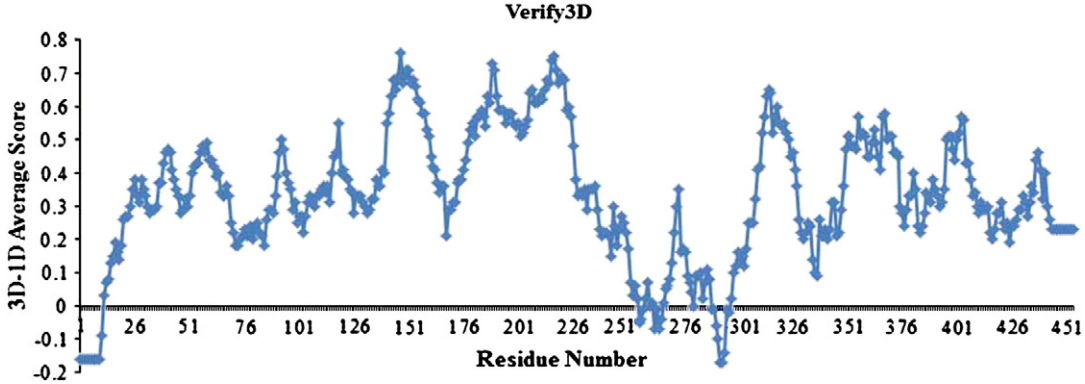

Fig. 7. The energy profile obtained from Verify 3D. The compatibility score above zero corresponds to the acceptable side chain environment.

4.6. Total energy and RMSD calculations

Structure analysis was performed to evaluate the structural stability of native and mutant protein models. The web resource BLAST [40] was used to identify the 3D structure of proteins coded by concerned genes. The 3D structure of the EPHX1 protein was not available, so the homology modeling approach was used for its 3D structure prediction. The modeling was performed by using the homology modeling program, MODELLER v9.10 [46]. The following steps were followed: template structure search using BLAST [http://www.ncbi.nlm.nih.gov]. The FASTA sequence of EPHX1 was submitted to NCBI BLAST. Following BLAST query, the structure of epoxide hydrolase (PDB ID: 1QO7) was selected as the template sequence (Fig. 5). The template was used to build the 3D structure of EPHX1 using MODELLER. The validation of the structured model was performed by using PROCHECK [47] and energy minimization was performed by Verify3D [48] and NOMAD-Ref server [10]. The overall stereo chemical quality of the protein was assessed by Ramachandran plot analysis [49]. The structures were visualized using Swiss PDB viewer and energy minimization for 3D structures was performed by NOMAD-Ref server, which uses Gromacs as the default force field for energy minimization based on the methods of steepest descent, conjugate gradient and L-BFGS [50]. Conjugate gradient method was used for optimizing the 3D structures. RMSD values were computed for both mutant and native structures, to check the structural divergences between them [51,52].

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.ygeno.2012.04.006.

References

[1] P.D. Pharoah, A. Antoniou, M. Bobrow, R.L. Zimmern, D.F. Easton, B.A. Ponder, Polygenic susceptibility to breast cancer and implications for prevention, Nat. Genet. 31 (2002) 33–36.
[2] A.B. Spurdle, J.H. Chang, G.B. Byrnes, X. Chen, G.S. Dite, M.R. McCredie, G.G. Giles, M.C. Southey, G. Chenevix-Trench, J.L. Hopper, A systematic approach to analysing gene–gene interactions: polymorphisms at the microsomal epoxide hydrolase EPHX and glutathione S-transferase GSTM1, GSTT1, and GSTP1 loci and breast cancer risk, Cancer Epidemiol. Biomarkers Prev. 16 (2007) 769–774.
[3] K.M. Egan, Q. Cai, X.O. Shu, et al., Genetic polymorphisms in GSTM1, GSTP1, and GSTT1 and the risk for breast cancer: results from the Shanghai Breast Cancer Study and meta-analysis, Cancer Epidemiol. Biomarkers Prev. 13 (2004) 197–204.
[4] J.W. Sull, H. Ohrr, D.R. Kang, et al., Glutathione S-transferase M1 status and breast cancer risk: a meta-analysis, Yonsei Med. J. 45 (2004) 683–689.
[5] K. Gudmundsdottir, L. Tryggvadottir, J.E. Eyfjord, GSTM1, GSTT1, and GSTP1 genotypes in relation to breast cancer risk and frequency of mutations in the p53 gene, Cancer Epidemiol. Biomarkers Prev. 10 (2001) 1169–1173.
[6] J. Sarmanová, S. Sùsová, I. Gut, M. Mrhalová, R. Kodet, J. Adámek, Z. Roth, P. Soucek, Breast cancer: role of polymorphisms in biotransformation enzymes, Eur. J. Hum. Genet. 12 (2004) 848–854.
[7] R. Rajasekaran, C. Sudandiradoss, C. George Priya Doss, R. Sethumadhavan, Identification and in silico analysis of functional SNPs of the BRCA1 gene, Genomics 90 (2007) 447–452.
[8] D. Fredman, M. Siegfried, Y.P. Yuan, H. Lehväslaiho, A.J. Brookes, HGVbase: a human sequence variation database emphasizing data quality and a broad spectrum of data sources, Nucleic Acids Res. 30 (2002) 387–391.
[9] E.M. Smigielski, K. Sirotkin, M. Ward, T.S. Sherry, dbSNP: a database of single nucleotide polymorphisms, Nucleic Acids Res. 28 (2000) 352–355.
[10] J. Jaruzelska, V. Abadie, Y. Aubenton-Carafa, E. Brody, A. Munnich, J. Marie, In vitro splicing deficiency induced by a C to T mutation at position-3 in the intron 10 acceptor site of the phenylalanine hydroxylase gene in a patient with phenylketonuria, J. Biol. Chem. 270 (1995) 20370–20375.
[11] J. Sarmanova, S. Susova, I. Gut, et al., Breast cancer: role of polymorphisms in biotransformation enzymes, Eur. J. Hum. Genet. 12 (2004) 848–854.
[12] M.M. Johnson, J. Houck, C. Chen, Screening for deleterious nonsynonymous single-nucleotide polymorphisms in genes involved in steroid hormone metabolism and response, Cancer Epidemiol. Biomarkers Prev. 14 (2005) 1326–1329.
[13] E. Lindahl, C. Azuara, P. Koehl, M. Delarue, NOMAD-Ref: visualization, deformation and refinement of macromolecular structures based on all-atom normal mode analysis, Nucleic Acids Res. 34 (2006) W52–W56.
[14] C. George Priya Doss, R. Rajasekaran, C. Sudandiradoss, K. Ramanathan, R. Purohit, R. Sethumadhavan, A novel computational and structural analysis of nsSNPs in CFTR gene, Genomics Med. 2 (2008) 23–32.
[15] Y. Zhu, M.R. Spitz, C.I. Amos, J. Lin, M.B. Schabath, X. Wu, An evolutionary perspective on single-nucleotide polymorphism screening in molecular cancer epidemiology, Cancer Res. 64 (2004) 2251–2257.
[16] T. Xi, I.M. Jones, H.W. Mohrenweiser, Many amino acid substitution variants identified in DNA repair genes during human population screenings are predicted to impact protein function, Genomics 83 (2004) 970–979.
[17] D. Chasman, R.M. Adams, Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation, J. Mol. Biol. 307 (2001) 683–706.
[18] S. Sunyaev, V. Ramensky, I. Koch, W. Lathe III, A.S. Kondrashov, P. Bork, Prediction of deleterious human alleles, Hum. Mol. Genet. 10 (2001) 591–597.
[19] M. Cargill, D. Altshuler, J. Ireland, P. Sklar, K. Ardlie, N. Patil, N. Shaw, C.R. Lane, E.P. Lim, N. Kalyanaraman, J. Nemesh, L. Ziaugra, L. Friedland, A. Rolfe, J. Warrington, R. Lipshutz, G.Q. Daley, E.S. Lander, Characterization of single nucleotide polymorphisms in coding regions of human genes, Nat. Genet. 22 (1999) 231–238.
[20] J.S. Palmer, D.L. Duffy, N.F. Box, J.F. Aitken, L.E. O'Gorman, Melanocortin-1 receptor polymorphisms and risk of melanoma: Is the association explained solely by pigmentation phenotype? Am. J. Hum. Genet. 66 (2000) 176–186.
[21] A.R. Brooks-Wilson, P. Kaurah, G. Suriano, Germline E-cadherin mutations in hereditary diffuse gastric cancer: assessment of 42 new families and review of genetic screening criteria, J. Med. Genet. 41 (2004) 508–517.
[22] E.Y. Zhang, D.J. Fu, Y.A. Pak, T. Stewart, N. Mukhopadhyay, S.A. Wrighton, K.M. Hillgren, Genetic polymorphisms in human proton-dependent dipeptide transporter PEPT1: implications for the functional role of Pro586, J. Pharmacol. Exp. Ther. 310 (2004) 437–445.
[23] S. Sunyaev, V. Ramensky, P. Bork, Towards a structural basis of human non-synonymous single nucleotide polymorphisms, Trends Genet. 16 (2000) 198–200.
[24] S. Sunyaev, J. Hanke, A. Aydin, U. Wirkner, I. Zastrow, J. Reich, P. Bork, Prediction of nonsynonymous single nucleotide polymorphisms in human disease-associated genes, J. Mol. Med. 77 (1999) 754–760.
[25] P. Yue, Z. Li, J. Moult, Loss of protein structure stability as a major causative factor in monogenic disease, J. Mol. Biol. 353 (2005) 459–473.
[26] P. Yue, J. Moult, Identification and analysis of deleterious human SNPs, J. Mol. Biol. 356 (2006) 1263–1274.
[27] P.D. Stenson, E.V. Ball, M. Mort, A.D. Phillips, J.A. Shiel, N.S. Thomas, S. Abeysinghe, M. Krawczak, D.N. Cooper, Human gene mutation database (HGMD): update, Hum. Mutat. 21 (2003) 577–581.
[28] R. Calabrese, E. Capriotti, P. Fariselli, P.L. Martelli, R. Casadio, Functional annotations improve the predictive score of human disease-related mutations in proteins, Hum. Mutat. 30 (8) (2009) 1237–1244.
[29] R.D. Finn, J. Mistry, J. Tate, P. Coggill, A. Heger, J.E. Pollington, O.L. Gavin, P. Gunasekaran, G. Ceric, K. Forslund, L. Holm, E.L. Sonnhammer, S.R. Eddy, A. Bateman, The Pfam protein families database, Nucleic Acids Res. 3 (2010) D211–D222.
[30] B. Rost, PHD: predicting one-dimensional protein structure by profile-based neural networks, Methods Enzymol. 266 (1996) 525–539.
[31] A. Krogh, B. Larsson, G. von Heijne, E.L. Sonnhammer, Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes, J. Mol. Biol. 305 (2001) 567–580.
[32] M. Delorenzi, T. Speed, An HMM model for coiled-coil domains and a comparison with PSSM-based predictions, Bioinformatics 18 (2002) 617–625.
[33] K. Peng, P. Radivojac, S. Vucetic, A.K. Dunker, Z. Obradovic, Length-dependent prediction of protein intrinsic disorder, BMC Bioinforma. 7 (2006) 208.
[34] J. Thusberg, A. Olatubosun, M. Vihinen, Performance of mutation pathogenicity prediction methods on missense variants, Hum. Mutat. 32 (4) (Apr 2011) 358–368.
[35] A. Saxena, V.S. Dhillon, M. Raish, M. Asim, S. Rehman, N.K. Shukla, S.V. Deo, A. Ara, S.A. Husain, Detection and relevance of germline genetic polymorphisms in glutathione S-transferases (GSTs) in breast cancer patients from northern Indian population, Breast Cancer Res. Treat. 115 (3) (Jun 2009) 537–543 [Epub 2008 Jun 24].
[36] T. Pongtheerat, M. Treetrisool, W. Purisa, Glutathione s-transferase polymorphisms in breast cancers of Thai patients, Asian Pac. J. Cancer Prev. 10 (1) (Jan–Mar 2009) 127–132.
[37] A. Unlü, N.A. Ates, L. Tamer, C. Ates, Relation of glutathione S-transferase T1, M1 and P1 genotypes and breast cancer risk, Cell Biochem. Funct. 26 (5) (Sep–Oct 2008) 643–647.
[38] M. Alanazi, Z. Abduljaleel, W. Khan, A.S. Warsy, M. Elrobh, Z. Khan, A.A. Amri, M.D. Bazzi, In silico analysis of single nucleotide polymorphism (SNPs) in human b-globin gene, PLoS One 6 (2011) e25876.
[39] P.C. Ng, S. Henikoff, Predicting deleterious amino acid substitutions, Genome Res. 11 (2001) 863–874.
[40] S.F. Altschul, T.L. Madden, A.A. Schaffer, Z. Zhang, W. Miller, D.J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res. 25 (1997) 3389–3402.
[41] K. Sjolander, K. Karplus, M. Brown, R. Hughey, A. Krogh, I.S. Mian, D. Haussler, Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology, Comput. Appl. Biosci. 12 (1996) 327–345.
[42] V. Ramensky, P. Pork, S. Sunyaev, Human non-synonymous SNPs: server and survey, Nucleic Acids Res. 30 (2002) 3894–3900.
[43] P. Yue, E. Melamud, J. Moult, SNPs3D: candidate gene and SNP selection for association studies, BMC Bioinforma. 7 (2006) 166.
[44] B. Boeckmann, A. Bairoch, R. Apweiler, M.C. Blatter, A. Estreicher, E. Gasteiger, M.J. Martin, K. Michoud, C. O'Donovan, I. Phan, S. Pilbout, M. Schneider, The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003, Nucleic Acids Res. 31 (2003) 365–370.

[45] B. Li, V.G. Krishnan, M.E. Mort, F. Xin, K.K. Kamati, D.N. Cooper, S.D. Mooney, P. Radivojac, Automated inference of molecular mechanisms of disease from amino acid substitutions, Bioinformatics 25 (21) (2009) 2744–2750.
[46] A. Sali, T.L. Blundell, Comparative protein modeling by satisfaction of spatial restraints, J. Mol. Biol. 234 (3) (Dec 5 1993) 779–815.
[47] R.A. Laskowski, J.A. Rullmannn, M.W. MacArthur, R. Kaptein, J.M. Thornton, AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR, J. Biomol. NMR 8 (1996) 477–486.
[48] J.U. Bowie, R. Lüthy, D. Eisenberg, A method to identify protein sequences that fold into a known three-dimensional structure, Science 253 (1991) 164–170.
[49] G.N. Ramachandran, C. Ramakrishnan, V. Sasisekharan, Stereochemistry of polypeptide chain configurations, J. Mol. Biol. 7 (1963) 95–99.
[50] M. Delarue, P. Dumas, On the use of low-frequency normal modes to enforce collective movements in refining macromolecular structural models, Proc. Natl. Acad. Sci. 101 (2004) 6957–6962.
[51] J.H. Han, N. Kerrison, C. Chothia, S.A. Teichmann, Divergence of inter domain geometry in two-domain proteins, Structure 14 (2006) 935–945.
[52] S.D. Varfolomeev, I.V. Uporov, Bioinformatics and molecular modeling in chemical enzymology. Active sites of hydrolases, Biochemistry. (Mosc.) 67 (2002) 1099–1108.