Identification of Lapatinib Derivatives and Analogs to Control Metastatic Breast Cancer-specific to South Asian Population-a Pharmacogenomic Approach

HIMA VYSHNAVI A M[1], P K KRISHNAN NAMBOORI[*]
Computational Chemistry Group (CCG), Amrita Molecular Modeling And Synthesis Research lab,
Amrita School of Engineering, Coimbatore, Amrita Vishwa Vidyapeetham, INDIA

Abstract—Introduction: The need for designing and developing personalized drugs for various diseases has become a challenging research topic at present. The individual variation towards susceptibility of a drug depends upon the genomic, epigenomic, metagenomic and environmental genomic factors. Areas covered: The 'Single Nucleotide Variant (SNV)' has been identified as the functional feature corresponding these factors. The need for personalized drug designing for the ERBB2 mutation related to Breast Cancer has been proposed by taking the South Asian (SA) population as the test sample. The SNVs corresponding to SA population for the ERBB2 mutation has been identified. The 'convolution neural network-based deep learning technique' (DeepCNN) has been used for computing the clinical significance of the SNVs, whose clinical significance values are unknown, using the functional variants as the attributes for the ethnic group. Expert opinion: The population has been classified into four groups based upon the probability of variants. The population-specific gene models and protein models have been designed. The potential molecules that control ERBB2 mutation specific to the South Asian population have been identified through docking/interaction score values

1. Introduction

THE 'triple negative breast cancer (TNBC)' is a type of breast cancer characterized by its unique level of metastasis and molecular pattern and with clinically negative response for expressions of estrogen and progesterone receptors and HER2 proteins (That is why it is known as triple negative breast cancer). The common suggested drugs for controlling the over expression of HER2 are Trastuzumab, Pertuzumab, Lapatinib, Neratinib and Afatinib[1]. However, in most cases of HER2 mutations, a very fast anti-cancer drug resistance is developed, making further treatment difficult [2]. However, Lapatinib has been suggested as a potential drug for controlling the 'metastatic breast cancer (MBC)' [3]. The Lapatinib has been identified as a dual 'tyrosine kinase inhibitor' against the mutations, HER1 (EGFR) and HER2 (ERBB2) and is found to be useful especially for patients who develop resistance towards the common drugs, Trastuzumab and Taxanes [4], while taking along with endocrine therapies.

The drug inhibits phosphorylation of the gene and prevents the signaling mutations 'ERK1/2 and P13l/AKT' by interacting with the 'ATP binding site' of the receptor [5]. It has been found that Lapatinib is sufficiently active while using it as a single drug or in combination with Trastuzumab, Capecitabine and Ixabepilone, especially on 'Trastuzumab-resistant cell lines' [6]. The prognosis and treatment strategies of breast cancer vary for different classes of the disease namely, metastatic, ductal carcinoma in situ (DCIS), invasive ductal carcinoma (IDC) and triple-negative breast cancer [7]. The ERBB2 amplification is found in most of the 'invasive breast carcinomas' resulted by stimulating growth factor signalling pathways such as P13K-AKT-mTOR pathway [8]. The HER2 mutation is generally unaffected by most of the common drugs/ligands [9], which may give way to the propagation of the 'growth factor signalling pathways', leading into unexpected outburst or continuation of the disease [10].

However, the drug response, to a large extend, is found to vary from person to person. The 'individual variation' towards proneness of the disease as well as gene expression profile of the mutations suggests a pharmacogenomic approach in designing a drug molecule. This necessitates the need for assessing the involvement of genomic, epigenomic, metagenomic and environmental genomic factors to study the susceptibility of Lapatinib, its derivatives and the analogs of both [11].

Among them, 'genomic factor' includes the populations-wise variations associated with mutations and identifying their significant genomic variants (genomic markers) [12]. The epigenomics involves the variations within the DNA structure and histone proteins such as DNA methylation, histone modification and miRNA mediated modification without undergoing mutations leading to some functional alterations [13]. The metagenomic includes the analysis of the

involvement of associated microbes in altering the functions of DNA [14]. The environmental genomics includes the involvement of environmental factors in gene mutations [15]. In fact, identification of the genetic markers corresponding to these 'omic factors' is a significant step in Pharmacogenomics [16].

The 'Single Nucleotide Variation (SNV)' has been identified as the most suitable variant for incorporating individual variation [17-19]. However, the significance of each SNV for any ethnic group depends upon its occurrence, leading into the need for a 'frequency-wise prioritisation and classification' of the variants. The 'CNN based deep learning technique' (deep convolution neural network-DeepCNN) can be used as a useful platform for predicting the significance of the SNVs if they are unknown or not clinically computed [20].

The analysis of clinical trials suggests that the progression of the disease and death are found to be decreasing among patients from the US population while taking Lapatinib [21]. However, the effect of consuming Lapatinib among Indian population is not that profound or in some cases at least it is having the reverse effect [22].This is a clear case of variation of drug action among different ethnic groups. The 'South Asian (SAS)' population has been chosen for the analysis as well as the drug action of Lapatinib, its derivatives and analogs.

The Side effects are also reported as varying among patients in both the ethnic groups [23]. Hence there is a need to identify a new strategy to study the individual variation towards the drug susceptibility among the populations and design and develop a population-specific potential molecule to down-regulate *ERBB2* mutation.

2.   Materials and Methods

An outline of the methodology adopted for the pharmacogenomic analysis has been included in Supplementary material 1. The details of which are included in the following sections.

2.1 Genomics

The genes related to breast cancer have been identified using Gene Cards and NCBI [24, 25]. Expression Atlas, a database with the data consisting of expression analysis of genes in various disease tissues has been used for collecting the microscopic images and computing the gene expression profile of TNBC [26]. The SNVs corresponding to genomics, epigenomics, metagenomics and environmental genomics for the gene *ERBB2* have been identified for the SA population using the tool, 'SNPNexus' from the1000 genome project. The frequency of occurrence and the conservation score values for the above SNVs have been noted down [27].

2.2 Functional analysis of variants

The functional analysis of SNVs has been carried out using the tools *SIFT*(measuring the effect of amino acid substitution on protein function based on sequence homology and the physical properties), *POLYPHEN* (finding the impact of amino acid substitutions on the structure and function of human proteins), *Combined Annotation-Dependent Depletion -CADD* (scoring the deleteriousness of single nucleotide variants as well as insertion/deletions variants in the human genome.), *Mutation Assessor* (quantifying the functional impact of protein mutations), *FATHMM* (measuring the functional consequences of non-synonymous single nucleotide variants (nsSNVs) and non-coding variants.), *FATHMKL* (functional Consequences of Non-Coding and Coding Single Nucleotide Variants) and *Pmut* scores (pathological variants on proteins) [28]. Based upon the clinical significance values, the SNVs have been classified into four functional classes namely, class-1 (1.0-0.80), class-2 (0.79-0.60), class-3 (0.59 to 0.40) and class-4 (0.39-0.00). However, out of 60 SNVs identified for the gene, only 35 variants were having experimental clinical significance values. The clinical significance values of the remaining 25 SNVs have been predicted using deep CNN based deep learning technique following a multiclass classification strategy. Few of the SNVs with their functional significance score values have been included in (Table 1). The complete list of the SNVs has been included in the Supplementary material 2.

Table 1. The SNVs with their functional significance score
Where A (Freq. SAS), B (SIFT), C (Polyphene2), D (Pmut score), E (Conservation Score), F (FATHMM), G (FATHMMK), H (Raw Score), I (PHRED score) and J (Clinical significance) are the parameters of variants.

SN VS	A	B	C	Mutation		F	G	CADD		
				D	E			H	I	J
I	0	0.1	0.47	0.69	0.000	0.390	0.007	0.400	8.398	0.503
II	0	0.03	0.12	2.38	0.025	0	0	0	0	0.626
III	0.13	0.04	0.73	2.19	0.007	0.118	0.033	0	0	0.766
IV	0	0.06	0.10	1.04	0.001	0	0	0	0	0.869
V	0	1	0.00	0.69	0.003	0	0	0	0	0.890
VI	0	0.15	0.01	1.44	0.029	0	0	0	0	0.744
VII	0	0.22	0.13	3.66	0.152	0.079	0.001	0.514	9.513	0.873
VIII	0	0.03	0.84	2.99	0.012	0.017	0.007	-0.083	1.561	0.844

2.3 Classification of variants

The SNVs have been added in the *ERBB2* gene based upon their 'clinical significance scores values (CSSV)' generating four classes of model genes for the population namely, CSSV between 0.80 and 1.00 as class 1 (or probability of variation 80-100%), CSSV between 0.60-0.79 as class 2 (or probability of variation 60.00-0.79%), CSSV between 0.40-0.59 as class 3 (or probability of variation 40-59%) and CSSV between 0.00-0.39 as class 4 (or probability of variation 0-39%).

The DeepCNN model has been used to predict the CSSV of SNVs whose values are unknown or not yet clinically characterized [29]. The reported variants were considered as training set and the variants whose CSSV are unknown was included in the testing set. The model has been designed on python platform and has been optimized for 1000 epochs.

2.4 Generation of Model

The mutant gene models for each class of SNVs have been generated by inducing the variation into the original gene sequence across the population. The corresponding protein models have been generated by transcription and translation of the gene models using Expasy Translate [30] and homology modelling of the translated amino acid sequences through Swiss Model. The model structures have been evaluated through the GQME score, LQE ('Local Quality Estimate')and 'global quality estimate-QMean' (denoting the 'degree of nativeness'), 'Ramachandran plot' (which gives the chemical stability of residues presentbased on the torsional angles - phi (φ) and psi (ψ) - of the amino acid residues) [31].

2.5 Docking Studies

All the docking studies have been carried out using the CDOCKER algorithm [32] using Accelrys Discovery Studio. The parameters used are: Force Field: CHARMm, Fold Potential was set as False, Grid Extension: 8, Ligand Partial charge method: CHARMm, Final Minimization: Full Potential and the Sphere size was set to: -0.0493, 3.29459, 10.9589, 17.

2.6 The design of personalized drugs

The inadequacy of Lapatinib in binding with the model protein molecules as efficiently as 'the control standard protein', 1XKK suggests the requirement for designing new molecules specific to the model proteins or some derivatives of the drug-specific to the SAS population [33]. The 'structure-based pharmacophore mapping' of the model proteins has been carried out using ZINCPharmer [34] to identify molecules in the ZINC database having complementary pharmacophoric similarity with the model proteins, which gave 385 molecules.

The drug action of Lapatinib is due to the binding of the '-NH-CH2-CH2-(SO2)CH3' end with the ATP-binding domain of the protein. Further, the kinase domain of Lapatinib has been identified and modifications have been made within the domain by substituting 9 groups (Fig 1): -NH2, -CH3, CH3CH2OCO-, CH3(CH2)3OCO-, CH3(CH2)2OCO-, CH3SO2CH2CH2-, -CH(CH3)2 and cyclohexyl as shown in (Fig 2), generating 9 Lapatinib derivatives [35].
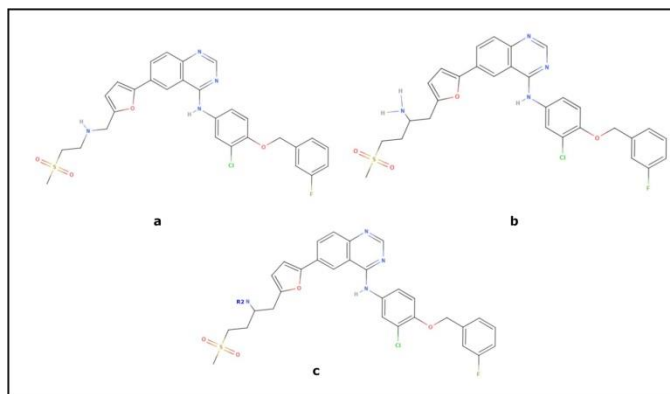

Fig 1. Lapatinib and its kinase domain. a. Lapatinib, b. Lapatinib with kinase domain highlighted with the -NH group repaced by (-CH2-NH2) and c. Lapatinib with a substitution in kinase domain-proposing substitutions in the -R group
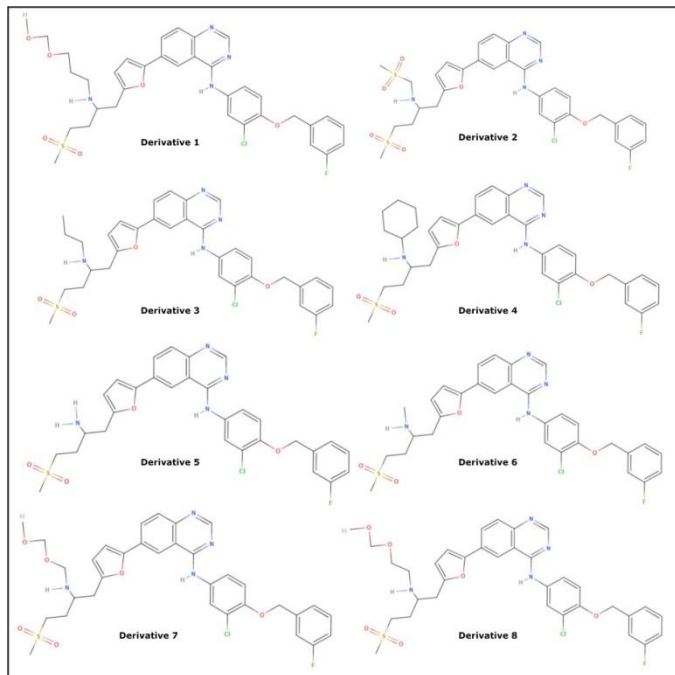
Fig 2. The proposed derivatives of Lapatinib

Moreover, 70 pharmacophoric analogs of these derivatives have been generated from Zinc database. Altogether, Lapatinib, 9 derivatives of Lapatinib, 70 pharmacophoric analogs of these derivatives and 384 structure based pharmacophoric analogs of model proteins, (a total of 464) have been taken for the docking studies.

The molecules with high docking scores were further subject to Molecular Dynamic simulation and ADMETox screening to identify the potential drugs/analogs for SAS population to control ERBB2 mutation [36].

2.7 Evaluating docking studies by MD simulation of ligand-target complex

The stability of ligand-target complex has been studied by analysing the retention of ligand molecule in the functional domain of the target protein through Molecular Dynamic (MD) simulation. The simulation has been carried out with NVE ensemble using Molecular Operating Environment (MOE) where total number of particles in the system (N), volume of the system (V), as well as the total energy of the system (E) is kept constant [37]. This meets the kinetic requirement of the drug action and has been carried out. The molecules were further subjected to pharmacokinetic screening using SwissADME and ProToxII [38, 39], drug likeness and QSAR analysis using MOE [40].

3. Results

3.1 Genomics

It has been identified that approximately1600 genes are directly or indirectly involved in causing breast cancer. Hence to study the influence of each mutation in causing the disease, a gene expression profile has been carried out from the microscopic images of the cancer cell line samples kept in the 'Expression Atlas. It has been found that *TOP2A* and *ERBB2* genes are highly expressed in breast cancer cell lines Fig 3. However, it is reported that the *ERBB2* mutation is functionally more significant than *TOP2A* for causing metastatic cancer. Hence *ERBB2* has been included in the analysis.
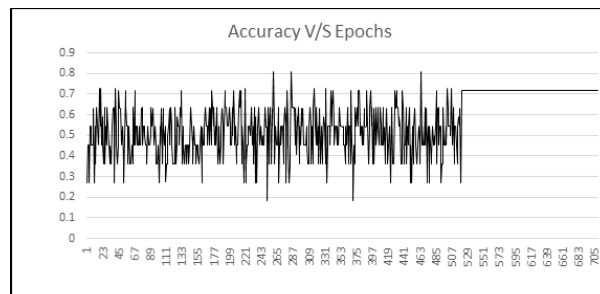

Fig 3. The accuracy versus epoch (iteration) graph of the DeepCNN model

The population analysis of *ERBB2* variations showed that; 314 SNVs were found in SAS, 318 SNVs in AMR, 257 in EUR, 485 SNVs in AFR and 239 variations among EAS populations. There were 43 variations found commonly among all the 5 populations; rs2517959, rs2643194, rs1565923, rs2934967, rs2517960, rs2517951, rs34284966, rs4252639, rs2517961, rs2643195, rs148150809, rs2952156, rs148325095, rs188492389, rs34006795, rs2952157, rs2904765, rs111700892, rs2952155, rs11653998, rs2904766, rs2934971, rs4252608, rs111947409, rs55717377, rs35542329, rs1810132, rs115334808, rs2904768, rs79747793, rs4252624, rs2088126, rs185106091, rs12947247, rs4252627, rs1136201, rs35797841, rs4252596, rs113231716, rs903506, rs1058808, rs4252661 and rs2952158.

3.2 Functional analysis and Classification of variants

During the analysis, 133 SNVs for AFR, 80 SNVs among AMR, 200 variants among EAS, 230 variants within EUR population and sixty SNVs for the South Asian (SAS) population of the gene *ERBB2* have been identified to be involved in epigenomic and metagenomic modifications. Other than 25 SNVs out of a total of 60 for the SAS population, the CSSV of SNVs of all other populations has been reported. To include the CSSV values of the remaining 25 SNVs of SAS population, the Deep-CNN based predictive model has been used, which shows a prediction accuracy of 72.56% (Supplementary material 3).

After the classification, for SAS population, 28 SNVs are found in 'class1' (with CSSV 0.80-1.00), 13 SNVs are in 'class-2' (with CSSV 0.60-0.79), 14 SNVs are in 'class-3' (with CSSV 0.40-0.59) and the remaining 5 SNVs are in 'class-4' (with CSSV 0.00-0.39) (Table 2). The SNVs

included in 'class-1' are the most frequent and significant SNVs corresponding to susceptibility towards drug action among each population. Hence the 'primary gene model' to be designed should be for this class by adding the variants of these SNVs to the ERBB2 gene. The classification models for other populations have been included in the supplementary materials.

Table 2. The SNVs after classification

Model	Class 1	Class 2	Class 3	Class 4
SAS	rs376524324, s144533600, rs113619125, s199668084, rs4252633, rs547041600, rs527779103, s566053951, rs373824622, s568793816, rs531563820, s574436396, rs141116145, s149210045, rs4252656, rs1136201, rs377649991, s539683889, rs564064363	rs376524324, rs144533600, rs113619125, rs199668084, rs4252633, rs547041600, rs527779103, rs566053951, rs373824622, rs568793816, rs531563820, rs574436396, rs141116145, rs149210045, rs4252656, rs1136201, rs377649991, rs539683889, rs564064363	rs565582943, rs35466868, rs1801201, rs182572604, rs142783371, rs199905364, rs562773948, rs546886845	rs193171026, rs554689197, rs534825855, rs4252596, rs201470725
EAS	rs2517961, rs2952158	rs2517960, rs11653998 rs4252627, rs903506 rs1058808, rs2904765 rs2517951, rs2904766 rs2517959, rs55717377	NIL	rs192213366, rs117641771, rs569374038, rs562212409, rs182525806
AFR	rs2952156, rs2088126 rs2952155, rs2934971 rs2952157, rs1565923 rs2643194, rs2952158 rs2517961	rs2643195, rs2517954	NIL	rs545122442, rs542280223, rs531509829, rs188039765, rs553800936, rs564868050, rs547297049, rs570062004, rs367611030, rs535959708, rs548386033, rs563737584, rs531034426, rs546172166, rs529439569, rs558278977, rs552970665, rs528451367
EUR	rs2517960, rs11653998	rs2517960, rs11653998 rs4252627, rs903506 rs1058808, rs2904765 rs2517951, rs2904766 rs2517959,	NIL	rs2132141, rs4252646 rs4252647, rs4252651 rs4252652, rs535716947 rs150165942, rs4252660 rs148325095,

		rs5571737 7 rs2904768 , rs1810132 rs2952157 , rs2643195 rs2934971 , rs1565923 rs2934967 , rs2952156 rs2088126 , rs2643194 rs2952155 , rs2517961 rs2952158		rs1851060 91 rs1832866 52, rs5654080 41 rs1833734 85, rs5660671 22 rs1177741 98, rs1402721 56 rs1810587 40, rs5515110 55 rs5358053 29, rs1510273 02 rs1411161 45, rs5623094 75 rs5401162 45, rs1851183 99 rs1924866 70 ,rs143319 500
AMR	rs295215 5, rs251796 1 rs295215 8	rs5571737 7, rs2904768 rs2904765 , rs2643195 rs2517954 , rs2517956 rs2517959 , rs2904766 rs1810132 , rs2952156	rs4252627, rs11653998 rs903506, rs1058808 rs2517955, rs2517960 rs2517951	rs5660540 09, rs5440007 28 rs5777676 74, rs5626019 00 rs5322125 27, rs1880397 65 rs1917605 37, rs5321562 55 rs3676110 30

3.3 Generation of Model

De By transcription and translation, sequences corresponding to protein models for SAS, AFR, AMR, EUR and EAS have been generated (Supplementary material 4). The four protein models of different gene classes for SAS

population are as follows:

>model 1

RPPKVLGLQVMAVVCDGWEAVWCLGVCGLPYPLSVP
LSPGSIRDGWCGLPICLPPSGHLPDIHGAAGDTAYALWL
PLRPCPGKPQKPGLPGPAELVYADCQGMHLGSLQVSPE
QTPMSTRGMSYLEDVRLVHRDLAARNVLVKSPNHVKI
TDFGLARLLDIDETEYHADGGKVRMCGVMVCDGGCW
EGWVRSHGWREDESWDGENYGATSACEGREGAACAP
PCRVCALPRIRERPGRVCLLASHLPLLPAMMLDSMTFG
AKPYDGIPAREIPDLLEKGERLPQPPICTIDVYMIMVKC
AWLSCAGCLEEGGRSWVEEPTRGMKGDQDVCRPRSPS
MLGASKPSCIPFTVKVQSHETIAPLKSSLGDRAKPHLKK
IKKQTKRKKKLKGKLEEMPKVLAEDPRVWCYFSTTMI
DSECRPRFRELVSEFSRMARDPQRFVVIQVLGLCAPSLP
VAKSTLLQRVGRRDES

>model 2

LKNMRGRTFAVLFTTLPPAQNLAYGRRMVLHMNKTTP
MDIMSNILKCSNAPHMITTHCNLNLPDSSDPSISASMPC
MWVPYSSNVYMNKALYFSWSEEGNRPMNYINGGRAI
QWKAEQKLIGMGQICNVLAWMLVRLSMVVIFENISRKI
KKKVEIVLASGGKWDYVPDQMGFLHVGQAGLELPTSG
DPSASASQSARITGVSHRAWPGTLFIFVPLKVCTVCVYN
QLKTKKSMPVIPALKEAEAGGRSPEVRSSRPAMQMIFL
VAFSFLMDCLLLISILVNSRLLVFFVLM

>model 3

SPAFWASMAASMSGKTAEAWAPRTCMMLDSMTFGAK
PYDGIPAREIPDLLEKGERLPQPPICTIDVYMIMVKCAW
LSCAGCLEEGGRSWVEEPTRGMKGDQDVCRPRSPSML
GASKPSCIPFTVKVQSHSMIAPLKSSLGDRAKPHLKKIK
KQTKRKKKLKGKLEEMPKVLAEDPRVWCYFSTTMIDS
ECRPRFRELVSEFSRMARDPQRFVVIQVLGLCAPSL

>model 4

YPCPQEAYVMAGVGSPYVSRLLGICLTSTVQLVTQLMP
YGCLLDHVRENRRSLGSQDLLNWCMQTPATCHDARLL
SRTSGSVHMCVAELCWLPGGGWEVLGGGAHKGHERG
PGCMMSAKIPGVGVMLGWGGATILPL

The 3D protein target models have been generated by 'homology modelling'. Three structures have been obtained for each model protein sequence. The model structures for different population are shown in Supplementary material 5.

The quality of the model structures were evaluated using the parameters, *GQME (Global Model Quality Estimation),* QMean (the degree of nativeness), Cbeta, All-atom and Solvation and Torsion angle potential. The Z-score values (standard deviation from the mean value) of these parameters have been included in (Table 3).

The high GQME score values of the models support reliability of the structures. The QMEAN score measures the degree of nativeness. The low (close to zero) QMEAN score value suggests the model structure to be comparable with the experimental ones. Similarly, the local geometry of the models has been studied using a short-range torsion angle potential (over three consecutive amino acids) and a long-range interaction using C-beta level and all atoms. The solvation energy is calculated to investigate the possibility for solvation. All these evaluation techniques support the quality of model

structures [41].

Table 3. Protein Model Evaluation Parameters

Protein		GQME-Z-score	QMEAN-Z-score	C beta-Z-score	All atom-Z-score	Solvation energy-Z-score	Torsion angle energy-Z-score
SAS	Model 1	0.79	-0.46	-0.21	-0.04	-0.08	-2.08
	Model 2	0.8	-0.3	-0.17	-0.13	-0.17	-8.25
	Model 3	0.54	-0.74	-0.13	-0.22	-0.26	-7.42
	Model 4	0.32	0.02	-0.09	-0.31	-0.35	-7.59
AMR		0.02	-2.82	-0.31	-2.64	-1.54	-2.17
EUR		0.8	-2.31	-0.11	-0.15	-0.08	-2.12
AFR		0.02	-2.63	-5.14	1.10	-1.28	-1.37
ESA		0.02	-5.32	-4.02	-4.02	-2.09	-3.84

The model is further evaluated through MD simulation by studying the time-evolved behaviour of the protein [42]. The deviation of model structure from the template can be studied through 'root mean square deviation (RMSD)' during the 'molecular dynamics (MD)' simulations. The smaller the deviation, higher will be the stability of the structure of protein. Screened models were having RMSD variations below 3 **Å** suggesting the structures to be stable.

The kinase domain of the protein models has been used for the docking studies (Fig.4). The reported kinase protein, 1XKK has been used as the standard target.

The protein characterization has been carried out for the protein models of different populations. It has been found that the number of amino acids, molecular weight, PI value, GRAVY value, and aliphatic index varied for different protein models. The protein structure stability has been studied through Ramachandran plot as well as molecular dynamic simulation. It has been identified that the properties of protein models for EAS and SAS populations were similar where as it is different for other populations (Supplementary material 6.)

3.4 Pharmcophore Mapping and Interaction analysis

NThe amino acid residues LEU-718, GLY-719, SER-720, GLY-721, ALA-722, PHE-723, GLY-724, THR-725 and MAT-726 were considered as the pharmacophore of the protein model. The pharmacophore mapping has been carried out against Zinc database which consists of structures of synthesized chemical molecules including FDA approved drugs. 2400 molecules were found to have pharmacophores complementary to the target protein. These molecules were further subjected to descriptor analysis, interaction analysis. Lapatinib was having comparatively good interaction with the target proteins when compared to the other molecules. Thus Lapatinib has been considered for further analysis.

The druggability has been primarily evaluated through docking studies using 1XKK and the model proteins as the target, Lapatinib, 9 derivatives of Lapatinib, 70 pharmacophoric analogs of these derivatives and 384 potential drug analogs predicted from the 'structure based pharmacophoric mapping' of model proteins of different population (SAS, AMR, EUR, AFR and EAS) and with ligand molecules included in the Zinc database. The 384, drug analogs have been further exposed to screening based upon the docking and interaction score values and nine molecules (Fig 5) with the top score values have been included in the Table 4.
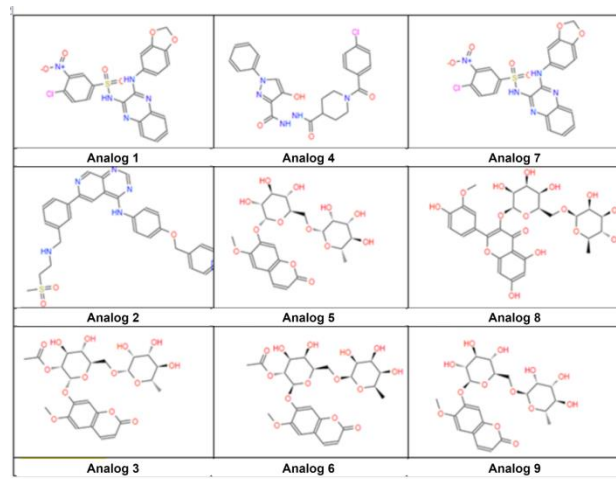

Fig 5. The nine analogs molecules with the top score Analogue

It has been identified that Lapatinib is most suitable for American population and least suitable for South Asian Population. The molecule had good interaction with the model_protein_AMR with an interaction score 45.11 followed by model_protein_EUR (interaction score 28.98), model_protein_AFR (interaction score 18), model_protein_EAS (interaction score 15.98) and model_protein_SAS (interaction score 15.01).

Table 4: The docking score and interaction score of Lapatinib, its derivative and analog molecules with 1xkk and the model proteins

Ligands	IXKK		Model 1		Model 2	
	DS	IS	DS	IS	DS	IS
Lapatinib	*19.09*	*42.80*	*10.11*	*30.25*	*12.96*	*55.21*
Derivative 1	9.83	35.30	34.83	60.78	34.07	61.22
Derivative 2	19.11	43.29	35.34	56.51	36.46	59.15
Derivative 3	19.73	45.01	35.94	63.96	36.17	61.17
Derivative 4	17.89	44.13	36.00	58.90	36.17	61.17
Derivative 5	13.01	38.80	35.94	63.96	35.93	60.47
Derivative 6	13.01	38.80	36.46	59.15	37.33	60.02
Derivative 7	22.62	48.55	35.40	60.00	35.64	56.92
Derivative 8	23.62	46.49	35.51	61.27	34.96	56.29
Derivative 9	24.05	48.31	36.19	61.14	35.83	56.56
Analog 1	35.17	60.74	36.74	63.98	35.17	57.93
Analog 2	35.17	60.74	36.83	59.77	36.83	59.77
Analog 3	37.23	61.17	35.19	56.50	36.83	59.77
Analog 4	35.29	59.29	34.14	60.81	36.83	59.77
Analog 5	35.17	60.74	36.83	59.77	35.25	56.88
Analog 6	35.17	60.74	35.19	56.50	35.25	56.88
Analog 7	36.83	59.77	35.96	60.41	36.74	63.98
Analog 8	35.29	59.29	35.29	59.29	35.17	60.74
Analog 9	35.29	59.29	35.17	60.74	35.17	60.74

for all the complexes were found to be below 0.9suggesting it to be able to retain in the complex for enough time. The molecules were further subjected to ADMETox analysis. The molecules were screened based on 'Lipinski rule of five'. The derivatives and the analogs are found to be better absorbed by the body. However, derivative 8 with IUPAC name (4-{3-Chloro-4-[(m-fluorophenyl)methoxy]phenylamino}-6-(5-{[2-(methylsulfonyl)ethylamino]methyl}-2-furyl)quinazoline) and analog 3 with IUPAC name (4,5-Dihydroxy-2-(6-methoxy-7-coumarinyloxy)-6-[(3,4,5-trihydroxy-6-methyltetra hydro-2H-pyran-2-yloxy)methyl]tetrahydro-2H-pyran-3-yl acetate) are found to be keeping higher solubility and oral absorption (Table 5).

Table 5. The ADMETox prediction of Lapatinib, its derivative and analogueue molecules

Ligands	Absorption	Solubility	cardiotoxicity	hepatotoxicity
Lapatinib	2	1	0.993	0.994
Derivative 1	2	1	0.990	0.993
Derivative 2	2	1	0.989	0.993
Derivative 3	2	1	0.993	0.992
Derivative 4	2	1	0.991	0.992
Derivative 5	2	1	0.991	0.995
Derivative 6	2	1	0.992	0.992
Derivative 7	2	1	0.991	0.992
Derivative 8	3	2	0.991	0.995
Derivative 9	3	2	0.991	0.992
Analogue 1	3	2	0.000	0.982
Analogue 2	3	2	0.981	0.000
Analogue 3	3	3	0.000	0.000
Analogue 4	3	1	0.986	0.000
Analogue 5	3	3	0.988	0.000
Analogue 6	3	1	0.000	0.000
Analogue 7	3	2	0.998	0.000
Analogue 8	3	2	0.995	0.000
Analogue 9	3	1	0.999	0.000

The derivative 7 had a good interaction with EUR, derivative 8 showed good interactions with SAS and EAS, derivative 9 interacts with AFR and the derivative 6 interacts with AMR model proteins. It has been identified that the analogue 3 interacts with SAS and EAS protein models, analogue 5 with EUR model protein, analogue 6 showed a good interaction with AFR protein model and analogue 7 had a good interaction with AMR protein model

3.5 Computational evaluation

UsThe molecules were evaluated by subjecting the ligand-target complexes obtained from interaction analysis to MD simulation. The derivative 7, derivative 8, derivative 9, analogue 3, analogue 5 and analogue 6 are found to be stable during the NVT MD simulation. The average RMSD values

The zero level cardio toxicity and hepatotoxicity suggest analog 3 to be a potential ERBB2 'down regulator' for SAS and EAS populations when compared to other derivative and analogue molecules interacting with AFR, EUR and AMR population models subject to further in-vitro/in-vivo evaluations

4. Discussion

4.1 Genomics

The gene, *ERBB2* was found to be involved in promoting metastatic breast cancer. The germline mutations of *ERBB2* gene; c.338G>A: p.R113Q has been identified as the potential risk factor that may result in breast cancer development and progression [43]. The *HER2* germline mutation A270S is reported to keep a negligible survival among breast cancer patients [44]. The genomic analysis of the HER2 gene and identification of its 'single nucleotide polymorphisms/variations (SNPs/SNVs)' could improve the efficacy of anti-HER2 treatments [45]. Though *Lapatinib* has been identified as the potential drug inhibiting *ERBB2* mutation, its drug action varies among ethnic groups due to individual variations. The clinical trials reports of the drug suggest that, the progression of the disease and death rates are decreasing among US population. However, the drug is not found to be that useful in controlling the disease among SAS populations. The Side effects also differ among patients in both the populations. With the implementation of pharamcogenomic analysis, a population-specific potential personalized drug molecule could be designed. The variations (SNVs) involved in each pharmacogenomic factors have been identified for all the populations.

4.2 Variation analysis

The variations could be classified into different classes based on population analysis. There were 43 variations that were found commonly among all the five populations namely; South Asian (SAS), American (AMR), European (EUR), African (AFR) and East Asian (EAS) according to the '1000 genome population data'. The variations among SAS were further classified into 4 classes based on its frequency of occurrence and their clinical significance. The 'deepCNN based prediction model' helps to assign clinical significance values to all the SNVs. The designed model is found to be giving an accuracy of 72.5%.

4.3 Model generation

Mutant gene models have been generated for each population by incorporating their corresponding variations. These models are further subjected to transcription, translation and homology modelling to generate the corresponding 'target protein models'. Homology modeling has been found to be a coherent technique to develop the 3D structure of unknown protein molecules by our research group. This technique was found to be efficient in generating the remote protein homologies of p-glycoprotein, 'methicillin-Resistant Staphylococcus aureus' of mecA gene and the structure of the 'E1 domain of Q1El92_CHIKV viruses'. [46-49]. In this work, the models structures have been evaluated through molecular dynamic (MD) simulation and by measuring the QMean score, the 'local quality estimate (LQE)',protein mobility score, number of bad angles and bonds as well as finding the percentage of residues present in the favored region (torsional angles - phi (φ)and (psi (ψ)) [50]. The protein models varied within and across the population group both in structure as well as sequence properties. This suggests the importance of considering genetic variations. The physicochemical properties of the protein models have been studied to identify its functional properties. The isoelectric point (PI), negative GRAVY (Grand Average Hydropathy) and relatively high 'aliphatic index values' suggests that the molecules are functionally significant, thermodynamically stable and can survive in hydrophilic environment [51, 52].

4.4 Pharmacophore mapping and interaction analysis

Among molecules obtained from target-based pharmacophore mapping, Lapatinib has been identified with complementary pharmacophore features and anti-breast cancer mechanisms. It was found to be interacting with the amino acid residues present in the kinase domain of the model protein (Fig 4). The derivatives generated through substitution in the kinase domain of the Lapatinib molecule and their pharmacophore analogs were subjected to the docking studies with the CDOCKER algorithm.
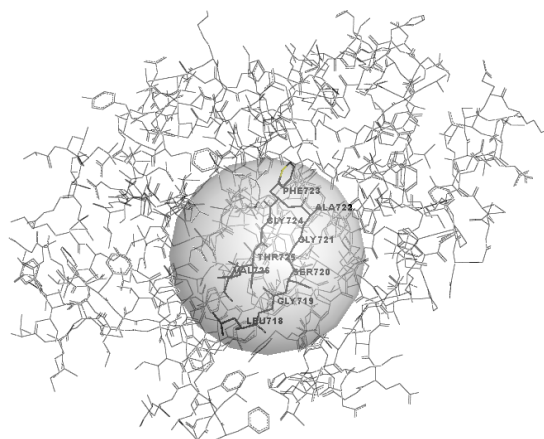

Fig 4. The docking score and interaction score of Lapatinib, its derivative and analogue molecules with 1xkk and the model proteins

4.5 Computational Evaluation

The interaction study and the ADMETox analysis suggest few 'customized potential molecules', more interacting and with higher docking score values with the model proteins. The transport of drug to and from central nervous system is measured through the logarithmic ratio of the compound concentration in brain and blood known as log (BB), stands for the 'Blood Brain Barrier' [53-55].

The comparative analysis between the five populations suggest that, s is having good interaction with mutant protein model for AMR and least interacting with mutant protein model for SAS. The most supporting molecule for South Asian population are derivative-8 ((4-{3-Chloro-4-[(m-fluorophenyl)methoxy]phenylamino}-6-(5-{[2-(methylsulfonyl)ethylamino]methyl}-2-furyl)quinazoline) and the drug analog-3 ((4,5-Dihydroxy-2-(6-methoxy-7-coumarinyloxy)-6-[(3,4,5-trihydroxy-6-methyltetrahydro-2H-pyran-2-yloxy)methyl]tetrahydro-2H-pyran-3-yl acetate)). Moreover, these molecules are found to be more interacting than the standard Lapatinib molecule. The absorption and solubility properties of the molecules suggest them to be acceptable. The toxicity predictions of these molecules support them to be potential drug molecules to down-regulate *ERBB2* mutation associated with metastatic breast cancer, subject to further invitro and invivo evaluations.

All the results and conclusions received in this manuscript are based upon the observations made by 'the Insilco-analysis' and 'pharmacogenomic prediction techniques'. However, all the conclusions can be made only by evaluating the results through in-vivo and experimental techniques. The authors would like to illustrate the importance of 'customized drug designing' especially to terminal diseases. The illustration of the strategy for designing and developing 'predictive, preventive, participatory and personalized (P4) drugs are the primary motive of this work.

5. Conclusion

The population analysis of ERBB2 variations have been carried out and classified the SNVs into different population groups based on the 1000 genome population analysis – American (AMR), European (EUR), South Asian (SAS), African (AFR) and East Asian (EAS). The variants within SAS have been further classified into 4 groups based on their frequency of occurrence among the patients. The population specific mutant gene models have been generated by inducing corresponding group of variations. The 3D structures of the mutant protein have been modelled through transcription, translation and homology modelling. The structure based pharmacophore mapping suggests Lapatinib to be the standard ligand which was further modified to identify suitable derivative and analogs for South Asian population. The computational evaluation suggests the derivative, (4-{3-Chloro-4-[(m-fluorophenyl)methoxy]phenylamino}-6-(5-{[2-(methylsulfonyl) ethylamino]methyl}-2-furyl)quinazoline) and the analogue of *Lapatinib* derivative, (4,5-Dihydroxy-2-(6-methoxy-7-coumarinyloxy)-6-[(3,4,5-trihydroxy-6-methyltetrahydro-2H-pyran-2-yloxy)methyl]tetrahydro-2H-pyran-3-yl acetate have been found to be potential molecule that could down-regulate ERBB2 mutation within breast cancer patients among South Asian population subject to further in-vitro, in-vivo and clinical evaluations.

References

[1] G Nan Jiang, Jing-Jing Lin, Jun Wang, Bei-Ning Zhang, Ao Li, Zheng-Yang Chen, Song Guo, Bin-Bin Li, Yu-Zhong Duan, Ru-Yi Yan, Hong-Feng Yan, Xiao-Yan Fu, Jin-Lian Zhou, He-Ming Yang, and Yan Cui. Novel treatment strategies for patients with HER2-positive breast cancer who do not benefit from current targeted therapy drugs. Exp Ther Med. 2018; 16(3): 2183–2192.

[2] Nevada Cancer Institute. Lapatinib and ixabepilone for the treatment of metastatic breast cancer. Pharmacotherapy. 2018;28(10):1255-66

[3] Frans L. Opdam, Henk-Jan Guchelaar, Jos H. Beijnen and Jan H.M. Schellens. Lapatinib for Advanced or Metastatic Breast Cancer. Oncologist. 2012; 17(4): 536–542.

[4] Beverly Moy and Paul E. Goss.Lapatinib: Current Status and Future Directions in Breast Cancer.The Oncologist.2006;11(10):1047-1057

[5] Sorlie T, Tibshirani R, Parker J, Hastie T, Marron JS, Nobel A, Deng S, Johnsen H, Pesich R, Geisler S, Demeter J, Perou CM, Lonning PE, Brown PO, Borresen-Dale AL, Botstein D. Repeated observation of breast tumor subtypes in independent gene expression datasets. Proc Natl Acad Sci USA. 2003: 100: 8418-8423.

[6] Yarden Y, Sliwkowski MX. Untangling the ErbB signalling network. Nat Rev. 2001; 2:127–137.

[7] Yarden Y. The EGFR family and its ligands in human cancer. signalling mechanisms and therapeutic opportunities. Eur J Cancer. 2001; 37(Suppl 4):S3–S8.

[8] Devika Gajria and Sarat Chandarlapat. HER2-amplified breast cancer: mechanisms of trastuzumab resistance and novel targeted therapies.Expert Rev Anticancer Ther. 2011; 11(2): 263–275.

[9] Wanyi Tai, Rubi Mahato, and Kun Cheng. The role of HER2 in cancer therapy and targeted drug delivery.J Control Release. 2010; 146(3): 264–275.

[10] Namboori PK, Vineeth KV, Rohith V, Hassan I, Sekhar L, Sekhar A, Nidheesh M.. The ApoE gene of Alzheimer's disease (AD). Funct Integr Genomics. 2011; 11(4):519-22.

[11] Lakshmi Anand C, Hima Vyshnavi A M, Gilu Francs, Sanjay Kumar P, Karthikeyan S, P K Krishnan Namboori. Population Wise Variation of Breast and Ovarian Cancer-

A Pharmacogenomic Approach. Materials today: proceedings. 2018; 5(8):16106-16110

[12] Preethi M Iyer, Karthikeyan S, Sanjay Kumar P, Karthikeyan S, P K Krishnan Namboori.Comprehensive strategy for the design of precision drugs and identification of genetic signature behind proneness of the disease-a pharmacogenomic approach. Functional & Integrative Genomics. 2017; 17(4), 375-385.

[13] Yitian Zhou, Souren Mkrtchian, Masaki Kumondai, Masahiro Hiratsuka, Volker M. Lauschke. An optimized prediction framework to assess the functional impactof pharmacogenetic variants.The Pharmacogenomics Journal (2019) 19:115–126

[14] Abu Z Dayem Ullah, Jorge Oscanoa, Jun Wang, Ai Nagano, Nicholas Lemoine, Claude Chelala, SNVnexus: assessing the functional relevance of genetic variation to facilitate the promise of precision medicine, Nucleic Acids Research, 2018, 46(W1):W109-W113

[15] Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. dbSNV: the NCBI database of genetic variation. Nucleic Acids Res. 2001; 29(1):308-11.

[16] Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Research, 2018, 47(D1):D886-D894.

[17] Reva B, Antipin Y, Sander C. Predicting the Functional Impact of Protein Mutations: Application to Cancer Genomics. Nucleic Acids Research (2011); 39(17):e118

[18] Shihab HA, Gough J, Mort M, Cooper DN, Day INM, Gaunt, TR. (2014). Ranking Non-Synonymous Single Nucleotide Polymorphisms based on Disease Concepts. Human Genomics, 8:11(2014).

[19] Víctor Lopez-Ferrando, Andrea Gazzo, Xavier de la Cruz, Modesto Orozco, and Josep L Gelpi. PMut: a web-based tool for the annotation of pathological variants on proteins, 2017 update. Nucleic Acids Res. 2017; 45: W222–W228.

[20] Dassault Systèmes BIOVIA, BIOVIA Workbook, Release 2017; BIOVIA Pipeline Pilot, Release 2017, San Diego: Dassault Systèmes, [2017].

[21] Wishart DS, Feunang YD, Guo AC, Lo EJ1, Marcu A, Grant JR, Sajed T, Johnson D, Li C, Sayeeda Z, Assempour N, Iynkkaran I, Liu Y, Maciejewski A, Gale N, Wilson A, Chin L, Cummings R, Le D, Pon A1, Knox C1, Wilson M1. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res. 2017;46(D1):D1074-D1082.

[22] David Ryan Koes, Carlos J. Camacho. ZINCPharmer: pharmacophore search of the ZINC database. Nucleic Acids Research, 2012; 40(W1), Pages W409–W414.

[23] Abreu RM, Froufe HJ, et al. ChemT, an open-source software for building template-based chemical libraries. SAR QSAR Environ Res. 2011;22(5-6):603-10.

[24] Gasteiger E., Gattiker A., Hoogland C., Ivanyi I., Appel R.D., Bairoch A. ExPASy: the proteomics server for in-depth protein knowledge and analysis. Nucleic Acids Res. 31:3784-3788(2003).

[25] Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F.T., de Beer, T.A.P., Rempfer, C., Bordoli, L., Lepore, R., Schwede, T.

[26] Benkert P, Biasini M, Schwede T. "Toward the estimation of the absolute quality of individual protein structure models." Bioinformatics. 2011; 27(3):343-50.

[27] Bahl A, Singh R, Wadhwa J, et al. Practical consensus recommendations regarding the management of HER2 neu positive early breast cancer. South Asian J Cancer. 2018;7(2):102-105.

[28] Oakman C, Pestrin M, Zafarana E, et al. Role of lapatinib in the first-line treatment of patients with metastatic breast cancer. Cancer Manag Res. 2010; 2:13-25.

[29] Stelzer G, Rosen R, Plaschkes I, et al. The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analysis, Current Protocols in Bioinformatics. 2016;54:1.30.1 - 1.30.33.

[30] Pruitt KD, Katz KS, Sicotte H, et al. Introducing RefSeq and LocusLink: curated human genome resources at the NCBI. Trends Genet. 2000;16(1):44-7.

[31] Irene Papatheodorou, Pablo Moreno, Jonathan Manning, et al. Expression Atlas update: from tissues to single cells. Nucleic Acids Research. 2020;48(D1):D77–D83.

[32] Abu Z Dayem Ullah, Jorge Oscanoa, Jun Wang, et al. SNPnexus: assessing the functional relevance of genetic variation to facilitate the promise of precision medicine, Nucleic Acids Research. 2018;46(W1):W109-W113.

[33] Robert Vaser, Swarnaseetha Adusumalli, Sim Ngak Leng, et al. SIFT missense predictions for genomes. Nat. Protoc. 2009;4:1073–1081.

[34] Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. Nat Methods. 2010; 7(4):248-249.

[35] Philipp Rentzsch, Daniela Witten, Gregory M Cooper, et al. CADD: predicting the deleteriousness of variants throughout the human genome, Nucleic Acids Research. 2019;47(D1):D886–D894.

[36] Boris Reva, Yevgeniy Antipin, Chris Sander. Predicting the functional impact of protein mutations: application to cancer genomics, Nucleic Acids Research. 2011; 39(17):e118.

[37] Shihab HA, Gough J, Cooper DN, et al. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. Hum Mutat. 2013;34(1):57-65.

[38] Mark F Rogers, Hashem A Shihab, Matthew Mort, et al. FATHMM-XF: accurate prediction of pathogenic point mutations via extended features, Bioinformatics. 2018;34(3):511–513.

[39] Torada L, Lorenzon L, Beddis A, et al. ImaGene: a convolutional neural network to quantify natural selection from genomic data. BMC Bioinformatics. 2019; 20:337.

[40] Yang Liu1, Duolin Wang, Fei He, et al. Phenotype Prediction and Genome-Wide Association Study Using Deep Convolutional Neural Network of Soybean.Front.Genet. 2019;10:1091.

[41] Bojian Yin, Marleen Balvert, Rick A A van der Spek, et al. Using the structure of genome data in the design of deep neural networks for predicting amyotrophic lateral

sclerosis from genotype, Bioinformatics. 2019;35(14):i538–i547.

[42] Gasteiger E, Gattiker A, Hoogland C, et al. ExPASy: the proteomics server for in-depth protein knowledge and analysis. Nucleic Acids Res. 2003;31:3784-3788.

[43] Waterhouse A, Bertoni M, Bienert S, et al. SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res. 2018;46(W1):W296-W303.

[44] Studer G, Rempfer C, Waterhouse A M, et al. QMEANDisCo - distance constraints applied on model quality estimation. Bioinformatics. 2020;36:1765-1771.

[45] Ho BK, Thomas A, Brasseur R. Revisiting the Ramachandran plot: hard-sphere repulsion, electrostatics, and H-bonding in the alpha-helix. Protein Sci. 2003;12:2508–2522.

[46] Lovell SC, Davis IW, Arendall WB, et al. Structure validation by Calpha geometry: phi,psi and Cbeta deviation. Proteins. 2003;50:437–450.

[47] Ho BK, Coutsias EA, Seok C, et al. The flexibility in the proline ring couples to the protein backbone. Protein Sci, 2005;14:1011–1018.

[48] Berman HM, Westbrook J, Feng Z, et al.The Protein Data Bank. Nucleic Acids Research. 2000;28:235–242.

[49] Wu G, Robertson DH, Brooks CL 3rd, Vieth M. Detailed analysis of grid-based molecular docking: A case study of CDOCKER-A CHARMm-based MD docking algorithm. J Comput Chem. 2003;24(13):1549-1562.

[50] Anantharaju PG, Reddy BD, Padukudru MA, et al. Naturally occurring benzoic acid derivatives retard cancer cell growth by inhibiting histone deacetylases (HDAC). Cancer Biology & Therapy. 2017;18(7):492-504.

[51] Kumar A, Petri ET, Halmos B, Boggon TJ. Structure and clinical relevance of the epidermal growth factor receptor in human cancer. J Clin Oncol. 2008; 26(10):1742-1751.

[52] Scaltriti M, Verma C, Guzman M. et al. Lapatinib, a HER2 tyrosine kinase inhibitor, induces stabilization and accumulation of HER2 and potentiates trastuzumab-dependent cell cytotoxicity. Oncogene. 2009;28:803–814.

[53] Kancha RK, von Bubnoff N, Bartosch N, et al. Differential Sensitivity of ERBB2 Kinase Domain Mutations towards Lapatinib. PLoS ONE. 2011;6(10):e26760.

First Author: Dr. P. K. Krishnan Namboori currently serves as an Associate Professor at the Center for Excellence in Computational Engineering and Networking (CEN) and the head of the Computational Chemistry Group and Amrita Molecular Modeling and Synthesis (AMMAS) Research Lab, Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Coimbatore campus. He is also an Adjunct Faculty at the School of Biotechnology at Amritapuri Campus, Amrita Vishwa Vidyapeetham. He is presently working on pharmacogenomics and 'design and development of personalized drugs' for diseases such as cancer, life style diseases etc. A number of international journal publications, conference proceedings, four published patents and one international book are to his credit.
ORCID ID: 0000-0003-1296-2621, SCOPUS ID: 57208055836

Second Author: Hima Vyshnavi A. M, after completion of her M. Tech in Computational Engineering and Networking, is currently pursuing her Ph. D in 'design and development of precision drugs for breast cancer move towards P5 pharmacogenomic approach' at AMMAS research lab, Amrita School of Engineering, Amrita Vishwa Vidyapeetham. She has two conference publications and two peer reviewed journal publications to her credit. She is the recipient of the prestigious 'Visvesvaraya Scholarship' of the 'Ministry of Electronics and Information Technology', Government of India. Her areas of interest include Bioinformatics, Cancer Research and Life Sciences.
ORCID ID: 0000-0003-3628-3922, SCOPUS ID: 57207305367

Author Contribution:

Dr. P K Krishnan Namboori: Concepts, Guidance, Design of the work, Manuscript review

Hima Vyshnavi A M: Implementation, Literature search, Data acquisition, Manuscript preparation and Manuscript editing