CARDIO-PRED: an *in silico* tool for predicting cardiovascular-disorder associated proteins

Prerna Jain¹ · Nitin Thukral¹ · Lokesh Kumar Gahlot¹ · Yasha Hasija¹

Received: 4 March 2015 / Accepted: 6 March 2015 / Published online: 14 March 2015
© Springer Science+Business Media Dordrecht 2015

Abstract Interactions between proteins largely govern cellular processes and this has led to numerous efforts culminating in enormous information related to the proteins, their interactions and the function which is determined by their interactions. The main concern of the present study is to present interface analysis of cardiovascular-disorder (CVD) related proteins to shed lights on details of interactions and to emphasize the importance of using structures in network studies. This study combines the network-centred approach with three dimensional studies to comprehend the fundamentals of biology. Interface properties were used as descriptors to classify the CVD associated proteins and non-CVD associated proteins. Machine learning algorithm was used to generate a classifier based on the training set which was then used to predict potential CVD related proteins from a set of polymorphic proteins which are not known to be involved in any disease. Among several classifying algorithms applied to generate models, best performance was achieved using Random Forest with an accuracy of 69.5 %. The tool named CARDIO-PRED, based on the prediction model is present at <http://www.genomeinformatics.dce.edu/CARDIO-PRED/>. The predicted CVD related proteins may not be the causing factor of particular disease but can be

involved in pathways and reactions yet unknown to us thus permitting a more rational analysis of disease mechanism. Study of their interactions with other proteins can significantly improve our understanding of the molecular mechanism of diseases.

Keywords Interface properties · Cardiovascular-disorder · Machine learning · 3-Dimensional structure · Network analysis

Introduction

The non-modifiable major risk factors associated with CVD include advancing age, genetic predisposition, gender and ethnicity, with aging being the major risk factor. Despite of numerous studies in the cardiovascular field many unanswered questions remain as to how the genetic pathways that regulate aging in model organisms influence cardiovascular aging. A number of longevity genes and their related pathways identified in model organisms have been assessed for their role in CVD (North and Sinclair 2012). The association between genes and diseases has been well studied in the past and has been documented in various databases such as Online Mendelian Inheritance in Man (OMIM, <http://omim.org/>), Genetic Association Database, Disease ontology and many more. These databases not only provide genes associated with the disorders but also, insight about the common genes or essential genes which are involved in pathways and processes of multiple disorders. The human disease gene network has already established link between the genetic disorders with the genes. An important conclusion inferred from this network is that genes associated with similar disorders show higher likelihood of physical interaction

Electronic supplementary material The online version of this article (doi:10.1007/s11693-015-9164-z) contains supplementary material, which is available to authorized users.

✉ Yasha Hasija
yashahasija@gmail.com

¹ Department of Biotechnology, Delhi Technological University, Shahbad Daultpur, Main Bawana Road, Delhi 110042, India

between their products, thus forming a hub of essential genes and their products (Goh et al. 2007).

Documentation of genes with the associated disease is not enough to understand the biological details of pathogenesis and disease progression. It is important to identify various molecules and mechanism triggering, participating and controlling the biological process. This understanding of molecular mechanism is a complex process and not much is known about the mechanism (Gonzalez and Kann 2012). Now with the tremendous increase in human interaction data, it is important to understand how any biological function is exerted over the body and this can only be made possible by the inclusion of structural details in the networks.

Undoubtedly, sequence based annotation is important in unravelling the encoded information but the finer details of the molecular mechanism within the cell is possible only with the structural information (Marini and Brandi 2010). Protein structures provide a higher resolution of information and a more sensitive approach for detecting similarities among proteins by including details about structures of the interacting proteins in the network, protein hubs and protein interfaces. Structural profile of the proteins provides the opportunities in understanding the cellular functioning in terms of structural scaffolds which facilitate the underlying molecular recognition events. Moreover, a protein's structure is more conserved than its amino acid sequence (Choura and Rebaï 2011).

Protein–protein interactions (PPI) network is a way of representing how two or more proteins interact with each other in a cell, and biological processes, are essentially interactions between multiple proteins (Zhang and Ouellette 2011) with PPI networks controlling the flow of information both within and between the cells. Network representation where proteins are nodes and interactions are edges is useful indicator of biological process and protein function (Choura and Rebaï 2011). It is well known that proteins are the main agents of biological function based on their interactions with other proteins and to other biomolecules such as DNA and RNA, mediating metabolic and signalling pathways and cellular processes. Protein interaction do not necessarily mean the physical association of proteins but can be metabolic interactions where proteins are involved in the same pathway or genetic interactions displaying co-expressed and co-regulated proteins (Gonzalez and Kann 2012). Any protein rarely acts alone but as mentioned above interacts with other protein to carry out a specific function. This pair of interacting protein forms a protein complex. The protein fragments within the complex that actually comes together and interact forms a special region called protein interface (Aung 2006). Thus, protein interaction networks are useful resources to gain knowledge about the evolution of proteins

and about the different systems in which they are involved. Alteration in this network can potentially lead to various disorders because disorders are often caused by alteration in the binding sites or allosteric changes in the protein. For example, mutation in the zinc finger domain present in the oncoprotein MDM2 can disrupt the interaction of MDM2 with L5 AND L1 (ribosomal proteins) that mediates p53 degradation and leads to cancer (Zhang and Ouellette 2011). Moreover, in protein–protein network, most of the proteins have few interactions whereas some proteins may have multiple interactions, these protein are central to the stability and normal functioning of the proteins in the network for example p53, p21, p27, BRCA1, ubiquitin, calmodulin are extensively involved in different forms of cancer, and the product of these genes forms a hub. Deletion of these hub proteins, is lethal to the organism, hence it can be inferred that these proteins are encoded by essential genes and are important targets for molecular and structural studies (Choura and Rebaï 2011). In some cases, during the binding process, interactions between two proteins might involve order–disorder transitions in partially disordered regions of the interacting proteins. These unstructured or disordered regions have been found to be involved in many disease mechanisms. For instance, the cancer suppressor BRCA1 has been shown to contain intrinsically disordered regions through which it binds to several proteins (Kann 2007). Other than the interference in protein–protein interaction network, several diseases are caused due to disruption in protein–DNA interaction, protein folding mechanism, protein–RNA disruption, can enable pathogen host protein interaction or can lead to new undesired protein interactions (Gonzalez and Kann 2012).

Protein interactions are mediated by specific recognition of distinct binding regions on the surface of interacting protein. Such recognition should be of sufficient affinity to effectively bind fragments of proteins, which is decided by specificities such as interfaces. Thus, interface properties and the topology of the protein involved play an important role in deciphering their function in the network. Various studies have been carried to integrate PPI network with the structural properties. A significant work in this regard has been done in the field of cancer biology. The cancer genes were mapped onto human PPI network and the network was studied with respect to structural features such as interface properties and study showed successful results in classifying cancer and non-cancer proteins (Kar et al. 2009). A novel integrated approach was proposed named CAERUS, for identification of gene signatures to predict cancer outcomes based on domain interaction network in human proteome (Zhang and Ouellette 2011). Further, a method was developed to find sub network-based signatures by incorporating PPI networks and gene expression profiles to classify metastatic and non-metastatic tumors

(Chuang et al. 2007). Studies have shown that a molecular mechanism leading to diseases is dependent on interlinked proteins and networks, but we are far from unravelling the difference in molecular interactions in healthy and diseased organism. Mechanism of genetic malfunctioning, that leads to one or several diseases can be understood when the molecular level of the protein interactions are known. Understanding the interactions and hence binding between proteins is critical for the rational design of new therapeutic agents targeted to disrupt the interactions that cause disorders (Gonzalez and Kann 2012).

PPIs not only identify disease-associated interacting proteins but also the potential disease-associated gene candidates (genes coding for interacting proteins are putative disease causing genes). This theory has been used in the present work where all the protein associated with CVD have been retrieved and a classifier using machine learning algorithms is used to identify potential proteins which are associated with CVDs. This classifier is based on the interface properties of the interacting complexes, to differentiate CVD related proteins from non-CVD associated proteins.

Methods

The detailed protocol for the study has been mentioned below and is depicted as a flow chart in Fig. 1.

Building datasets

OMIM (OMIM, <http://omim.org/>) Morbid Map was used to obtain training and test sample. The training sample: consisted of two subsets; one set consisting of all the proteins associated with the cardiovascular disorders and the other set consisting of all the protein associated with diseases other than CVDs. There are a total of 124 diseases in the class of cardiovascular disorders. For each disease, concerned proteins were listed. Test sample lists polymorphic proteins which are not known to be involved in any disease. Proteins were retrieved from humsavar.txt–UniProt. It is text file describing human polymorphism and disease mutation and contains disease variants, polymorphisms and unclassified variants. In the present study, we used protein variants under polymorphism as test set (Magrane 2011).

Retrieval of the interacting proteins

Wang et al. found high quality binary PPI along with the atomic resolution interface. They combined reliable literature-curated binary interactions and well verified yeast two-hybrid interactions to produce human protein

interaction network (Wang et al. 2012). Further, they structurally resolved the interfaces of these interactions using iPfam and 3did utilizing homology modelling approaches. This resulted in a human structural interaction network (hSIN). This information is stored in the database with the gene id and Pfam id. hSIN was used to determine the structurally relevant interactions to the proteins involved in the cardiovascular interactions. Gene ids for the disease associated proteins were retrieved from NCBI, and the interacting partners with reliable structure were obtained from hSIN.

Structural interaction network

Once the structurally interacting partners have been obtained, a structure interaction network was formed by using Interactome3D. It is a web service for the structural annotation of protein–protein interaction networks (Mosca et al. 2013). PDB codes with the interacting chains between proteins were retrieved.

Interface property analysis

For interface analysis, 2P2I inspector (a protein–protein interface analysis tool) was used that invokes VMD, NACCESS and SURFNET. It is a complete tool that computes the interaction properties from 3D structure of interacting complexes. The various descriptors provided by this tool includes accessible surface area (ASA), Gap Volume, percentage charged residues, secondary structure contribution, number of hydrogen bonds, number of salt bridges and number of disulphide bonds. All these parameters are important in determining the specificity and the strength of the interface. Input can be a PDB code or the PDB file along with the chains involved in the complex (Basse et al. 2013).

Balancing and building the random training dataset

Before running Weka (Hall et al. 2009), Cygwin tool was used, which is a popular GNU development tools for Microsoft Windows. With Cygwin, it is possible to easily port many Unix programs without the need for extensive changes to the source code. With Cygwin, a series of random files were created by combining the positive and negative data set. These files were then used in Weka for classification and finding the prediction probability of test or unknown dataset.

Classification analysis: algorithm and analysis

Comparison between the proteins involved in CVDs and those which are not known to be involved in these diseases

were made using interface features which describe the protein complexes. By using machine learning algorithm, an automatic classifier capable of identifying genes more likely to be involved in CVDs based on the interfacial patterns was made. Different algorithms were tested and parameters such as true positive, true negative, false positive, false negative, precision, recall and accuracy were used to determine the best algorithm. Random forest with iteration 30 was used to generate the model for prediction. Tens runs of 10-fold cross validation were performed on the classifier algorithm used. The model generated was evaluated and used to predict the likelihood of polymorphic proteins associated with CVDs.

Development of user interface

For the prediction of involvement of unknown proteins in CAD, user-friendly web interfaces was developed. XHTML and CSS was used for creating presentation layer of CARDIO-PRED, the web server used was Apache/2.4.4 (Win 32) Open SSL/0.9.8y PHP/5.4.19. For backhand, database server version 5.5.32 MySQL community server (GPL) was used. It is hosted at <http://www.genomeinformatics.dce.edu/CARDIO-PRED/>. The screenshots from the tool are as shown in the Fig. 2.

Results

Proteins associated with cardiovascular disorders

A total of 124 diseases under cardiovascular class of disorder were extracted from OMIM Morbid Map. Proteins associated with the specific diseases were identified from UniProt KB. For each protein interactions which are structurally relevant i.e. in which both the interacting partners have Pfam id, were considered, out of a total of 124 proteins, only 47 interactions were found to be structurally relevant (Table 1 in supplementary material). For diseases other than the CVD, a total of 1200 interactions were studied and only 400 interactions were found to be structurally relevant. In the test set, almost 1500 polymorphic proteins were studied for interactions out of which only 600 interactions were structurally relevant.

Structural interaction and interface analysis network

A network with the PDB structures mapped onto interacting proteins was obtained from Interactome3D (Fig. 3).

Fig. 1 Flow chart of the methodology carried out

Fig. 2 Overview of CARDIO-PRED web page interface (a) a screenshot of home page of CARDIO-PRED, (b) the Tool page of CARDIO-PRED representing the attributes—interface area, % charged residues, gap volume index, no. of hydrogen bonds, gap

volume, secondary structure at interface, disulphide bonds, no. of salt bridges, (c) an example: using “FGG” and selecting different attributes to be displayed, (d) result page showing the output for “FGG”

This network was generated separately for both training and the test set. A complete analysis of network helped in elucidating individual structures of proteins as well as the protein structures of interacting complexes. Along with the structures, this tool provides a detailed description of chains from both the proteins responsible for complex formation, which forms an important part of our study (Fig. 4).

A total of eight descriptors were obtained from 2P2I inspector namely accessible surface area, gap volume, gap volume index, % charged residues, number of hydrogen bonds, number of salt bridges, number of disulphide bonds and secondary structure at interface (Tables 2, 3 and 4 in supplementary materials). The parameters of interfaces for analysing interactions in the present study are as follows:

- *Accessible surface area (ASA)* this is the mean ASA buried by each domain of the complex.
- *Complementarity of interface—gap volume* provides a measure of complementarity and closeness of packing of the interface between two interacting proteins by measuring the volume of empty space between them (Kar et al. 2009).
- *Gap volume index* the gap volume index between two protein domains is calculated as ratio of gap volume to the interface area. It estimates the volume enclosed between any two molecules, delimiting the boundary by defining a maximum allowed distance from both interface (Kar et al. 2009).
- *Planarity* the planarity of interfaces analyzes the shape of the interface. It is defined as the rmsd of the interface

Fig. 3 Network chart of PPI with structures mapped on the network

atoms from the least-squares plane fitted through all interface atoms (Kar et al. 2009).

- *Hydrogen bonding* hydrogen bonds play key role in specificity of interactions between two proteins. The average number of hydrogen bonds is proportional to the area of subunit surfaces i.e. ASA: one bond per each 100–200 Å² (Moreira et al. 2007).
- *Number of salt bridges* the more the number of salt bridges, there will be more electrostatic interactions resulting in tighter packing within the residues.
- These descriptors were used to from the training and test sample. These files were used as input in Weka.

Classification and prediction analysis

A set of famous algorithms such as Naïve Bayes, Random Forest, J48, Random Forest with iteration 30, Bagging with Random Forest and Multilayer Perceptron were run on the dataset and comparisons were done between them (Table 1). Various statistical values were used in determining the optimum classifier for our training sample. In Fig. 5, true positive rate, true negative rate, precision, recall and accuracy-all were highest for Random Forest with

iterations 30 (i.e. the number of trees were increased from 10 to 30). In the columns of false positive and false negative, Random Forest_30 scores the least. Thus, Random Forest_30 was found to be the best classifier in our dataset for the model building.

Prediction probability for test set

Predictions were made using the test sample to predict potential CVD related proteins from a set of polymorphic proteins. Further analysis of these may provide insight into yet unknown molecular mechanism various CVDs. List of all the polymorphic protein along with their prediction probability was obtained. The predicted proteins are not explicitly involved in causing CVD but may be involved in various pathways and mechanisms associated with CVDs. Table 2 shows the proteins with prediction value more than 0.8. A deep analysis of these proteins both computationally and experimentally can unravel the mechanisms not known to us. Further, the proteins listed in the table may be used to study gene prioritization by considering the proteins with the highest prediction probability as having the likelihood of being associated with cardiovascular disorders.

Fig. 4 Interaction between two proteins along with the interface residues

Table 1 Comparisons between different classifiers

Classifier	True positive rate	False positive rate	True negative rate	False negative rate	Precision	Recall	Accuracy (%)
Random Forest_10	0.77	0.4	0.59	0.23	0.67	0.77	68.34
Naïve Bayes	0.6	0.41	0.59	0.4	0.6	0.56	59.5
J48	0.74	0.39	0.61	0.26	0.67	0.74	67.5
Random Forest_30	0.77	0.38	0.62	0.23	0.7	0.77	69.5
Multilayer perceptron	0.71	0.41	0.6	0.28	0.65	0.71	65.5
Bagging_Random Forest	0.75	0.37	0.63	0.25	0.68	0.6	69

Discussion

Main concern of the present study was to present interface analysis of CVD related proteins to shed light on details of interactions and to emphasize the importance of using structures in network studies (Kar et al. 2009). The results obtained showed that interface properties can be used as descriptors for classifying purposes; be it two different diseases, a set of disease having same phenotypic effect from other set of disease with different phenotype, or for all the known hereditary diseases. This study provided insights into the usage of fine structural details in disorders related to cardiovascular system. A further exploration of this study can be utilized to study whether the classifier

could work equally efficient when used for classifying all the inherited disease.

These results can be used as evidence when searching for candidate gene in predefined disease loci and also for predictions, to identify novel genes involved in the mechanisms. Studies have been carried out where sequence features such as cDNA, number of exons, protein size were investigated between set of genes known to be involved in hereditary disease and those not (López-Bigas and Ouzounis 2004; Adie et al. 2006). Functional annotation shared between known diseases genes had also been used to design a classifier (Rao et al. 2012). The classification based on genomic sequences have disadvantages associated with them, as algorithms which are based on functional

Fig. 5 Bar graph with different classifiers and the statistical values

annotations are inherently biased towards a particular known subset of genes. The present study utilized human PPI network with structural properties to compare genes involved in diseases. It is assumed that structural properties provide a much wider coverage of the understanding of mechanism of disease than sequence analysis. A promising approach for disease gene discovery is to combine all the evidences available such as PPIs, structures, sequences, transcriptional expression, functional annotation to make a classifier that would be possible to predict accurately without any ambiguities (Calvo et al. 2006).

Protein–protein interaction network provide a vast area of study in which different criteria, different features can be utilized to devise new diagnostic tool or to understand a novel mechanism. Further, new prognostic tools can be created by identification of pathways or disease sub networks that get activated only in diseases states. For instance, a recent study integrated protein networks with cancer expression profile and identified pathway that get activated only during tumor progression discriminating metastasis better than earlier described markers (Gonzalez and Kann 2012).

Drug target identification and drug design can be made more comprehensive by including disease networks and structural details, for example, structural information on allosteric site or binding site can be used to design potential drug to affect protein function. Further, rebuilding different

Table 2 Proteins predicted as potentially associated with cardiovascular disorders and their associations described

Protein 1	Protein 2	PDB ID	Prediction probability
DSG1	DSG1	3IFQ, AC	0.924
ABI1	EPS8	1GRI, AB	0.894
BCAR1	BCAR1	1OV3, AB	0.894
CASP1	CASP1	3E4C, AB	0.892
CASP1	CASP1	2H4 W, AB	0.88
ACTN2	ACTN2	1HCI, AB	0.877
CASP1	CASP1	1ICE, AB	0.868
CASP1	CASP1	1RWK, AB	0.868
CASP1	CASP1	1RWO, AB	0.868
CASP1	CASP1	1RWV, AB	0.868
CASP1	CASP1	1SC3, AB	0.868
A2 M	CELA1	3HS0, AB	0.861
ACY1	ACY1	2ZOG, AB	0.861
CAND1	CAND1	1U6G, AB	0.861
FARSA	FARSB	3L4G, AB	0.861
EGFL6	EGFL6	2BO2, AB	0.859
FGG	FGG	1FZA, ab	0.858
CASP1	CASP1	2FQQ, AB	0.846
ELANE	ELANE	1EZX, AB	0.844
CASP1	CASP1	1IBC, AB	0.841
BHMT	BHMT	1LT8, AB	0.838
CASP1	CASP1	1RWM, AB	0.827
CASP1	CASP1	1SC1, AB	0.827
DCP2	DCP2	2DSO, AB	0.827
DHPS	DHPS	1ROZ, AB	0.827
DHPS	DHPS	1RQD, AB	0.827
BCR	BCR	2B3R, AB	0.826
BHMT	BHMT	1LT7, AB	0.824
FGG	FGG	2FFD, AB	0.813
ACVR1	ACVR1	3O96, AB	0.811
CASP1	CASP1	1BMQ, AB	0.81
ABL1	ABL1	2HYY, AB	0.809
FBP1	FBP1	3KC1, AB	0.809
FBP1	FBP1	1FTA, AB	0.809
ABLIM1	ABLIM1	2DFY, CX	0.806
BCL6	BCL6	1R2B, AC	0.806
BCL6	BCL6	3LBZ, AB	0.806
ABL1	ABL1	3CS9, AB	0.803
BAIAP2	BAIAP2	1Y2O, AB	0.802
FBP1	FBP1	2VT5, AB	0.8
BAIAP2	BAIAP2	1WDZ, AB	0.799
DOCK9	CDC42	2WM9, AB	0.794
FGG	FGG	3GHG, AB	0.794
ACY1	ACY1	1Q7L, AB	0.792
ABI1	ABI1	3P8C, AD	0.791
ADA	ADA	1WII, AB	0.79
ABL1	OSBPL5	3JV3, AB	0.782

Table 2 continued

Protein 1	Protein 2	PDB ID	Prediction probability
BIRC2	BIRC2	1I3O, AB	0.771
ADAM17	TIMP3	3CKI, AB	0.769
BDNF	BDNF	3BUK, AC	0.765
FBP1	FBP1	2JJK, AB	0.763
FECH	FECH	3HCN, AB	0.762
EMILIN1	EMILIN1	2KA3, AB	0.761
FBP1	FBP1	2WBB, AB	0.761
BFAR	BFAR	3HCU, AB	0.751
DUSP7	DUSP7	2GWO, AB	0.751
FECH	FECH	2HRC, AB	0.751
BCL6	BCL6	3E4U, AB	0.749
BLMH	BLMH	1CB5, AB	0.746
FGG	FGG	1LTJ, AB	0.743
ABCB8	ABCB8	2HYD, AB	0.742
FBP1	FBP1	3A29, AB	0.742
FBP1	FBP1	3KBZ, AB	0.742
FGG	FGG	3BVH, AB	0.742
EEA1	EEA1	1 JOC, AB	0.741
ELANE	ELANE	3KCG, HI	0.74
ELANE	ELANE	1OPH, AB	0.736
ELANE	ELANE	3F1S, AB	0.731
ELANE	ELANE	3F1S, AB	0.731
FASN	AASDHPPT	2JFK, AB	0.723
DR1	DR1	1N1J, AB	0.722
BCR	BCR	1K1F, AB	0.714
DOCK9	CDC42	2WMN, BA	0.707
ACVR1	UBXN1	2HAK, AB	0.706
BAG5	BAG5	3A8Y, AC	0.701
ABL2	ERBB3	2HCK, AB	0.697
ACAP1	ACAP1	3JUE, AB	0.693
ABI3	ABI3	2FPD, AB	0.691
BARD1	BARD1	3FA2, AB	0.689
ADH7	ADH7	1D1S, AB	0.687
FASLG	FAS	1D0G, AB	0.687
DNM1	DNM1	2X2E, AD	0.685
CADM1	CADM1	35C8, LH	0.683
ABL1	EPB41L3	2PSQ, AB	0.682
FGG	FGG	1LT9, AB	0.671
FGG	FGA	3HUS, AB	0.671
FGG	FGG	1RF0, AB	0.669
AANAT	YWHAZ	1IB1, AE	0.665
FGF9	FGFR3	1NUN, AB	0.665
ELANE	ELANE	1JMO, HA	0.661
ADH7	ADH7	1AGN, AB	0.654
ADH5	ADH5	1M6H, AB	0.653
ADH5	ADH5	1MC5, AB	0.653
ADH5	ADH5	1MP0, AB	0.653

Table 2 continued

Protein 1	Protein 2	PDB ID	Prediction probability
ABCB8	ABCB8	2IXF, AB	0.65
ACLY	ACLY	3MWE, AB	0.65
ADH5	ADH5	2FZE, AB	0.65
ADORA2A	DRD2	3ODU, AB	0.65
FASN	FASN	3HHD, AB	0.65
ECHS1	ECHS1	2HW5, EF	0.647
BDNF	BDNF	1WWW, VX	0.635
BFAR	BFAR	3EB6, AB	0.634
ACTB	ACTB	1HLU, AP	0.627
ADH1C	ADH1C	1HT0, AB	0.624
CAPN1	CAPN1	2ARY, AB	0.623
ADH5	ADH5	1M6 W, AB	0.62
ACLY	ACLY	3MWD, AB	0.618
ECHS1	ECHS1	2VRE, AB	0.618
FECH	FECH	1HRK, AB	0.618
ABL1	GRB2	3M7F, AB	0.612
BIRC2	BIRC2	3M1D, AB	0.612
ABL2	NCK1	2C0O, AB	0.608
DCTN1	DCTN1	2HKN, ab	0.602
FGG	FGG	2OYH, AB	0.599
BARD1	BARD1	2R1Z, AB	0.598
ABR	ABR	1BWN, AB	0.594
CADM1	CADM1	1S78, AB	0.592
FGG	FGG	3H32, AB	0.592
ACP6	ACP6	1CVI, AB	0.591
ADH1C	ADH1C	1U3W, AB	0.591
BIRC2	BIRC2	3MUP, AB	0.591
ABI1	EPS8	3EG1, AB	0.59
ADH5	ADH5	2FZW, AB	0.587
ADH5	ADH5	1MA0, AB	0.587
ABCD1	ABCD1	1R0Y, AB	0.582
ABR	ABL1	1DYN, AB	0.581
DNM1	DNM1	1DYN, AB	0.581
ELANE	ELANE	1H1B, AB	0.581
FBP2	FBP2	3IFA, AB	0.572
FASN	AASDHPPT	2PX6, AB	0.56
ADAM15	ADAM15	3G5C, AB	0.559
DCTN1	MAPRE1	2HKQ, AB	0.554
FASLG	TNFRSF6B	2HEY, FR	0.553
FGG	FGG	3E1I, AB	0.553
ABCD1	ABCD1	1G29, 1,2	0.552
EGLN3	EGLN3	3HQR, AS	0.545
ACVR2A	ACVR2A	3Q4T, AB	0.544
FAS	FADD	3EZQ, AB	0.544
ABCB8	ABCB8	1XF9, AB	0.541
FAU	FAU	2ZNV, AB	0.538
ABI1	NCKAP1	2V1Q, AB	0.527

Table 2 continued

Protein 1	Protein 2	PDB ID	Prediction probability
BAK1	BAK1	1BXL, AB	0.527
ABL1	NFASC	2IZV, AC	0.521
DDB1	CUL4A	2HYE, CA	0.519
BIRC2	BIRC2	3OZ1, AB	0.511
ABCD1	ABCD1	3C4 J, AB	0.51
ABLM1	ABLM1	2XJZ, AI	0.508
BCR	BCR	3BYI, AB	0.508
ACVR2A	ACVR2A	2QJA, AC	0.507
FGG	FGG	1FZB, AC	0.507
DDB1	DDB2	3EI4, FE	0.505
ARTN	GFRA3	2GYR, AB	0.504
DDX18	DDX18	3LY5, AB	0.504
DDX24	DDX24	3LY5, AB	0.504
ACVRL1	ACVRL1	3MY0, AM	0.5
CALM3	CALM3	2VAY, AB	0.5
CARD10	CARD10	3YGS, CP	0.499
EGFR	EGF	1NQL, AB	0.499
ADCYAP1	ADCYAP1R1	2JOD, AB	0.482
CALM3	CALM3	2X0G, AB	0.476
FASN	AASDHPPT	2JFD, AB	0.475
DCTN1	MAPRE1	1TXQ, AB	0.471
ABL2	EGFR	2K7A, AB	0.467
DVL3	DVL3	2UZC, AB	0.458
ABL2	ABL2	1D1Z, AB	0.457
ABCG2	ABCG2	2PZE, AB	0.455
BAX	bcl2l11	2K7 W, AB	0.454
ADAP1	KIF13B	3MDB, AC	0.451
CALM3	CALM3	3OXQ, AE	0.45
FASN	FASN	1XKT, AB	0.449
ACVR1	CDK14	2V55, AB	0.44
ABL1	PRICKLE3	2G2H, AB	0.439
BMP1	BMP1	1NZI, AB	0.432
ADAP1	KIF13B	3FM8, AC	0.429
BDNF	NTF3	1BND, AB	0.426
ELANE	ELANE	2R9P, EA	0.423
FBXW8	SKP1	1FQV, AB	0.415
ECHS1	ECHS1	2ZQR, AB	0.414
FGG	FGG	1RE3, AB	0.413
DVL3	DVL3	2PNT, AB	0.412
EGFL6	EGFL6	3GIS, EF	0.412
ARTN	ARTN	2GH0, CA	0.41
ABCD1	ABCD1	1Q1B, AB	0.405
BFAR	UBE2D1	3HCT, AB	0.403
DUSP7	DUSP7	2C46, CD	0.401
ABII	EPS8	1JU5, AC	0.399
ACVR2A	ACVR2A	3KFD, AE	0.393
BARD1	BARD1	2NTE, AB	0.391

Table 2 continued

Protein 1	Protein 2	PDB ID	Prediction probability
ABCB8	ABCB8	3C41, JK	0.39
DVL3	DVL3	1N99, AB	0.386
ACVRL1	ACVRL1	3C4E, AB	0.384
BIRC2	BIRC2	2NZ7, AB	0.383
ACVR2B	ACVR2B	1NYS, AB	0.382
ECHS1	ECHS1	1HZD, AB	0.381
ABCB8	ABCB8	3G61, AB	0.376
FGG	FGG	1FIC, AB	0.374
ACVRL1	ACVRL1	3GGF, AB	0.373
BAD	BCL2L1	1G5 J, AB	0.37
BCL2A1	BCL2A1	3PK1, AB	0.37
ABII	NCKAP1	2O88, AB	0.366
ACD	ACD	2I46, AB	0.364
BMP1	BMP1	1SZB, AB	0.359
ELANE	ELANE	1SR5, AC	0.359
BCL6	BCL6	1R28, AB	0.355
BIRC2	BIRC2	1XB0, AB	0.355
ASAP1	APC	2RQU, AB	0.351
FGG	FGG	2H43, AC	0.349
ACTB	ACTN3	3LUE, AK	0.347
BCL2A1	BCL2A1	1ZY3, AB	0.347
BCL2A1	BCL2A1	2VM6, AB	0.347
BCL6	BCL6	3BIM, AI	0.343
ADCYAP1	ADCYAP1R1	3IOL, AB	0.34
BCAR1	BCAR1	1WYX, AB	0.338
DCTN1	DCTN1	3E2U, GA	0.335
ACP5	ACP5	1FJM, AM	0.333
ACPP	ACPP	1ND6, AB	0.331
FASN	AASDHPPT	2CG5, AB	0.331
ADAM17	ADAM17	1ZXC, AB	0.328
ABL1	ASB16	2E2B, AB	0.327
DDX24	DDX24	3FE2, AB	0.324
ABL2	GRB3	2XYN, AB	0.322
ABR	ABR	1B55, AB	0.321
EHMT1	TP53	1YCS, AB	0.318
BAX	BAX	2P1L, AB	0.317
DVL3	DVL3	3GGE, BC	0.315
ABL1	ABL1	2HZI, AB	0.314
ACVR1	NKIRAS1	2H96, AF	0.314
BDNF	BDNF	1HCF, AX	0.314
ABL2	ERBB2	2D0 N, CD	0.31
ACVR1B	ACVR1B	2QJ9, AC	0.309
ELANE	ELANE	1EAW, AB	0.309
FBXW8	SKP1	1NEX, AB	0.308
ACVR1	FKBP1A	3H9R, AB	0.307
ACVRL1	ACVR2A	3H9R, AB	0.307
CAPN1	CAPN1	2NQA, AD	0.305

Table 2 continued

Protein 1	Protein 2	PDB ID	Prediction probability
BCL2	BCL2	2XA0, AB	0.294
ABL2	PLCG1	1CWE, CD	0.291
FGG	FGG	2A45, AB	0.291
ACVR1	CHN1	3MDY, AB	0.289
FBXW7	SKP1	2OVP, AB	0.284
ABL2	SRC	3N84, AG	0.283
edn1	edn1	1T7H, AB	0.283
DCTN1	MAPRE1	2HL5, AC	0.279
ACHE	ACHE	3LII, AB	0.277
DUSP7	DUSP7	1RXD, CA	0.276
FAU	FAU	3ALB, AB	0.275
BIRC2	BIRC2	3D9T, AB	0.273
ABR	ABL1	1PBW, AB	0.262
ABI1	ABL1	3SEM, AB	0.26
ABL1	HCK	1ABO, AB	0.258
ACVR1B	1NHBB	1REW, AC	0.255
BLMH	BLMH	2CB5, AB	0.253
FGG	FGG	1FZF, AB	0.245
BCL2	BCL2	2 W3L, AB	0.241
ABI1	ABI1	1JQQ, AB	0.229
ASAP1	APC	2W2W, AB	0.227
ACHE	ACHE	1VZI, AI	0.22
DVL3	DVL3	1OBY, PA	0.22
ABI1	ABL1	1ZUY, AB	0.214
ABL1	DNM1	1U5S, AB	0.207
BIRC2	BIRC2	1TW6, AC	0.206
FGF10	FGFR3	1RY7, AB	0.197
ABCB8	ABCB8	1G9X, AB	0.19
ABCG5	ABCG5	1L7 V, AC	0.189
FBXW8	SKP1	2ASS, AB	0.189
ECT2	ECT2	3L46, AB	0.188
ABL1	SOS2	3KFA, AB	0.187
ABL1	ERBB4	1OPL, AB	0.187
FES	FES	3CBL, AB	0.187
BFAR	UBE2D3	1FBV, AB	0.186
ABI1	ABI1	1BBZ, AD	0.185
EGFR	EGF	1IVO, AC	0.182
CADM1	CADM1	1Q1 J, LH	0.173
ABCG2	ABCG2	2AWN, AB	0.165
ACY1	ACY1	1CG2, AB	0.163
ADORA2A	ADORA2A	3KJ6, AL	0.162
DCTN1	DCTN1	2HL3, AC	0.16
ACPP	ACPP	2HPA, AB	0.155
ACVRL1	ACVR2A	1B6C, AB	0.153
ABL1	ASAP1	1GCQ, AC	0.15
EML1	EML1	1P22, AB	0.149
DVL3	DVL3	3KZE, CE	0.142

Table 2 continued

Protein 1	Protein 2	PDB ID	Prediction probability
DCTN1	DCTN1	2HQH, EA	0.14
ABCD1	ABCD1	2IT1, AB	0.133
ACCS	ACCS	1AHG, AC	0.132
ABL1	ABL1	2G1T, AE	0.122
FECH	FECH	2QD1, AB	0.121
ABCB8	ABCB8	2BBS, AB	0.075
ABL1	ERBB3	2SHP, AB	0.075
ADH5	ADH5	1TEH, AB	0.075
ELK3	ELK3	1HBX, AG	0.075
BARD1	BARD1	3C5R, AB	0.065

interaction networks such as metabolic, signalling etc. prediction about hub protein can be made which are involved in various pathways for proper functioning of the cell. This knowledge can be used while designing the drugs, i.e. whether to target hub proteins or not.

The prediction of CVD associated genes is of great importance to understand the mechanism behind PPI. A lot of research efforts have gone into classifying proteins either as disease associated or not associated with a disease. Several bioinformatics strategies have been explored to develop an efficient model for this classification. Machine learning tools have been applied using different classifiers and different kinds of features/attributes based on protein sequence, structure and binding affinities. Such approaches in machine learning were used to develop the most optimal model by exploring the contribution of different attributes to interaction.

Some studies have shown the association between age-related CVDs and osteoporosis with common etiology such as increased risk of hip fracture in women. Recent studies have shown that drugs have common effect for CVDs and osteoporosis, for example bone anti-resorptive drug reduces the risk of CVDs. Similarly, positive effects of statins, antihypertensive drugs have been shown on bone mass. These studies points to the common physiopathological molecular pathways between both the diseases. Statins for example, reduces cardiovascular mortality through reduction of LDL cholesterol levels, these have also been associated to increase bone mineralization in mice and reduction of fractures. All these examples suggest a link between the vascular and skeletal systems; therefore it is the main concern to understand the exact physiopathological mechanism shared between both the diseases as well as to determine common risk factors and genetic determinants (Marini and Brandi 2010).

Conclusion

The present study aimed at identifying and distinguishing proteins based on the interface parameters of the structure involved as CVD associated and CVD non-associated. It started with genes associated with disease to the structures associated with diseases. A classifier using Random Forest with 30 iterations was used for training based on the training set which included interface properties of proteins (with their interacting partners) associated with CVDs and proteins (with their interacting partners) associated with diseases other than CVDs.

To best of our knowledge, there are no prediction servers available based on interface properties of protein structure for predicting CVD associated proteins. In the present study, we presented a Random forest based method for prediction of CVD associated proteins with the accuracy of 69.5 %. The predicted proteins/genes may not be the causing factor of particular disease but can be involved in various pathways and mechanism yet unknown to us. Study of their interactions with other proteins can significantly improve our understanding in the molecular mechanism of diseases. Generally, mutations in proteins affect the structure of protein and hence it is becoming important to shift the focus from genetic studies to molecular studies. The classifying method has been implemented to develop a web based tool, CARDIO-PRED, which is available online at <http://www.genomeinformatics.dce.edu/CARDIO-PRED>.

Acknowledgments The study has been supported by SERB under OYS scheme via Project File No. SR/FT/LS-96/2011.

Conflict of interest The author(s) declare that they have no competing interests.

References

- Adie EA, Adams RR, Evans KL, Porteous DJ, Pickard BS (2006) SUSPECTS: enabling fast and effective prioritization of positional candidates. *Bioinformatics* 22(6):773–774
- Aung Z (2006) Computational analysis of 3D protein structures. Doctoral dissertation, School of Computing, National Institute of Singapore
- Basse MJ, Betzi S, Bourgeois R, Bouzidi S, Chetrit B, Hamon V, Morelli X, Roche P (2013) 2P2Idb: a structural database dedicated to orthosteric modulation of protein–protein interactions. *Nucleic Acids Res* 41:D824–D827
- Calvo S, Jain M, Xie X, Sheth SA, Chang B, Goldberger OA, Spinazzola A, Zeviani M, Carr SA, Mootha VK (2006) Systematic identification of human mitochondrial disease genes through integrative genomics. *Nat Genet* 38(5):576–582. doi:10.1038/ng1776
- Choura M, Rebaï A (2011) Structural analysis of hubs in human NR-RTK network. *Biol Direct* 6:49. doi:10.1186/1745-6150-6-49
- Chuang HY, Lee E, Liu YT, Lee D, Ideker T (2007) Network-based classification of breast cancer metastasis. *Mol Syst Biol* 3:140
- Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabási AL (2007) The human disease network. *Proc Natl Acad Sci* 104(21):8685–8690
- Gonzalez MW, Kann MG (2012) Protein interactions and disease. *PLoS Comput Biol* 8(12):e1002819. doi:10.1371/journal.pcbi.1002819
- Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH (2009) The WEKA data mining software: an update SIGKDD explorations 11(1)
- Kann MG (2007) Protein interactions and disease: computational approaches to uncover the etiology of diseases. *Brief Bioinform* 8(5):333–346
- Kar G, Gursay A, Keskin O (2009) Human cancer protein–protein interaction network: a structural perspective. *PLoS Comput Biol*. doi:10.1371/journal.pcbi.1000601
- López-Bigas N, Ouzounis CA (2004) Genome-wide identification of genes likely to be involved in human genetic disease. *Nucl Acids Res* 32(10):3108–3114
- Magrane M, The UniProt consortium (2011) UniProt knowledgebase: a hub of integrated protein data
- Marini F, Brandi ML (2010) Genetic determinants of osteoporosis: common bases to cardiovascular diseases. *Int J Hypertens*. doi:10.4061/2010/394579
- Moreira IS, Fernandes PA, Ramos MJ (2007) Hot spots—a review of the protein–protein interface determinant amino-acid residues. *Proteins* 68(4):803–812
- Mosca R, Céol A, Aloy P (2013) Interactome3D: adding structural details to protein networks. *Nat Methods* 10(1):47–53
- North BJ, Sinclair DA (2012) The intersection between aging and cardiovascular disease. *Circ Res* 110:1097–1108
- Online Mendelian Inheritance in Man, OMIM®. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD). <http://www.omim.org/>. Accessed April 2014
- Rao A, Bulusu G, Srinivasan R, Joseph T (2012) Protein–protein interactions and disease, protein interactions. Dr. Jianfeng Cai (Ed.), ISBN: 978-953-51-0244-1, InTech
- Wang X, Wei X, Thijssen B, Das J, Lipkin SM, Yu H (2012) Three-dimensional reconstruction of protein networks provides insight into human genetic disease. *Nat Biotechnol* 30(2):159–164. doi:10.1038/nbt.2106
- Zhang KX, Ouellette BF (2011) CAERUS: predicting CAnCER oUtcomeS using relationship between protein structural information, protein networks, gene expression data, and mutation data. *PLoS Comput Biol*. doi:10.1371/journal.pcbi.1001114