Structural assessment of the effects of Amino Acid Substitutions on protein stability and protein–protein interaction

Shaolei Teng

Department of Genetics and Biochemistry,
Clemson University,
Clemson, SC 29634, USA
E-mail: steng@clemson.edu

Anand K. Srivastava and Charles E. Schwartz

J.C. Self Research Institute of Human Genetics,
Greenwood Genetic Center,
Greenwood, SC 29646, USA
E-mail: anand@ggc.org
E-mail: ceschwartz@ggc.org

Emil Alexov

Department of Physics and Astronomy,
Clemson University,
Clemson, SC 29634, USA
E-mail: ealexov@clemson.edu

Liangjiang Wang*

Department of Genetics and Biochemistry,
Clemson University,
Clemson, SC 29634, USA
E-mail: liangjw@clemson.edu
Fax: +1 864 388 1707
*Corresponding author

Abstract: A structure-based approach is described for predicting the effects of amino acid substitutions on protein function. Structures were predicted using a homology modelling method. Folding and binding energy differences between wild-type and mutant structures were computed to quantitatively assess the effects of amino acid substitutions on protein stability and protein–protein interaction, respectively. We demonstrated that pathogenic mutations at the interaction interface could affect binding energy and destabilise protein complex, whereas mutations at the non-interface might reduce folding energy and destabilise monomer structure. The results suggest that the structure-based analysis can provide useful information for understanding the molecular mechanisms of diseases.

Copyright © 2010 Inderscience Enterprises Ltd.

Eli Lilly & Co.
Exhibit 2067
BPI Labs v. Eli Lilly
IPR2025-01346
Ex. 2067-001

Keywords: AASs; amino acid substitutions; homology modelling; folding energy; binding energy; protein stability; protein–protein interaction.

Reference to this paper should be made as follows: Teng, S., Srivastava, A.K., Schwartz, C.E., Alexov, E. and Wang, L. (2010) 'Structural assessment of the effects of Amino Acid Substitutions on protein stability and protein–protein interaction', *Int. J. Computational Biology and Drug Design*, Vol. 3, No. 4, pp.334–349.

Biographical notes: Shaolei Teng is currently a PhD student in the Department of Genetics and Biochemistry at Clemson University. He received his MS in Molecular Biology from Gyeongsang National University in South Korea, and his BS in Biotechnology from Soochow University in China. His current research interests include machine learning and structure modelling. He has published seven papers in peer-reviewed journals and conference proceedings.

Anand K. Srivastava is a Senior Research Scientist and Associate Director of the Center for Molecular Studies, J.C. Self Research Institute of Human Genetics, Greenwood Genetic Center and Professor (adjunct) in the Department of Genetics and Biochemistry, Clemson University. He received his MS in Biochemistry from the University of Allahabad, and PhD in Biochemistry from Banaras Hindu University, India. His research interest is to understand the etiology of inherited disorders of human brain development, cognitive function and epithelial organogenesis. He has published more than 65 peer-reviewed papers including reviews and book chapters.

Charles E. Schwartz is the Director of Research at the Greenwood Genetic Center and Head of the J.C. Self Research Institute. He received his PhD in Biochemistry from Vanderbilt University and his MS in Biochemistry from Oklahoma State University. For over 25 years his research has focused on the identification of genes that cause intellectual disability. He has published over 250 peer-reviewed articles and has been a reviewer for many journals.

Emil Alexov got his MS in Physics in 1984 and in 1990 received PhD in Plasma Physics from Sofia University, Bulgaria. Since 1990 he was an Assistant Professor at the Department of General Physics at Sofia University. In 1995 he came to USA as a Research Associate at the Department of Physics, City College of New York and later in 2000 joined Columbia University, New York, as Howard Hughes Bioinformatics Specialist. In 2005, he became Associate Professor at the Department of Physics, Clemson University. His research is focused on further development of DelPhi and modelling disease-causing missense mutations.

Liangjiang Wang is an Assistant Professor in the Department of Genetics and Biochemistry at Clemson University, and an adjunct faculty member at Greenwood Genetic Center. He received his PhD in Molecular Biology from the University of Georgia, and his MS in Computer Science from Mississippi State University. His research focuses on biological knowledge discovery and genomic data integration. He has published more than 40 peer-reviewed papers in journals and conference proceedings. He has been in the editorial board for three international journals, and a reviewer for more than 20 journals.

1 Introduction

Revealing the effects of Amino Acid Substitutions (AASs) on protein structure and function is critical for understanding the complex mechanisms of human disease caused by single amino acid mutations. There are 67,000–200,000 non-synonymous Single Nucleotide Polymorphisms (nsSNPs) in the human population (Cargill et al., 1999), which give rise to a large number of AASs in proteins. The residue changes at key sites within a protein may result in a series of conformation changes, including the breakage of salt bridges, alteration of interaction network, disruption of hydrogen bonds, which in turn may perturb the energy landscape. These changes can affect the kinetics of protein folding or cause protein aggregation and destabilisation (Dill et al., 1993). More than half of the monogenic diseases are caused by single mutations, and a common mechanism by which AASs cause human disease is protein stability change. Yue and Moult (2006) investigated the effect of AASs on protein stability, and estimated that approximately 25% of nsSNPs in the human population might be deleterious to protein function. Of the known disease-causing missense mutations, the majority (83%) resulted in alternation of protein stability (Wang and Moult, 2001).

AASs can also affect protein–protein interactions. Approximately 88% of disease-associated nsSNPs are found to be located in the voids/pockets important for protein-protein interactions (Stitziel et al., 2003). The AASs located at the binding interface or active site cleft could block the entrance to the active site, change the recognition, alter the specificity, or affect the binding affinity. For example, the substitution G2019S in leucine-rich repeat kinase 2 (LRRK2) was shown to be associated with familial and sporadic Parkinson's disease (Aasly et al., 2005). Structure analysis indicates that this mutation is located at the interface of LRRK2's N-terminal and C-terminal domains, which is important for positioning of Mg^{2+} within the active site of the kinase (Albrecht, 2005; Mata et al., 2006). This finding is in agreement with the experimental result that G2019S enhances kinase activity *in vitro* (Kachergus et al., 2005). Recently, Teng et al. (2009) examined the effects of nsSNPs at the interaction interfaces of 264 protein complexes using a homology modelling method and all atoms energy calculations. The results suggest that disease-causing mutations tend to destabilise protein–protein interactions. Therefore, understanding how AASs affect protein stability and protein–protein interactions can provide new insights into the molecular mechanisms of human genetic diseases.

Protein structure modelling methods have been widely used for predicting the effects of disease-causing mutations on protein stability and protein–protein interaction. For instance, to predict the effects of the mutations related to the genetic disorder galactosemia, more than 100 mutant structures of galactose-1-phosphate uridyltransferase were constructed using the homology modelling method, and the results suggested that most mutations might alter protein stability (Facchiano and Marabotti, 2009). By mapping disease-causing mutations onto known three-dimensional protein structures, Dimmic et al. (2005) have shown that about 70% of the deleterious mutations are located in the structurally or functionally important sites. However, the effects of mutations were analysed statically in these studies. The Free Energy Perturbation (FEP) calculation has been used to quantitatively assess the effects of AASs on protein stability. Dixit et al. (2009) used the AMBER force field

and solvent-accessible surface area solvation methods to calculate the protein stability changes in terms of free energy differences caused by cancer-associated mutations in the RET and MET kinases, and showed that the AASs could decrease the thermodynamical stability of the mutant structures. The FEP calculation was also used to assess the protein stability changes upon single AASs in membrane proteins (Park and Lee, 2005). Nevertheless, these studies on FEP calculation did not take into account the effects of AASs on protein–protein interactions.

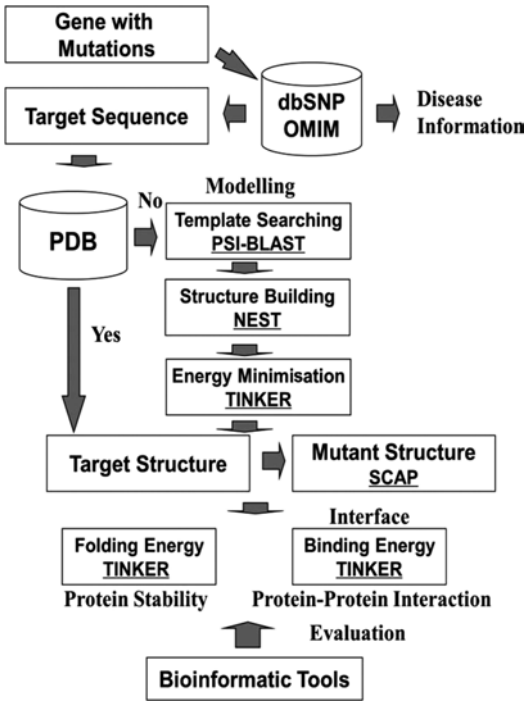
The advent of high-throughput sequencing technology makes it possible to identify a large number of nsSNPs in the human genome. The dbSNP database, one of the primary data resources for genetic studies, contains the information of more than 23 million human SNPs (Smigielski et al., 2000). The records in the dbSNP database are linked to the Online Mendelian Inheritance in Man (OMIM) database, which contains disease gene information, including genetic polymorphisms, map locations, inheritance patterns and clinical descriptions (Wheeler et al., 2007). Computational analyses provide an efficient way for examining the effects of nsSNPs on protein stability and function, and for identifying potential disease-causing mutations. Ng and Henikoff (2003) used a position-specific scoring matrix (PSSM)-based method called Sorting Intolerant From Tolerant (SIFT) to predict whether an AAS affects protein function. We have recently developed the MuStab web server for predicting protein stability changes upon AASs from sequence features (Teng et al., 2010). MuStab uses a Support Vector Machine (SVM) model to discriminate between destabilising and stabilising AASs in proteins. iPTREE-STAB (Huang et al., 2007) and I-Mutant 3.0 sequence version (Capriotti et al., 2008) are also available for sequence-based prediction of protein stability changes caused by point mutations. Structure-based methods, including PoPMuSiC-2.0 (Dehouck et al., 2009), Dmutant (Zhou and Zhou, 2002), Eris (Yin et al., 2007), I-Mutant 3.0 structure version (Capriotti et al., 2008) and FoldX (Schymkowitz et al., 2005), are available for examining the effects of mutations on protein stability and protein–protein interactions. In particular, the FoldX software tool can be used to provide quantitative estimations about the effects of AASs on the stability of proteins or protein complexes using the empirical force field calculation (Schymkowitz et al., 2005). Among these protein stability predictors, I-Mutant3.0 structure version, Dmutant and FoldX gave the best predictive performances (Khan and Vihinen, 2010).

The experimental approach for determining the effects of AASs on protein stability is to obtain the mutant proteins and measure their thermal stability changes by melting experiments. However, the experimental approach is time-consuming and thus may not be applied to a large number of AASs. In this study, a structure-based approach was performed for predicting the effects of AASs on protein stability and protein–protein interaction. The differences of folding energy and binding energy between the WT and mutant structures were calculated to predict the protein stability and protein–protein interaction changes caused by the mutations. The predictions were evaluated by using other bioinformatic methods. The results suggest that the structure-based approach can provide useful information for characterising disease-causing mutations in human genetic studies.

2 Methods

The schematic diagram of the structure-based approach is shown in Figure 1. The methodology was also investigated in two previous studies (Teng et al., 2008; Zhang et al., 2010). For a specific gene with mutations, the related sequence and disease information were extracted from the dbSNP and OMIM databases. If the structure of the target protein was available in the Protein Data Bank (PDB), no structure modelling was needed. Otherwise, target structures were constructed using the homology modelling method (Xiang, 2006). The suitable templates were identified in the PDB database using the PSI-BLAST program (Altschul et al., 1997), and then used to construct the target structures with the NEST program (Petrey et al., 2003). Energy minimisation was performed to obtain the optimal structure with the TINKER program (Ponder, 1999), and the mutant structure was constructed using the SCAP program (Xiang and Honig, 2001). The folding energy of the WT or mutant structure was calculated using TINKER to estimate the effects of the mutations on protein stability. For AASs located at the interface, the binding energy changes were also computed to predict the effects of the mutations on protein–protein interaction. At the end, the predictions were compared with several bioinformatics tools, including FoldX (Schymkowitz et al., 2005), PoPMuSiC-2.0 (Dehouck et al., 2009), Dmutant (Zhou and Zhou, 2002), Eris (Yin et al., 2007), MuStab (Teng et al., 2010), iPTREE-STAB (Huang et al., 2007) and I-Mutant 3.0 (both sequence and structure versions) (Capriotti et al., 2008).

Figure 1 Schematic diagram of the approach for assessing the effects of Amino Acid Substitutions on protein stability and protein-protein interaction. Underlined are the software tools used in this study



2.1 Protein structure modelling

Homology modelling was applied to the proteins with no structures available in the PDB database. The structures were modelled as follows:

- *Template searching*. The suitable templates were selected from the PDB database for the target protein. Position-Specific Iterated BLAST (PSI-BLAST) (Altschul et al., 1997) was used for the template searching. The structures with significant *E*-value ($<10^{-5}$) were selected as the suitable templates.
- *Structure building*. The program NEST was used to build structure models according to the sequence alignment between the target protein and its structural template (Petrey et al., 2003). NEST is an integrated model-building program, including the program LOOPY9 for loop prediction and SCAP10 for side-chain modelling.
- *Energy minimisation*. To generate the optimal structure, energy minimisation was performed by using the TINKER package (Ponder, 1999) with the CHARMM27 force field parameters (Brooks et al., 1983). The MINIMIZE program in TINKER was used to minimise structures with the algorithm of Limited Memory BFGS Quasi-Newton Optimisation (Ponder, 1999).

The mutant structures were derived *in silico* from the WT structure using the SCAP program (Xiang and Honig, 2001). The AASs were introduced by side-chain replacements with the rest of the structure kept rigid. The MINIMIZE program in the TINKER package was used to minimise the mutant structures.

2.2 Folding energy calculation

The effects of AASs on protein stability were assessed by the folding energy changes. The energy calculation was based on the monomer structure of the target protein, and was performed as described in the recent publication (Zhang et al., 2010). The folding energy is the energy difference between the folded and unfolded states:

$$\Delta G(folding) = G(folded) - G(unfolded) \tag{1}$$

where $G(folded)$ or $G(unfolded)$ is the total potential energy of the target protein in the folded or unfolded state, respectively.

The protein stability change ($\Delta\Delta G_{stability}$) is the folding energy difference between the WT structure and the structure with the AAS. It can be calculated using the following equation:

$$\begin{aligned}\Delta\Delta G_{stability} &= \Delta G(folding : WT) - \Delta G(folding : AAS) \\ &= [G(folded : WT) - G(folded : AAS)] - [G(unfolded : WT) \\ &\quad - G(unfolded : AAS)]\end{aligned} \tag{2}$$

However, the energy difference between the WT and mutant proteins in the unfolded state, $G(unfolded: WT) - G(unfolded: AAS)$, is difficult to calculate. In this study, we assume that the difference of energy in the unfolded state can be estimated by using the substitution site and its neighbouring residues. The total potential energy of the eleven-residue segment ($S11$) with the substitution site in the middle position was used to

represent the folding energy of the full-length protein in the unfolded state. Equation (2) can thus be rewritten as:

$$\Delta\Delta G_{stability} = [G(folded : WT) - G(folded : AAS)] - [G(folded : WT_{s11}) - G(folded : AAS_{s11})]. \quad (3)$$

All of the above-mentioned total potential energy terms were calculated using the ANALYZE program in the TINKER package. A positive value of $\Delta\Delta G_{stability}$ indicates that the AAS may make the protein more stable, whereas a negative value of $\Delta\Delta G_{stability}$ suggests that the mutation can destabilise the protein.

2.3 Binding energy calculation

For an AAS located at the interaction interface, the binding energy difference of the protein complex between the WT and mutant structures was used to assess the effect of the mutation on protein–protein interaction. As described in the previous study (Teng et al., 2009), the binding energy was calculated using the rigid body approach, in which the structures of the monomers were kept as they were in the dimer complex. The binding energy, $\Delta\Delta G(binding)$, was the difference between the total potential energy of the dimer complex and the individual monomers:

$$\Delta\Delta G(binding) = \Delta G(folding : complex) - \Delta G(folding : A) - \Delta G(folding : B) \quad (4)$$

where ΔG (*folding: complex*), ΔG (*folding: A*) and ΔG (*folding: B*) are the folding free energy values of the dimer complex, monomer A and monomer B, respectively. Since the internal mechanical energy values of the unbound and bound monomers are the same, the energy terms in the unfolded state can be cancelled out in equation (4). Thus, the binding free energy can be calculated as follows:

$$\Delta\Delta G(binding) = G(folding : complex) - G(folding : A) - G(folding : B) \quad (5)$$

where G(*folded: complex*), G(*folded*) and G(*folded*) are the total potential energy values of the dimer complex, monomer A and monomer B in the folded state, respectively.

In this study, the total potential energy was computed using the ANALYZE program in the TINKER package. The effect of an AAS on protein–protein interaction was assessed by using the binding energy difference between the WT structure and the structure with the AAS:

$$\Delta\Delta\Delta G(binding) = \Delta\Delta G(binding : WT) - \Delta\Delta G(binding : AAS). \quad (6)$$

A positive value of the binding energy change ($\Delta\Delta\Delta G_{binding}$) indicates that the AAS may strengthen the binding affinity and make the protein dimer complex more stable. In contrast, a negative value of $\Delta\Delta\Delta G_{binding}$ suggests that the mutation can weaken the binding affinity and destabilise the dimer complex.

2.4 Prediction evaluation

Several bioinformatic tools were used to evaluate the predictive power of the structure-based approach used in this study, and the predictions were considered to be reliable if a consensus was reached by most of the predictors. Sequence-based

prediction of the direction of protein stability change could give useful information. Three sequence-based tools were used to predict the directions of protein stability changes caused by AASs from primary sequence data, including iPTREE-STAB (http://210.60.98.19/IPTREEr/iptree.htm), MuStab (http://bioinfo.ggc.org/mustab/) and I-Mutant3.0 (sequence version) (http://gpcr2.biocomp.unibo.it/cgi/predictors/I-Mutant3.0/I-Mutant3.0.cgi).

Structure-based prediction methods could provide quantitative assessment of the effects of AASs on protein stability. Khan and Vihinen (2010) compared the predictive performances of different protein stability predictors, and showed that three structure-based tools, including I-Mutant3.0 (structure version) (http://gpcr2.biocomp.unibo.it/cgi/predictors/I-Mutant3.0/I-Mutant3.0.cgi), Dmutant (http://sparks.informatics.iupui.edu/hzhou/mutation.html) and FoldX (Schymkowitz, et al., 2005) were the most reliable predictors. These three tools were used in this study. Other two structure-based predictors, including PoPMuSiC-2.0 (http://babylone.ulb.ac.be/popmusic/) and Eris (http://eris.dokhlab.org), were also used to calculate the folding energy for monomer structures, respectively. The difference of the folding energy between the WT and mutant structures was used to assess the protein stability change caused by an AAS, and compared with the $\Delta\Delta G_{stability}$ value calculated using the approach applied in this paper. Furthermore, FoldX was also used to determine the interaction energy of complex protein. The effect of an AAS on protein–protein interaction was estimated by the interaction energy difference of the protein complex between the WT and mutant structures ($\Delta\Delta\Delta G_{FoldX}$), which was compared with $\Delta\Delta\Delta G_{binding}$ computed using the method utilised in this study.

In addition, ClustalX (Larkin et al., 2007) was used to perform the multiple sequence alignment for conservation analysis. Protein sequences from different species were downloaded from the NCBI Entrez database using the GENE search option with the gene name as the query.

3 Results and discussion

To evaluate the usefulness of the structure-based approach utilised in this paper, three case studies were carried out for four pathogenic mutations and one neutral nsSNP in three human genes (Table 1). One disease-causing mutation, A111V (dbSNP ID: rs28928889, OMIM ID: 141850.0029), and one neutral nsSNP, T119N (dbSNP ID: rs1058069), in the human *HBA2* gene (haemoglobin subunit alpha) were used to show their different effects on protein stability and protein–protein interaction. Two pathogenic mutations, Q61K (dbSNP ID: rs28933406, OMIM ID: 190020.0002) and A146T (dbSNP ID: rs104894231, OMIM ID: 190020.0008), in the human *HRAS* gene (v-Ha-ras Harvey rat sarcoma viral oncogene homologue) were analysed to assess the effects of mutations on different structural regions (interface or non-interface). The computational approach was also used to investigate the substitution, A693V, in the human *ZBTB20* gene (zinc finger and BTB domain containing 20). As discussed in the following sections, the results suggest that the pathogenic mutations make the monomer structures less stable ($\Delta\Delta G_{stability} < 0$), or weaken the binding affinity to destabilise the dimer structures ($\Delta\Delta\Delta G_{binding} < 0$). In contrast, the neutral nsSNP has only slight effects on protein stability and protein–protein interaction ($\Delta\Delta G_{stability}$ and $\Delta\Delta\Delta G_{binding}$ close to 0).

It was shown that the predictions agree well with the results given by most of the structure-based methods. However, the sequence-based tools often did not agree with the consensus predictions from the structure-based methods (Table 1). The structure-based predictors (I-Mutant3.0 structure version, Dmutant and FoldX) appeared to be more reliable for predicting protein stability changes caused by mutations (Khan and Vihinen, 2010). Thus, this study focused on the structure-based analyses.

Table 1 The effects of five amino acid substitutions on protein stability. The unit of energy change is kcal/mol

Amino Acid Substitution		*A111V* (*HBA2*)	*T119N* (*HBA2*)	*Q61K* (*HRAS*)	*A146T* (*HRAS*)	*A693V* (*ZBTB20*)
Structure-based tools	$\Delta\Delta G_{stability}$	–0.75	0.06	–4.42	–1.39	–2.69
	FoldX	–4.19	10.54	–2.74	–0.22	–0.68
	PoPMuSiC-2.0	–0.49	–0.50	–0.24	–0.38	–0.05
	Dmutant	–0.48	0.32	–0.38	–0.24	–0.34
	Eris	4.28	2.29	–2.63	–1.24	–0.72
	I-Mutant 3.0 (structure version)	0.13	–0.30	0.29	–0.79	–0.03
	Consensus	Decreased	Increased	Decreased	Decreased	Decreased
Sequence-based tools	I-Mutant 3.0 (sequence version)	Increased	Increased	Increased	Decreased	Increased
	MuStab	Increased	Decreased	Increased	Decreased	Increased
	iPTREE-STAB	Increased	Decreased	Increased	Decreased	Decreased
	Consensus	Increased	Decreased	Increased	Decreased	Increased

3.1 Pathogenic mutation and neutral nsSNP in haemoglobin

Haemoglobin molecules in red blood cells transport oxygen from the lung to the peripheral tissues, and thus are important for maintaining cell viability. Human haemoglobin is made up of symmetric dimers of polypeptide chains, the α/β-globin dimers (Kan, 1991). Several point mutations in α-globin have been shown to cause α-thalassemia, which can result in Hydrops fetalis (Chui and Waye, 1998). In this study, the two AASs of human haemoglobin subunit alpha (HBA2), A111V and T119N, were analysed to show the different effects of disease-causing and neutral AASs on protein stability and protein–protein interaction. The homodimer structure of HBA2 was built using the crystal structure of human deoxy haemoglobin (PDB: 1O1P) as the template.

The majority of disease-causing mutations cause protein destabilisation, whereas most neutral nsSNPs have limited effect on protein stability (Wang and Moult, 2001). In this study, the predicted effects of A111V (disease-causing) and T119N (neutral) on protein stability agree well with the previous observations. As shown in Table 1, the folding energy change ($\Delta\Delta G_{stability}$) caused by A111V is –0.75 kcal/mol, suggesting that the mutation may destabilise haemoglobin monomer structure. The decreased protein stability is also predicted for the A111V mutation by three structure-based tools including FoldX, PoPMuSiC-2.0 and Dmutant (Table 1). In contrast, the neutral nsSNP (T119N) is

predicted by our calculations and three structure-based tools (FoldX, Dmutant and Eris) to stabilise the protein monomer. PoPMuSiC-2.0 and I-Mutant3.0 (structure version) give the opposite predictions. The results suggest that T119N may not cause destabilisation of the monomer structure.

AASs at the interaction interface may result in binding affinity changes, and thus affect the structure of the protein complex. As shown in Figure 2(A), the pathogenic mutation, A111V, is located in the α-helix of the HBA2 binding interface. Although most regions of the WT and mutant structures are similar, the structures are not overlapped in the α-helix interface region. This structural change may significantly affect the binding energy, and make the protein complex unstable. The observation has been confirmed by the binding energy calculation using both TINKER and FoldX ($\Delta\Delta\Delta G_{binding} = -11.56$ kcal/mol and $\Delta\Delta\Delta G_{FoldX} = -1.41$ kcal/mol) (Table 2). In contrast, the neutral nsSNP (T119N) is located in the flexible loop region (Figure 2(B)). Since T119N is not located in the inner region of the interface, it may not significantly affect protein–protein interaction. The binding energy change caused by T119N is $\Delta\Delta\Delta G_{binding} = 0.90$ kcal/mol (Table 2), which is smaller than the absolute value of binding energy change caused by A111V.

In addition, the multiple sequence alignment shown in Figure 2(C) suggests that the residue, Ala 111, is well conserved, but Thr 119 is not conserved in *Xenopus laevis* and *Xenopus tropicalis*. The result agrees with the previous observation that pathogenic mutations tend to be located at evolutionarily conserved positions (Miller and Kumar, 2001).

Figure 2 Illustration of two amino acid substitutions (A111V and T119N) in human HBA2. (A) Structural representation of the A111V mutation. The Wild-Type chain A is shown in green colour, mutant chain A in cyan, Wild-Type chain B in yellow, and mutant chain B in orange. The amino acid residue Ala 110 (Wild-Type) is shown in magenta, and Val 110 (mutant) in white; (B) Structural representation of T119N. Chains A and B are shown in green and yellow, respectively. The residue Asn 119 (Wild-Type) is shown in pink, and Thr 119 (mutant) in blue and (C) Multiple sequence alignment of HBA2 with the amino acid substitution sites indicated (see online version for colours)

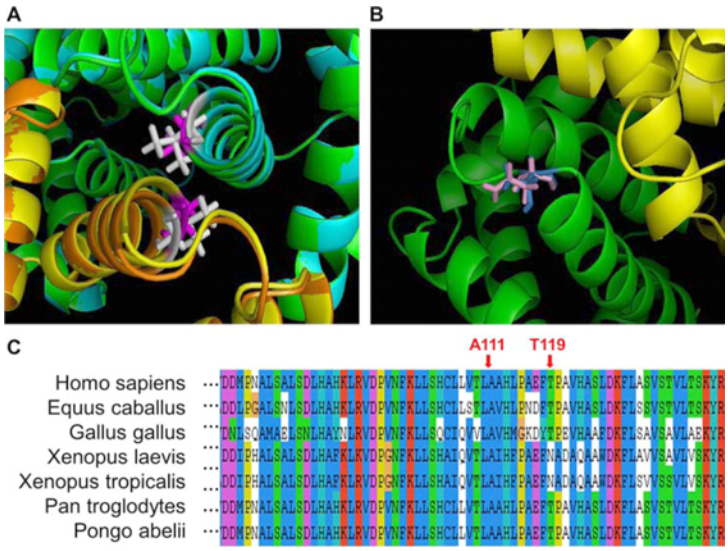


Table 2 The effects of five amino acid substitutions on protein-protein interaction. The unit of energy change is kcal/mol

Amino Acid Substitution	*A111V* (*HBA2*)	*T119N* (*HBA2*)	*Q61K* (*HRAS*)	*A146T* (*HRAS*)	*A693V* (*ZBTB20*)
$\Delta\Delta\Delta G_{binding}$	–11.56	0.90	–7.29	0.21	–0.31
$\Delta\Delta\Delta G_{FoldX}$	–1.41	1.39	–2.40	–0.11	0.00

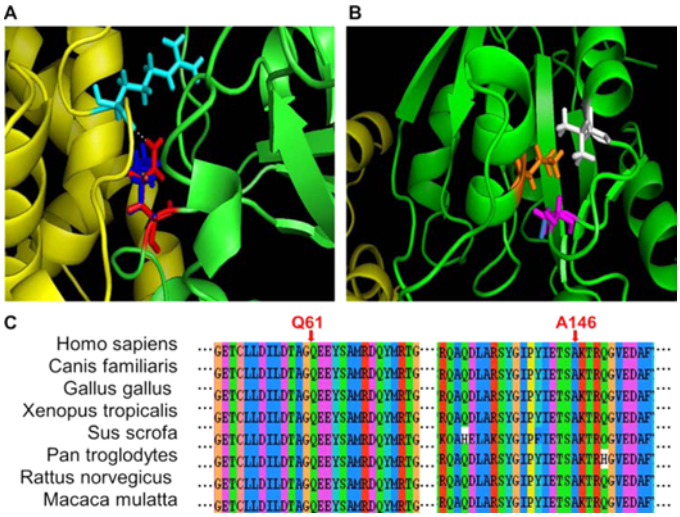
3.2 Pathogenic mutations at the interface or non-interface of HRAS

Follicular carcinoma is the second most common thyroid cancer, which accounts for about 15% of all thyroid malignancies. The v-Ha-ras Harvey rat sarcoma viral oncogene homologue (*HRAS*) encodes a follicular cancer-related protein located at the inner surface of cell membrane. The protein plays an important role in the transduction of signals arising from tyrosine kinase and G-protein-coupled receptors. One pathogenic mutation (Q61K) in *HRAS* was found to cause constitutive activation of the downstream signalling pathways (Nikiforova et al., 2003). Another disease-causing mutation (A146T) was identified in patients with Costello syndrome, and was shown to affect the GTP/GDP binding of HRAS (Zampino et al., 2007). In this study, the heterodimer structure of HRAS has been built using the crystal structure of the transforming protein RhoA (PDB: 1OW3) as the template. The AAS Q61K is located at the interaction interface (Figure 3(A)), and A146T lies in a non-interface region of HRAS (Figure 3(B)). These two mutations in different structural regions have been analysed to assess their effects on protein stability and protein–protein interaction.

Both amino acid residues in the HRAS protein, Gln 61 and Ala 146, are conserved in other species (Figure 3(C)), suggesting that they may be functionally important sites. As shown in Table 1, the folding energy changes ($\Delta\Delta G_{stability}$) caused by Q61K and A146T are –4.42 kcal/mol and –1.39 kcal/mol (Table 1), respectively, suggesting that both mutations may destabilise the HRAS monomer structure. Consistent with the above-mentioned results, the predictions made by structure-based tools show decreased protein stability for both mutations (excluding I-Mutant3.0 structure version for Q61K). Furthermore, all the sequence-based methods also predict that A146T could make HRAS protein unstable.

The Q61K mutation is located at the interaction interface (Figure 3(A)), and the binding energy change caused by Q61K is $\Delta\Delta\Delta G_{binding} = -7.29$ kcal/mol, or $\Delta\Delta\Delta G_{FoldX} = -2.40$ kcal/mol (Table 1), suggesting that the mutation may significantly affect protein–protein interaction. The distance between Gln 61 and its interaction partner, Arg 47 from the other chain, is only 1.88 Å, which is within the distance of hydrogen bond formation. When the polar residue Gln is replaced by positively charged residue Lys, the hydrogen bonds may be affected, and thus make strongly unfavourable interactions with Arg 47. In contrast, the A146T mutation located in a non-interface region (Figure 3(B)) does not appear to have a significant effect on protein–protein interaction. As shown in Table 1, the binding energy change caused by A146T is $\Delta\Delta\Delta G_{binding} = -0.21$ kcal/mol, or $\Delta\Delta\Delta G_{FoldX} = -0.11$ kcal/mol. Nevertheless, Ala 146 and its neighbouring residues (Leu 15 and Val 148) may form the hydrophobic pocket, which is involved in the binding of the purine ring of GTP/GDP. The substitution of Ala 146 by the polar residue Thr may alter the hydrophobic environment in the pocket, and thus affect the binding of GTP or GDP.

Figure 3 Illustration of two disease-causing mutations (Q61K and A146T) in human HRAS: (A) Structural representation of the Q61K mutation. Chains A and B are shown in green and yellow, respectively. The residue Gln 61 (Wild-Type) is shown in red, Lys 61 (mutant) in blue, and Arg 47 of chain B in cyan. The hydrogen bond is represented as a white dash line; (B) Structural representation of the A146T mutation. Chains A and B are shown in green and yellow, respectively. Ala 146 (Wild-Type) is shown in magenta, and Thr 146 (mutant) in blue. Two neighbouring residues, Leu 15 in orange and Val 148 in white, are also shown and (C) Multiple sequence alignment of HRAS with the amino acid substitution sites indicated (see online version for colours)



3.3 Application: the A693V substitution in ZBTB20

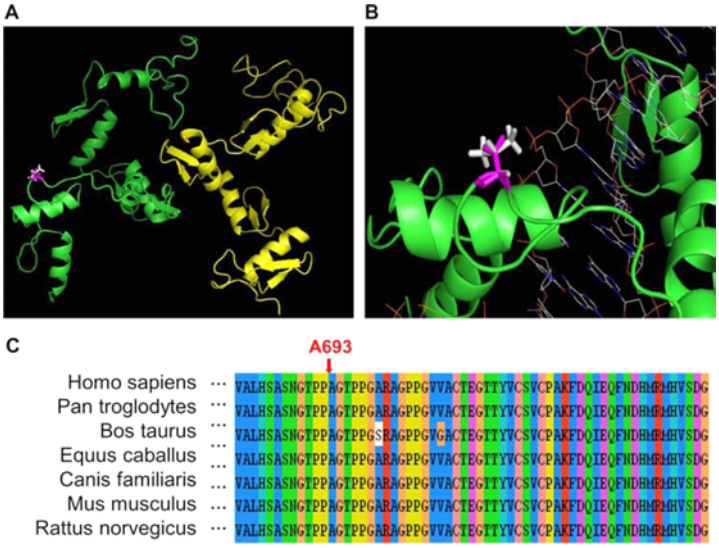
The structure-based approach was also used to investigate the amino acid substitution, A693V, in the human *ZBTB20* gene (zinc finger and BTB domain containing 20). *ZBTB20* plays important roles in neurogenesis (Mitchelmore et al., 2002), postnatal survival and glucose homeostasis (Sutherland et al., 2009). The A693V substitution is implicated to impair the function of ZBTB20 in the brain. Thus, predicting the effects of A693V on protein stability and function may help determine the pathogenic potential of the amino acid substitution.

The structure of the C-terminal region (560–739) of ZBTB20, including five zinc finger domains, was constructed using the homology modelling method with the six-finger zinc finger peptide (PDB: 2I13) as the template. As shown in Figure 4(A), although ZBTB20 may form a homodimer structure, the A693V mutation is not located at the interaction interface. The binding energy change caused by A693V is $\Delta\Delta\Delta G_{binding} = -0.31$ cal/mol, or $\Delta\Delta\Delta G_{FoldX} = 0$ kcal/mol (Table 2), suggesting that the AAS has little effect on dimer formation. The folding energy change was also calculated for A693V using TINKER ($\Delta\Delta\Delta G_{stability} = -2.69$ kcal/mol, Table 1). In addition, all of the structure-based methods predicted that A693V will decrease protein stability. Thus, the consensus prediction is that A693V will slightly destabilise the monomer structure of ZBTB20.

Since the ZBTB20 protein was previously shown to bind DNA (Mitchelmore et al., 2002), the structure of ZBTB20 in complex with DNA has been modelled using the six-finger zinc finger peptide (PDB: 2I13) as the template. As shown in Figure 4(B),

the amino acid residue, Ala 693, is located close to the phosphate group of DNA backbone. Therefore, another possibility is that the A693V substitution may be involved in protein–DNA interaction. The multiple sequence alignment shown in Figure 4(C) also suggests that Ala 693 is highly conserved in other species, and thus may be important for the normal function of ZBTB20.

Figure 4 Illustration of the A693V substitution in human ZBTB20: (A) Structural representation of the A693V substitution. Chains A and B are shown in green and yellow, respectively. Ala 693 (Wild-Type) is shown in magenta, and Val 693 (mutant) in white; (B) Representation of the modelled structure of ZBTB20 in complex with DNA. Shown are chain A in green, Ala 693 in magenta, Val 693 in white, and the DNA molecule as wireframe and (C) Multiple sequence alignment of ZBTB20 with the amino acid substitution site indicated (see online version for colours)



4 Conclusion

In this paper, a structure-based approach is described for assessing the effects of AASs on protein stability and protein–protein interaction. Homology modelling and free energy calculation methods were used to compute the differences of folding energy and binding energy between the WT and mutant structures. Three case studies showed that the disease-causing mutations at the interaction interface might reduce the binding energy, and thus weaken the affinity in the protein complex. The pathogenic mutations in the non-interface region could reduce the folding energy and thus destabilise the monomer structure. Therefore, the structure-based approach can be used to quantitatively assess the effects of AASs on protein stability and protein–protein interaction. The approach may be useful for understanding the molecular mechanisms by which gene mutations cause human diseases.

Acknowledgement

This work is supported by the CSREES/USDA, under project number SC-1700355. Emil Alexov and Charles E. Schwartz are supported by grants from NLM, NIH, grant numbers 1R03LM009748 and 1R03LM009748-S1.

References

Aasly, J.O., Toft, M., Fernandez-Mata, I., Kachergus, J., Hulihan, M., White, L.R. and Farrer, M. (2005) 'Clinical features of LRRK2-associated Parkinson's disease in central Norway', *Ann. Neurol.*, Vol. 57, No. 5, pp.762–765.

Albrecht, M. (2005) 'LRRK2 mutations and parkinsonism', *Lancet*, Vol. 365, No. 9466, p.1230.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) 'Gapped BLAST and PSI-BLAST: a new generation of protein database search programs', *Nucleic Acids Res.*, Vol. 25, No. 17, pp.3389–3402.

Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S. and Karplus, M. (1983) 'CHARMM: A program for macromolecular energy, minimization and dynamic calculations', *J. Comp. Chem.*, Vol. 4, pp.187–217.

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

Cargill, M., Altshuler, D., Ireland, J., Sklar, P., Ardlie, K., Patil, N., Shaw, N., Lane, C.R., Lim, E.P., Kalyanaraman, N., Nemesh, J., Ziaugra, L., Friedland, L., Rolfe, A., Warrington, J., Lipshutz, R., Daley, G.Q. and Lander, E.S. (1999) 'Characterization of single-nucleotide polymorphisms in coding regions of human genes', *Nat. Genet.*, Vol. 22, No. 3, pp.231–238.

Chui, D.H. and Waye, J.S. (1998) 'Hydrops fetalis caused by alpha-thalassemia: an emerging health care problem', *Blood*, Vol. 91, No. 7, pp.2213–2222.

Dehouck, Y., Grosfils, A., Folch, B., Gilis, D., Bogaerts, P. and Rooman, M. (2009) 'Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0', *Bioinformatics*, Vol. 25, No. 19, pp.2537–2543.

Dill, K.A., Fiebig, K.M. and Chan, H.S. (1993) 'Cooperativity in protein-folding kinetics', *Proc. Natl. Acad. Sci. USA*, Vol. 90, No. 5, pp.1942–1946.

Dimmic, M.W., Sunyaev, S. and Bustamante, C.D. (2005) 'Inferring SNP function using evolutionary, structural, and computational methods', *Pac. Symp. Biocomput.*, Vol. 10, pp.382–384.

Dixit, A., Torkamani, A., Schork, N.J. and Verkhivker, G. (2009) 'Computational modeling of structurally conserved cancer mutations in the RET and MET kinases: the impact on protein structure, dynamics, and stability', *Biophys. J.*, Vol. 96, No. 3, pp.858–874.

Facchiano, A. and Marabotti, A. (2009) 'Analysis of galactosemia-linked mutations of GALT enzyme using a computational biology approach', *Protein Eng. Des. Sel.*, Vol. 23, No. 2, pp.103–113.

Huang, L.T., Gromiha, M.M. and Ho, S.Y. (2007) 'iPTREE-STAB: interpretable decision tree based method for predicting protein stability changes upon mutations', *Bioinformatics*, Vol. 23, No. 10, pp.1292–1293.

Kachergus, J., Mata, I.F., Hulihan, M., Taylor, J.P., Lincoln, S., Aasly, J., Gibson, J.M., Ross, O.A., Lynch, T., Wiley, J., Payami, H., Nutt, J., Maraganore, D.M., Czyzewski, K., Styczynska, M., Wszolek, Z.K., Farrer, M.J. and Toft, M. (2005) 'Identification of a novel LRRK2 mutation linked to autosomal dominant parkinsonism: evidence of a common founder across European populations', *Am. J. Hum. Genet.*, Vol. 76, No. 4, pp.672–680.

Kan, Y.W. (1991) 'Molecular biology of hemoglobin: its application to sickle cell anemia and thalassemia', *Schweiz. Med. Wochenschr. Suppl.*, Vol. 43, pp.51–54.

Khan, S. and Vihinen, M. (2010) 'Performance of protein stability predictors', *Hum. Mutat.*, Vol. 31, No. 6, pp.675–684.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J. and Higgins, D.G. (2007) 'Clustal W and Clustal X version 2.0', *Bioinformatics*, Vol. 23, No. 21, pp.2947–2948.

Mata, I.F., Wedemeyer, W.J., Farrer, M.J., Taylor, J.P. and Gallo, K.A. (2006) 'LRRK2 in Parkinson's disease: protein domains and functional insights', *Trends Neurosci.*, Vol. 29, No. 5, pp.286–293.

Miller, M.P. and Kumar, S. (2001) 'Understanding human disease mutations through the use of interspecific genetic variation', *Hum. Mol. Genet*., Vol. 10, No. 21, pp.319–328.

Mitchelmore, C., Kjaerulff, K.M., Pedersen, H.C., Nielsen, J.V., Rasmussen, T.E., Fisker, M.F., Finsen, B., Pedersen, K.M. and Jensen, N.A. (2002) 'Characterization of two novel nuclear BTB/POZ domain zinc finger isoforms. Association with differentiation of hippocampal neurons, cerebellar granule cells, and macroglia', *J. Biol. Chem.*, Vol. 277, No. 9, pp.7598–7609.

Ng, P.C. and Henikoff, S. (2003) 'SIFT: predicting amino acid changes that affect protein function', *Nucleic Acids Res.*, Vol. 31, No. 13, pp.3812–3814.

Nikiforova, M.N., Lynch, R.A., Biddinger, P.W., Alexander, E.K., Dorn 2nd, G.W., Tallini, G., Kroll, T.G. and Nikiforov, Y.E. (2003) 'RAS point mutations and PAX8-PPAR gamma rearrangement in thyroid tumors: evidence for distinct molecular pathways in thyroid follicular carcinoma', *J. Clin. Endocrinol. Metab*., Vol. 88, No. 5, pp.2318–2326.

Park, H. and Lee, S. (2005) 'Prediction of the mutation-induced change in thermodynamic stabilities of membrane proteins from free energy simulations', *Biophys. Chem.*, Vol. 114, Nos. 2–3, pp.191–197.

Petrey, D., Xiang, Z., Tang, C., Xie, L., Gimpelev, M., Mitros, T., Soto, C., Goldsmith-Fischman, S., Kernytsky, A., Schlessinger, A. Koh, I., Alexov, E. and Honig, B. (2003) 'Using multiple structure alignments, fast model building, and energetic analysis in fold recognition and homology modeling', *Proteins*, Vol. 53, pp.430–435.

Ponder, J.W. (1999) *TINKER-software Tools for Molecular Design*, Version 3.7, Washington University, St. Luis.

Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F. and Serrano, L. (2005) 'The FoldX web server: an online force field', *Nucleic Acids Res*., Vol. 33 (Web Server issue), pp.W382–W388.

Smigielski, E.M., Sirotkin, K., Ward, M. and Sherry, S.T. (2000) 'dbSNP: a database of single nucleotide polymorphisms', *Nucleic Acids Res.*, Vol. 28, No. 1, pp.352–355.

Stitziel, N.O., Tseng, Y.Y., Pervouchine, D., Goddeau, D., Kasif, S. and Liang, J. (2003) 'Structural location of disease-associated single-nucleotide polymorphisms', *J. Mol. Biol.*, Vol. 327, No. 5, pp.1021–1030.

Sutherland, A.P., Zhang, H., Zhang, Y., Michaud, M., Xie, Z., Patti, M.E., Grusby, M.J. and Zhang, W.J. (2009) 'Zinc finger protein Zbtb20 is essential for postnatal survival and glucose homeostasis', *Mol. Cell Biol.*, Vol. 29, No. 10, pp.2804–2815.

Teng, S., Srivastava, A.K. and Wang, L. (2010) 'Sequence feature-based prediction of protein stability changes upon amino acid substitutions', *BMC Genomics*, Vol. 11, Suppl. 2, p.S5.

Teng, S., Madej, T., Panchenko, A. and Alexov, E. (2009) 'Modeling effects of human single nucleotide polymorphisms on protein-protein interactions', *Biophys. J.*, Vol. 96, No. 6, pp.2178–2188.

Teng, S., Michonova-Alexova, E. and Alexov, E. (2008) 'Approaches and resources for prediction of the effects of non-synonymous single nucleotide polymorphism on protein function and interactions', *Curr. Pharm. Biotechnol.*, Vol. 9, No. 2, pp.123–133.

Wang, Z. and Moult, J. (2001) 'SNPs, protein structure, and disease', *Hum. Mutat*., Vol. 17, No. 4, pp.263–270.

Wheeler, D.L., Barrett, T., Benson, D.A., Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., DiCuccio, M., Edgar, R., Federhen, S., Geer, L.Y., Kapustin, Y., Khovayko, O., Landsman, D., Lipman, D.J., Madden, T.L., Maglott, D.R., Ostell, J., Miller, V., Pruitt, K.D., Schuler, G.D., Sequeira, E., Sherry, S.T., Sirotkin, K., Souvorov, A., Starchenko, G., Tatusov, R.L., Tatusova, T.A., Wagner, L. and Yaschenko, E. (2007) 'Database resources of the National Center for Biotechnology Information', *Nucleic Acids Res.*, Vol. 35 (Database issue), pp.D5–D12.

Xiang, Z. (2006) 'Advances in homology protein structure modeling', *Curr. Protein Pept. Sci.*, Vol. 7, No. 3, pp.217–227.

Xiang, Z. and Honig, B. (2001) 'Extending the accuracy limits of prediction for side-chain conformations', *J. Mol. Biol.*, Vol. 311, pp.421–430.

Yin, S., Ding, F. and Dokholyan, N.V. (2007) 'Eris: an automated estimator of protein stability', *Nat. Methods*, Vol. 4, No. 6, pp.466–467.

Yue, P. and Moult, J. (2006) 'Identification and analysis of deleterious human SNPs', *J. Mol. Biol.*, Vol. 356, No. 5, pp.1263–1274.

Zampino, G., Pantaleoni, F., Carta, C., Cobellis, G., Vasta, I., Neri, C., Pogna, E.A., De Feo, E., Delogu, A., Sarkozy, A., Atzeri, F., Selicorni, A., Rauen, K.A., Cytrynbaum, C.S., Weksberg, R., Dallapiccola, B., Ballabio, A., Gelb, B.D., Neri, G. and Tartaglia, M. (2007) 'Diversity, parental germline origin, and phenotypic spectrum of de novo HRAS missense changes in Costello syndrome', *Hum. Mutat.*, Vol. 28, No. 3, pp.265–272.

Zhang, Z., Teng, S., Wang, L., Schwartz, C.E. and Alexov, E. (2010) 'Computational analysis of missense mutations causing Snyder-Robinson syndrome', *Hum. Mutat.*, Vol. 31, No. 9, pp.1043–1049.

Zhou, H. and Zhou, Y. (2002) 'Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction', *Protein Sci.*, Vol. 11, No. 11, pp.2714–2726.