ISSN NO. 2320-5407

Journal Homepage: -www.journalijar.com

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH (IJAR)

**Article DOI:**10.21474/IJAR01/1337
**DOI URL:** http://dx.doi.org/10.21474/IJAR01/1337

*RESEARCH ARTICLE*

# IN-SILICO BASED PREDICTION OF DELETERIOUS NON-SYNONYMOUS SINGLE NUCLEOTIDE POLYMORPHISMS IN THE HUMAN CDK1 GENE

**Shikha Suman[1], Gyan Prakash Tiwari[2], Ashutosh Mishra[1], and Anurag Kulshrestha[3].**

1. Indian Institute of Information Technology, Allahabad-211012, India.
2. Dr. Ram Manohar Lohia Institute of Medical Sciences, Lucknow-226010, India.
3. CSIR-National Botanical Research Institute, Lucknow-226001, India.

## Manuscript Info

*Manuscript History*

Received: 15 June 2016
Final Accepted: 10 July 2016
Published: August 2016

***Key words:-***
Polymorphism, Non-synonymous, Deleterious

## Abstract

Cyclin dependent kinase 1 (CDK1) is involved in regulation of cell cycle in eukaryotes. Polymorphism in this CDK1 have been associated with various cancers in human. Using structure, function and evolution based bioinformatics approaches, most damaging single nucleotide polymorphisms (SNPs) in CDK1 gene and their impact on the structure and function of CDK1 protein were predicted. Among the 23 non-synonymous SNPs (nsSNPs) retrieved from the Uniprot and Ensemble databases, 2 nsSNPs were concluded to be most deleterious SNPs after the prediction analysis of 6 computational tools namely SIFT, Polyphen-2, PROVEAN, mCSM, AlignGVGD and Mutation assessor. I136N and R275Q were predicted to be highly damaging and hence affecting the structure, function and stability of CDK1 protein, which were further cross-checked using I-Mutant and DUET tools. These SNPs can be further considered for wet lab validation.

## Introduction:-

Cyclin dependent kinase 1(CDK1) is an important gene involved in cell cycle regulation (Malumbres, 2011). Activation of CDK1 is the target point of many signaling pathways responsible for the commencement of cell division (Rhind et al., 2012). Dysregulation and mutation of these genes may result in serious consequences like uterine cancer, lung cancer, breast cancer etc. (Deshpande et al., 2005). The human CDK1 protein contain 297 amino acids. Single nucleotide polymorphism (SNP) is a common genetic variation in humans (Cargill et al., 1999). A number of polymorphisms have been reported in CDK1 till date. Non-synonymous SNPs causing change in the corresponding protein, are mostly associated with disorders and disease in humans. These nsSNPs not only affect protein structure and function but also influences the stability of protein.

CDK1 (cyclin dependent kinase 1) is the member of serine/threonine kinase family and has been known to modulate mitotic onset and centrosome cycle thus controlling the eukaryotic cell cycle (Malumbres, 2011). Its association with wide spectrum of cancers such as squamous cell carcinomas of neck and head, tongue and larynx,

**Corresponding Author:-Shikha Suman**
Address:-Indian institute of Information Technology, Allahabad-211012, India.

esophagus, uterine cervix cancer, lung cancer, breast cancer and many other has already been well established (Deshpande et al., 2005; Cargill et al., 1999; Casimiro et al., 2012).

A number of computational methods employing various algorithms have been developed over the years for demarcating the damaging SNPs from the neutral. These approaches uses chemical, structural, functional, evolutionary and physical properties of the proteins. Combining the analysis from various computational methods may increase the prediction accuracy, instead of following single tool to prioritize the deleterious nsSNPs. This type of approaches for variant analysis can further improve our understanding of the mechanism with which these mutations alter the function of the proteins thus can leading to disease. This study aimed at identifying the deleterious non-synonymous single nucleotide polymorphisms which affects the structure and function of CDK1 protein. Myriad of computational tools were used to identify the pathogenic variants of CDK1 and scrutinize its effect on the structure, function and stability of the protein.

## Methodology:-

### Primary sequence analysis:-

The amino acid sequence of human CDK1 protein was obtained from Uniprot (http://www.uniprot.org) with accession number P06493. Basic Local alignment search tool (BLAST) was performed on CDK1 protein against non-redundant protein sequences to identify the protein family. Conserved domain search (CD Search) (Marchler-Bauer et al., 2014) was utilized to identify conserved domain of CDK1. Search tool for Retrieval of Interacting Gene/Proteins (STRING) (Von Mering et al., 2003) version 10 was employed for identifying the interactions of CDK1 though its vast resource of interaction information obtained from high-throughput experiments, genomic context, co-expression and previous knowledge.

### SNPs retrieval:-

The information on polymorphisms in the coding region of CDK1 was retrieved from databases such as dbSNP (http://www.ncbi.nlm.nih.gov/snp) and Uniprot. Only large scale (LS) studies were considered for analysis in Uniprot. To increase our knowledge of the nsSNPs in CDK1 protein, the datasets form both the databases were combined. The three dimensional structure of CDK1 (PDB Id: 4YC6) protein was downloaded from Protein Data bank (PDB) to study the effect of nsSNPs on the structure and function of protein.

### SNP variant effects:-

To elucidate the effect of amino acid substitutions on protein structure and function, the nsSNPs were scrutinized using six prediction tools: SIFT (Kumar et al., 2009), Polyphen-2 (Adzhubei et al., 2010), AlignGVGD (Mathe et al., 2006), Mutation Assessor (Reva et al., 2011), PROVEAN (Choi et al., 2015) and mCSM (Pires et al., 2014).

SIFT classifies the substitution as deleterious or tolerated depending on the amino acid substitution type using sequence homology based features. The normalized probability of discovering a new amino acid at a particular position is referred to as SIFT score and ranges from 0 to 1. Variants with a SIFT score less than 0.05 indicate a deleterious effect of nsSNP on protein function while a score greater than 0.05 is assumed tolerated.

Polyphen-2 predicts the effect of amino acid substitution on structure and function of protein by utilizing protein structure and multiple sequence alignment (MSA) procured information. Polyphen-2 is a Naive Bayes classifier that assigns a PSIC (Position Specific Independent Count) score to the variants ranging from 0 to 1. A substitution is classified as “probably damaging” if the probabilistic score is greater than 0.85 and “possibly damaging” if the score is in between 0.15 to 0.85. The rest mutations are classified as benign.

PROVEAN (Protein Variation Effect Analyzer) predicts the change in biological function of a protein upon amino acid substitution or INDEL. It assigns a delta alignment score based on sequence clustering of reference and variant protein sequences. The substitutions with score less than -2.5 were termed as deleterious to protein function.

Mutation Assessor predicts the functional impact of amino acid substitution in proteins based upon evolutionary conservation of amino acids in homologous proteins. It provides a functional impact score (FIS) based on evolutionary knowledge classifying nsSNPs as low, medium, high and neutral.

AlignGVGD integrates biophysical attributes of amino acids together with MSA to predict the whether an amino acid substitution is deleterious or neutral. Residues are exchanged with the frequency of substitution to obtain and compare physical as well as chemical properties. It provides a “C-score” which ranges from C0 to C65, with class 65 denoting deleterious mutations while class 0 represents neutral mutations.

Mutation Cutoff Scanning Matrix (mCSM) predicts the effect the amino acid substitution on the stability of protein through graph based signatures. The PDB structure of CDK1 protein was used as input along with mutated site and residue information to predict stability change (ΔΔG).

### Evolutionary conservation of amino acids:-

Every residue contributes to the maintenance of protein structure, highlighting the importance of identifying evolutionary conservation of amino acids. Consurf (Celniker et al., 2013) was employed to calculate the degree of conservation of amino acids in CDK1 protein using MSA derived information and then determining the conservation rate by Bayesian inference. Consurf score ranges from 1 to 9 with 1 denoting variable residues and 9 denoting evolutionary conserved residues.

### Biophysical characterization of protein:-

SNPeffect (De Baets et al., 2011) integrates 4 tools namely, TANGO (prediction of aggregation prone regions), WALTZ (prediction of amyloid forming regions), LIMBO (Prediction of chaperone binding sites) and FoldX (analysis of protein stability) for molecular characterization of polymorphic variants in proteins. Mutations are categorized into mutations that decrease (dTANGO< -50), increase (dTANGO> 50) and do not affect (dTANGO between -50 and 50) the tendency of protein aggregation. Also, mutations that increase (dWALTZ> 50), decrease (dWALTZ< -50) and do not affect (dWALTZ between -50 and 50) the propensity of amyloid region in protein.

### Protein stability studies:-

An amino acid substitution in a protein sequence can cause significant change in protein stability. I-Mutant 2.0 (Capriotti et al., 2005) is a support vector machine (SVM) based classifier trained with extensive dataset derived from Protherm. A free energy change (DDG) is calculated between the wild and mutant variants based on Gibbs free energy. If DDG $< 0$, the mutation causes a decrease in stability of protein while DDG $> 0$ denotes an increase in stability of protein upon mutation.

SDM (Site Directed Mutator) (Worth et al., 2011) is a server for assessing the effect of amino acid substitution on protein stability and function through potential energy function that uses environment specific amino acid substitution frequency within homologous families for calculation of a stability score. This stability score is similar to the free energy change difference between wild and mutant protein.

DUET (Pires et al., 2014) sever integrates two methodologies namely, mCSM and SDM to compute a consensus prediction using SVM regression with a radial basis kernel function for prediction of protein stability upon amino acid substitution.

## Results and Discussion:-

The human CDK1 protein was predicted to be a member of protein kinase superfamily by BLASTp against non-redundant protein sequences. A putative conserved domain was identified between 3-287 positions incorporating catalytic domain of serine/threonine kinase. CDK1 protein was found to be interacting with 10 other proteins such as Cyclin B1 (CCNB1) and Cyclin A2 (CCNA2) with confidence score greater than 0.900 (Figure 1).

A total of 23 non-coding synonymous SNPs in CDK1 gene were retrieved from dbSNP and Uniprot. The nsSNPS were scrutinized and categorized using six tools such as SIFT, Polyphen-2, Provean, Mutation Assessor, mCSM and AlignGVGD. SIFT predicted 12 nsSNPS as deleterious or affecting the protein function. Out of which 7 nsSNPs (D73P, Y15C, V18L, I136N, S178L, D211G AND R275Q) were predicted as highly deleterious with SIFT score of 0.0. Out of the 23 polymorphic variants, mCSM predicted 20 nsSNPs as damaging or destabilizing the protein structure. 19 SNPs were predicted to be damaging the protein function by PROVEAN with PROVEAN score less than -2.5. 13 nsSNPs were classified as deleterious by AlignGVGD, 10 variants were predicted to be affecting the protein structure and function by Polyphen-2. Mutation Assessor classified 2 nsSNPs with high confidence as damaging to protein structure and function. Based on the six prediction tools, 2 SNPs namely, I136N and R275Q were predicted to be deleterious to protein structure and function.

Further analysis revealed the impact of these mutations on the structure and function of the protein. Thus, suggesting the significant role in initiation of the diseases.

To confirm the deleterious effect of the two variants, the effect of substitutions on the stability of protein was predicted using three protein stability analysis tools namely, I-mutant, DUET and SDM. Both the nsSNPs were predicted to destabilize the CDK1 protein by yielding negative free energy change (ΔΔG) values (Table 1). I136N was found to form extended beta sheet and R275Q formed the loop or the irregular structure in the protein. Evolutionary conservation information of amino acids of CDK1 protein revealed that isoleucine and arginine at 136th and 275th position respectively were present in the highly conserved region.

Molecular phenotypic analysis of the deleterious variants was performed to understand the molecular characteristics of disease causing mutations. SNPeffect predicted the reduction in protein stability as a result of these two mutations. However, no effect on aggregation tendency, amyloid propensity and chaperone binding tendency of protein was predicted.

The impact of the amino acid substitutions on the local structural environment of protein was studied using HOPE (Venselaar et al., 2010) server. Replacement of isoleucine with asparagine at 136th position was termed as deleterious by all the six prediction tools (Figure 4). Asparagine is bigger and less hydrophobic as compared to isoleucine. Isoleucine being a small residue was found to be buried in the core of the protein but the bigger residue, asparagine, may not fit in the core of protein. The position at which the mutation happened was the site for protein kinase domain. Moreover, MAPK docking motif was predicted at the site of mutation. The mutation can thus damage the motif and hence its function. Overall, the mutant residue has entirely different property than the wild type, it can disturb this domain and abolish its function.

Another substitution at 275th position, where arginine is replaced with glutamine was considered as deleterious by all the six prediction tools. Arginine is a bigger and positively charged residue whereas glutamine is neutral and smaller. This change in charge can cause empty space at the core of the protein. Arginine forms the hydrogen bond with glutamic acid at position 173, proline at position 184 and isoleucine at position 269. Due to the size difference between glutamine and arginine, glutamine does not make hydrogen bonds. Also, the arginine forms salt bridge with glutamic acid at position 173 and aspartic acid at position 271. Since the natural and wild type residues have different charge properties, ionic interactions in the CDK1 protein are also disturbed. This mutation is also located in the domain that functions as protein kinases, which is hence abolished. The PKA phosphorylation site is damaged due to mutation as the mutation occurs in this motif.

The variant I136N is strongly associated with lung cancer whereas R275Q variant was found to be associated with breast cancer and colon cancer. This was found from the uniprot databases about non-synonymous SNPs, classified as pathogenic.

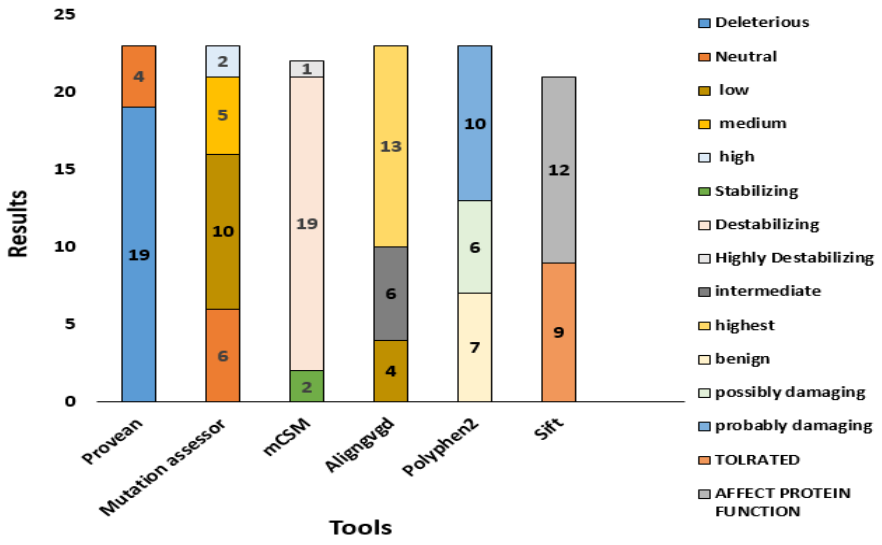


**Figure 1:-** Prediction results of 23 nsSNPs in the CDK1 gene by six tools.

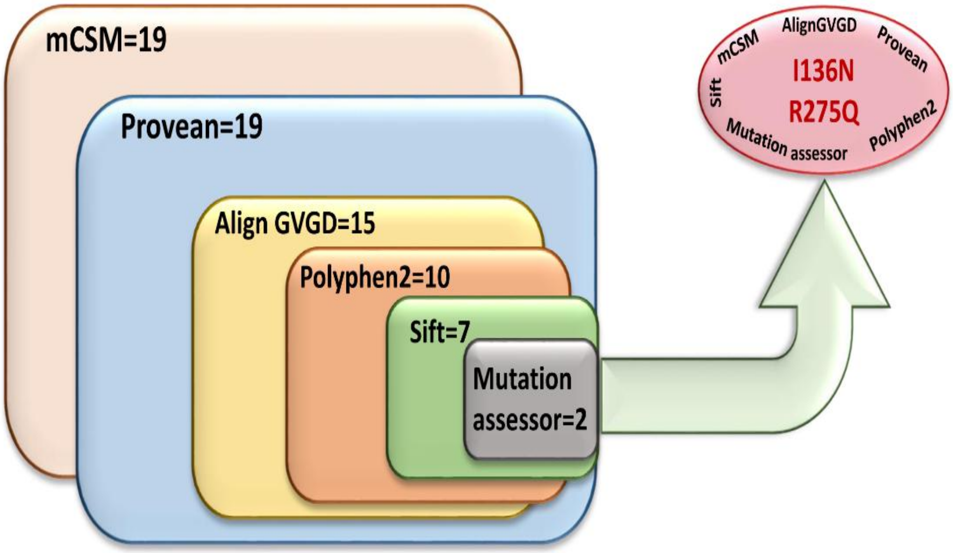


**Figure 2:-** Consensus prediction of all the six tools.

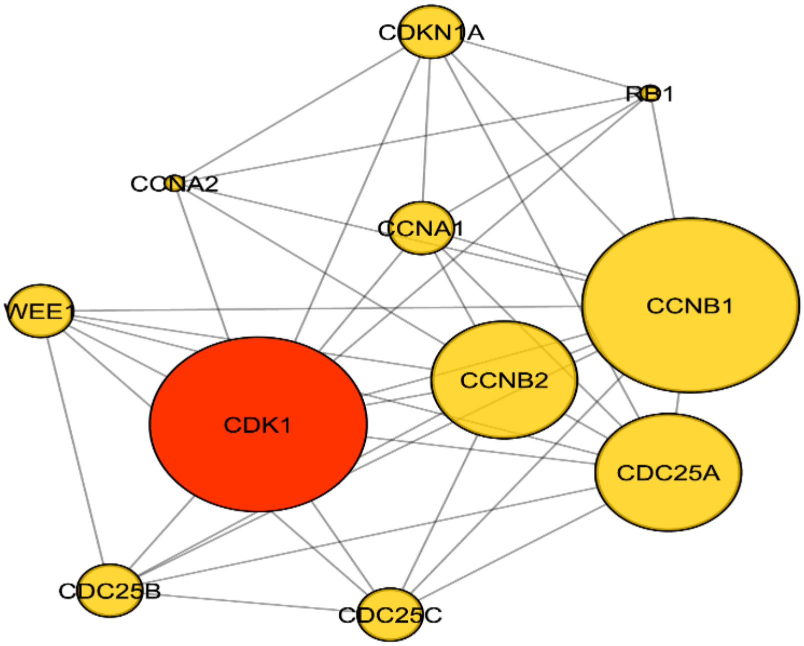


**Figure 3:-** protein-protein interaction of CDK1.

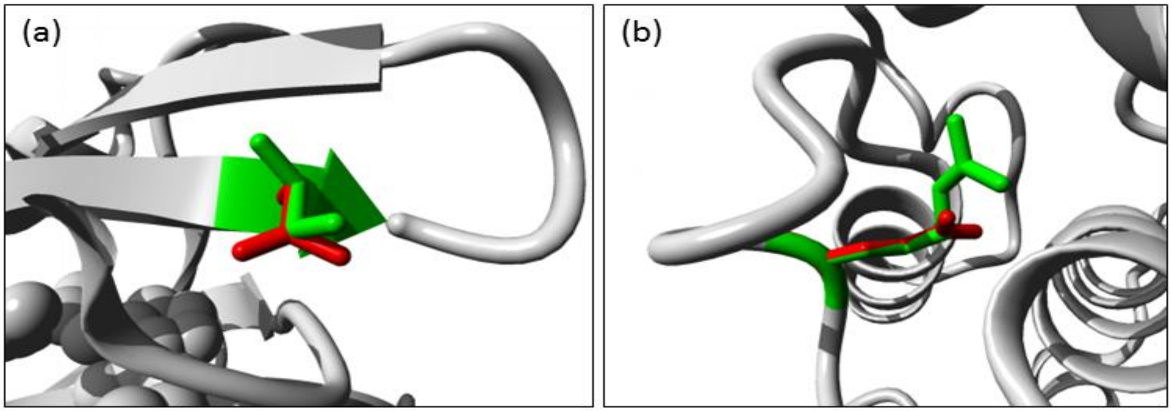
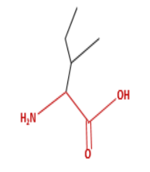
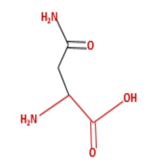
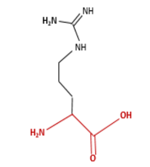
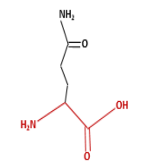


**Figure 4:-** secondary structure depiction of mutation in (a) I136N and (b) R275Q. Wild and mutated structure is represented by green and red color respectively.

**Table 1:-** Variants with free energy change values with their conservation score and type.

| **Variant** | **Position** | **I-mutant (ΔΔG)** | **DUET (ΔΔG)** | **SDM (ΔΔG)** | **Conservation score** | **Conservation type** |
|---|---|---|---|---|---|---|
| I/N | 136 | -2.76 | -2.84 | -5.26 | 9 | Conserved |
| R/Q | 275 | -2.27 | -1.54 | -1.89 | 9 | Conserved |

**Table 2:** structural and chemical properties of the two highly deleterious SNPs.

| **Variant** | **Residue** | **Amino Acid** | **Molecular weight (g/mol)** | **Charge** | **Hydrophobicity** | **Structure** | **Stability** |
|---|---|---|---|---|---|---|---|
| I136N | Wild type | Isoleucine (I) | 131.173 | Neutral | Hydrophobic |  | Destabilize |
| | Mutant | Asparagine (N) | 132.118 | Neutral | Hydrophilic |  | |
| R275Q | Wild type | Arginine (R) | 174.201 | Positive | Hydrophilic |  | Destabilize |
| | Mutant | Glutamine (Q) | 146.145 | Neutral | Hydrophilic |  | |

## Conclusion:-

Numerous tools using different computational algorithms were utilized to predict the deleterious mutations in human CDK1 gene. CDK1 genes encodes an important protein responsible for cell cycle regulation. Combination of the prediction results from the 6 structural, functional and evolution based methods identified two mutations namely, I136N & R275Q as deleterious non-synonymous SNPs of CDK1. These mutations were also found to affect the stability of the CDK1 protein. This highly pathogenic mutation may contribute in understanding of many cancer mechanisms and can be further validated by experimental confirmation.

## Acknowledgments:-

Computing amenities of IIIT Allahabad is greatly acknowledged.

## References:-

1. Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S. and Sunyaev, S.R. (2010): A method and server for predicting damaging missense mutations. Nat. methods., 7(4): 248-9.
2. Capriotti, E., Fariselli, P. and Casadio, R. (2005): I-Mutant2. 0: predicting stability changes upon mutation from the protein sequence or structure. Nucleic. Acids. Res., 33(suppl 2): W306-10.
3. Cargill, M., Altshuler, D., Ireland, J., Sklar, P., Ardlie, K., Patil, N., Lane, C.R., Lim, E.P., Kalyanaraman, N., Nemesh, J. and Ziaugra, L. (1999): Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nat. genet., 22(3): 231-8.
4. Casimiro, M.C., Crosariol, M., Loro, E., Li, Z. and Pestell, R.G.(2012): Cyclins and cell cycle control in cancer and disease. Genes. Cancer., 3(11-12): 649-57.
5. Celniker, G., Nimrod, G., Ashkenazy, H., Glaser, F., Martz, E., Mayrose, I., Pupko, T., and Ben-Tal, N. (2013): ConSurf: Using Evolutionary Data to Raise Testable Hypotheses about Protein Function. Isr. J. Chem., 53(3-4): 199-206.
6. Choi, Y. and Chan, A.P. (2015): PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. Bioinformatics., 31(16): 2745-2747.
7. De Baets, G., Van Durme, J., Reumers ,J., Maurer-Stroh, S., Vanhee, P., Dopazo, J., Schymkowitz, J. and Rousseau, F. (2011): SNPeffect 4.0: on-line prediction of molecular and structural effects of protein-coding variants. Nucleic. Acids. Res., 40(Database issue): D935-9.
8. Deshpande, A., Sicinski, P. and Hinds, P.W. (2005): Cyclins and cdks in development and cancer: a perspective. Oncogene., 24(17): 2909-15.
9. Kumar, P., Henikoff, S. and Ng, P.C. (2009): Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat. protoc., 4(7): 1073-81.
10. Malumbres, M. (2011): Physiological relevance of cell cycle kinases. Physiol. Rev., 91(3): 973-1007.
11. Marchler-Bauer, A., Derbyshire, M.K., Gonzales, N.R., Lu, S., Chitsaz, F., Geer, L.Y., Geer, R.C., He, J., Gwadz, M., Hurwitz, D.I. and Lanczycki, C.J. (2014): CDD: NCBI's conserved domain database. Nucleic. Acids. Res., 43(Database issue): D222-6.
12. Mathe, E., Olivier, M., Kato, S., Ishioka, C., Hainaut, P. and Tavtigian, S.V. (2006): Computational approaches for predicting the biological effect of p53 missense mutations: a comparison of three sequence analysis based methods. Nucleic. Acids. Res., 34(5): 1317-25.
13. Pires, D.E., Ascher, D.B. and Blundell, T.L. (2014): DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. Nucleic. Acids. Res., 42(Web Server issue): W314-9.
14. Pires, D.E., Ascher, D.B. and Blundell, T.L. (2014): mCSM: predicting the effects of mutations in proteins using graph-based signatures. Bioinformatics., 30(3): 335-42.
15. Reva, B., Antipin, Y. and Sander, C. (2011): Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic. Acids. Res., 39(17): e118.
16. Rhind, N. and Russell, P. (2012): Signaling pathways that regulate cell division. Cold. Spring. Harbor. Perspect. Boil., 4(10).
17. Venselaar, H., teBeek, T.A., Kuipers, R.K., Hekkelman, M.L. and Vriend, G. (2010): Protein structure analysis of mutations causing inheritable diseases. An e-Science approach with life scientist friendly interfaces. BMC. Bioinformatics., 11(1): 1.
18. Von Mering, C., Huynen, M., Jaeggi, D., Schmidt, S., Bork, P. and Snel, B. (2003): STRING: a database of predicted functional associations between proteins. Nucleic. Acids. Res., 31(1): 258-61.
19. Worth, C.L., Preissner, R. and Blundell, T.L. (2011): SDM—a server for predicting effects of mutations on protein stability and malfunction. Nucleic. Acids. Res., 39(Web Server issue): W215-22.