SANCHO SCIENCE

Journal of Bioinformatics & Research 1(1) 2012: 10-14

Research Article
Open Access

ISSN: 2277-3363

Computational analysis of active site amino acid dihedral angles of CDK2 towards ligand binding

Srinivasu Nulaka[1]*, Appa Rao Allam[2]

[1]*Department of Computer Science (MCA), KGRL College, Bhimavaram, AP, India. *E-mail: nulaka@gmail.com*
[2]*Vice-chancellor, JNTUK, Kakinada, AP, India E-mail : apparaoallam@gmail.com*

ABSTRACT

Cell proliferation is a consequence of positive signals which promote cell division and negative signals which suppress the process. Key factors in this signaling cascade are a series of cyclin dependent kinases (CDKs). It has been identified experimentally that CDK enzymes are highly flexible and the ligand binding orientations are primarily influenced by side chain torsions of amino acids in active site region. Hence to address the importance of backbone and side chain phi, psi and chi angle contributions upon ligand binding, various computational softwares and approaches have been utilized to recognize the influential dihedral angles towards ligand binding. The dihedrals angles (phi, psi, chi1, chi2, chi3, chi4) of all 135 enzymes from protein data bank were calculated using DANG software. The effect of changes in the backbone and side chain torsion angles (phi, psi and chi) on ligand binding within CDK2 is predicted using multiple regression analysis. After removing few data as outliers, 121 proteins as training set and 7 proteins as validation (test) set resulted in 19 variable model with regression coefficient, r: 0.765 R^2: 0.586 and Cross Validation, r^2(CV): 0.977. The results showed that 19 out of 85 independent variables (torsion angles) are highly influential towards ligand binding with in CDK2 proteins.

KEYWORDS: CDK2, regression, dihedral angle, docking, validation

INTRODUCTION

Cell proliferation is a consequence of positive signals which promote cell division and negative signals which suppress the process. Key factors in this signaling cascade are a series of cyclin dependent kinases (CDKs) [1]. Cyclin-dependent kinases are a family of serine/threonine kinases which play a crucial role in cell cycle control and are involved in diverse cellular processes, in regulation of cell division (CDKs1, 2, 3, 4, 6 and 7), transcription (CDKs7, 8 and 9) or maintenance of the structure of the cytoskeleton (CDK5) [2].

Cyclin dependent kinases control the cell cycle progression operating at the transition from G_2 to M, G_1 to S phases, and progression through S phase, regulated by a complex set of mechanisms, including the presence of activating cyclins, regulatory phosphorylations, and endogenous CDK inhibitors at checkpoints (Figure 1) [3]. Cell cycle progresses by the activation of Cyclin and CDK complexes. These cyclins and CDKs function as check points regulating the transition from one phase of cell cycle to another. Structural studies have explored the active and inactive states of CDK2 [4].

*Corresponding Author email: *nulaka@gmail.com*

Monomeric form was inactive, while association of Cyclin A with CDK2 and Thr160 phosphorylation results active CDK2 [5-6]. Activation of CDK2 results in rotation of N- and C-terminal domains leading to a slight widening of ATP cleft. The movement of PSTAIRE helix and Glu51 and the subsequent reorganization leads to reshaping of the phosphate-binding site [7].

PSTAIRE helix**:** This is an alpha helix in the amino-terminal lobe of CDK2, which interacts with cyclin and is moved inward upon cyclin binding, resulting in reorientation of key active-site residues. The name of this helix comes from its amino-acid sequence, which is conserved among all major CDKs [8-9].

All CDK inhibitors studied so far act by competing with ATP for binding in the CDK ATP binding pocket [10]. Most of the inhibitor contacts with the active site residues of CDK2 are hydrophobic and the complexes present few intermolecular interactions. Analysis of the contact area between inhibitor and CDK2 indicates that inhibitors with low IC_{50} values (higher affinity for CDK2) present higher contact areas and higher number intermolecular hydrogen bonds. CDK's are considered a potential target for anticancer medication [11]. If it is possible to selectively interrupt the cell cycle regulation in cancer cells by interfering with CDK action, the cell will die. Currently, some CDK inhibitors are undergoing clinical trials.

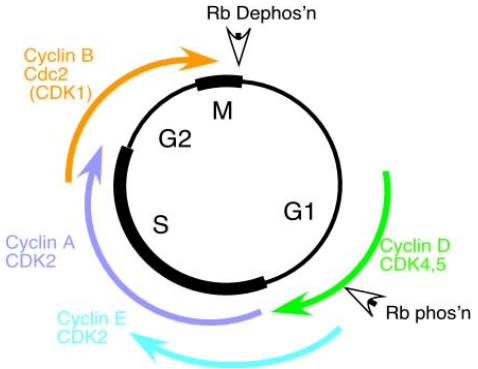

Figure 1: Various CDK/cyclin complexes involved at different stages of cell cycle

Dihedral Angles of Proteins

The atoms along a protein backbone (where the peptide bond takes place) are Cα-C-N-Cα-C-N-Cα in a repeating sequence, repeating every third atomThe backbone dihedral angles of proteins are called φ (phi, involving the backbone atoms C'-N-C^{α}-C'), ψ (psi, involving the backbone atoms N-C^{α}-C'-N) and ω (omega, involving the backbone atoms C^{α}-C'-N-C^{α}) (Figure 2). Thus, φ controls the C'-C' distance, ψ controls the N-N distance and ω controls the C^{α}-C^{α} distance.

The planarity of the peptide bond usually restricts ω to be 180° (the typical trans case) or 0° (the rare cis case. The atoms along the side chain are named with Greek letters: α, β, γ, δ, ϵ and so on. C_{α} refers to the carbon atom closest to the carbonyl group of that amino acid, C_{β} the second closest and so on. The C_{α} is usually considered a part of the backbone. The dihedral angles around the bonds between these atoms are named χ1, χ2, χ3 etc. E.g. the first and second carbon atom in the side chain of lysine is named α and β, and the dihedral angle around the α-β bond is named χ1 (chi1) [12].

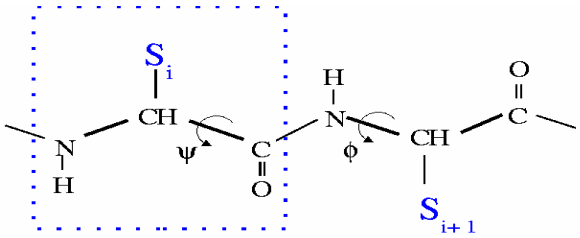

Figure 2: Image showing phi and psi torsion angles along a protein backbone

MATERIALS AND METHODS

Selection CDK2 Proteins from PDB

About 135 experimentally determined structures of CDK2 in complex with inhibitors were assembled from Protein data bank**.** Various CDK2 proteins selected from Protein Data Bank are: 1AQ1, 1B39, 1BUH, 1CKP, 1DM2, 1E1V, 1E9H, 1FIN, 1FVT, 1FVV, 1GIH, 1H01, 1H07, 1H08, 1H0V, 1H0W, 1H1P, 1H1Q, 1H1R, 1H1S, 1H24, 1H25, 1H26, 1H27, 1H28, 1HCK, 1HCL, 1JST, 1JVP, 1KE5, 1KE6, 1KE7, 1KE8, 1KE9, 1OGU, 1OIT, 1OI9, 1OIU, 1OIY, 1PF8, 1PF8, 1PKD, 1PW2, 1PXI, 1PXJ, 1PXL, 1PXM, 1PXN, 1PXO, 1PXP, 1QMZ, 1R78, 1URC, 1V1K, 1VYW, 1VYZ, 1W8C, 1W98, 1Y8Y, 1Y91, 1YKR, 2A0C, 2A4L, 2B53, 2B54, 2B55, 2BHE, 2BKZ, 2BPM, 2BTR, 2BTS, 2CCI, 2C4G, 2C5N, 2C5P, 2C5V, 2C5X, 2C68, 2C69, 2C6I, 2C6K, 2CCH, 2CJM, 2CLX, 2DUV, 2EXM, 2FVD, 2G9X, 2IW6, 2IW9, 2I40, 2J9M, 2JGZ, 2R3F, 2R3G, 2R3M, 2R3N, 2R3O, 2R3P, 2R3Q, 2R3R, 2R64, 2UUE, 2UZB, 2UZD, 2UZE, 2UZL, 2UZN, 2UZO, 2V0D, 2VTA, 2VTH, 2VTI, 2VTJ, 2VTL, 2VTN, 2VTO, 2VTP, 2VTQ, 2VTR, 2VTS, 2VTT, 2VU3, 2VV9, 2W05, 2W06, 3BHT, 3BHU, 3BHV, 3DDP, 3DDQ, 3DOG, 3EID, 3EJ1, 3EOC.

Detection of Consensus Active Site Residues

WebLab-Viewer software is used to find the amino acid residues within 8 Aº from the center of the ligand which is bound to a CDK2 protein. The fasta sequences (A chain) of the above 135 different CDK2 proteins were collected from PDB and multiple sequence alignment was conducted using ClustalW tool, to find the consensus active site residues. The following 23 amino acids are found to be consensus active site residues: Ile10, Gly11, Glu12, Gly13, Thr14, Val18, Lys20, Ala31, Lys33, Val64, Phe80, Glu81, Phe82, Leu83, His84, Gln85, Asp86, Lys129, Gln131, Asn132, Leu134, Ala144 and Asp145

DATA SET-1(Torsion Angles of active site residues of 135 CDK2 proteins)

The dihedrals angles (phi, psi, chi1, chi2, chi3, chi4) were calculated using DANG software. The Dang software reads coordinates from a Protein Data Bank molecular structure file and generates the torsion angles for each amino acid. These torsion angles (considered as independent variables) constitute the initial data set for Multiple Linear Regression using TSAR software.

DATASET-2 (Activity values obtained by performing docking experiments with reference ligands)

The activity values of all the 135 CDK2 proteins are calculated by performing docking experiments with reference ligands using MVD software. The Docking wizard option was used with all default parameters to perform docking with the reference ligand based on mol dock optimizer algorithm. The resulting dock scores is transformed into activity values based the formulae, Activity = log (1/dock-score). These values guarantee linear distribution of data. These activity values are taken as dependent variables for Multiple Linear Regression using TSAR software.

Multivariate Regression Analysis

Stepwise multiple linear regressions were performed using TSAR Version 3.3 software. This methodology is used to find the influence of flexibility of active site residues on activity. In other words the affect of changes in the backbone and sidechain torsion angles (phi, psi and chi) on ligand binding within CDK2 is predicted using multiple regression analysis.

The dihedral angles of the 23 consensus active site residue in all 135 CDK2 proteins are considered as independent variables. The dependent variables are the activity values obtained by transforming the docking scores resulted during the docking experiments of CDK2

with their reference ligands. The relationship between dependent variable (activity) and the independent variables (various torsion angles of active site residues) was established by multiple linear regression analysis.

QSAR models were constructed on complete and training sets, respectively. Validation was done internally using leave-one-out (LOO) technique and externally by predicting the activities of validation set. The relationship between dependent variable (log1/C) and independent variables was established by linear multiple regression analysis using Tsar. Significant descriptors were chosen based on the statistical data of analysis. Statistical quality of the generated QSAR equation was judged based on the parameters like correlation coefficient (r), standard error of estimate (s), F-value, cross-validation r2 (q2) and predictive residual sum of squares (PRESS). Cross-validation was calculated using leave-one-out (LOO) technique over 7 random trials with F to leave and F to enter being 4 in F stepping to include the most significant variables in generating the QSAR model.

Predictive Ability of QSAR model

Predictive ability of the generated model was estimated externally by predicting the activities of validation set. This criterion may not be sufficient for a QSAR model to be truly predictive [13].

An additional condition for high predictive ability of QSAR model is based on external set cross-validation r2, (R2 cv,ext) and the regression of observed activities against predicted activities and vice versa for validation set, if the following conditions are satisfied [13-14].

$R^2cv,ext > 0.5$ (1)

$R2 > 0.6$ (2)

$(R2 – R02) / R2 < 0.1$ or $(R2 – R0'2) / R2 < 0.1$ (3)

$0.85 \_ k \_ 1.15$ or $0.85 \_ k' \_ 1.15$ (4)

Calculations relating to R2cv,ext, R02 and the slopes, *k* and *k'* are based on regression of observed values against predicted values and vice versa. They were discussed in detail in ref.13.

RESULTS AND DISCUSSION

To select the most influential backbone or sidechain torsion angles of aminoacids in active site towards ligand binding, multivariate regression analysis was performed. Initially regression analysis was carried out on all the 135 proteins, considering all the 91 independent variables.

Cross-validation was calculated using leave-one-out (LOO) technique over 1 random trial with F to leave being 2 and F to enter being 1 in F stepping to include the most significant variables in generating the model. This resulted in a 24 variables less significant model with regression coefficient, r: 0.778573, r^2: 0.606176 and Cross Validation, r^2(CV): 0.620674. In order to increase the predictive power of the regression model, the outliers are removed by calculating the standard residual values.

Outlier Detection

The data set was investigated for outliers by calculating the standard residuals. Standardized residuals greater than 2 and less than -2 are usually considered as outliers. Generally outliers have larger residuals than non-outliers. The following seven proteins are removed as outliers after initial regression test: 1H1P, 1PXK, 2R3F, 2R3H, 2R64, 2VTM, 2VTS respectively and can safely be excluded from the data set. Outliers were removed in order to obtain the best statistical result [15]. After removing the outliers the complete set was decreased to 128 proteins.

Regression Analysis after Removing Outliers

MLR technique was then applied on the remaining 128 proteins. The F test stepping values are taken as, F to enter value- 2, F to leave value -1 and random trails-1 (leave out one row cross validation) resulted in 15 variables model with regression coefficient, r: 0.765802 R^2: 0.586453and Cross Validation, r^2(CV): 0.977166.

The actual and predicted values in the result file of regression analysis after removing the outliers are selected to plot a correlation graph by taking actual values on x-axis and predicted values on y-axis. From the graph those proteins that lie closer and on the regression line are randomly selected as the test set. The following seven proteins are taken as the test set: 1F1N, 1HOV, 1KE5, 1PXN, 2AOC, 2J9M, 2VTR. Thus the complete set after removing the outliers is divided into training set and test set (validation set). This resulted in a 121 molecule training set and a 7 molecule validation (test) set. The test set is used for cross validation of results obtained by performing regression analysis on training set.

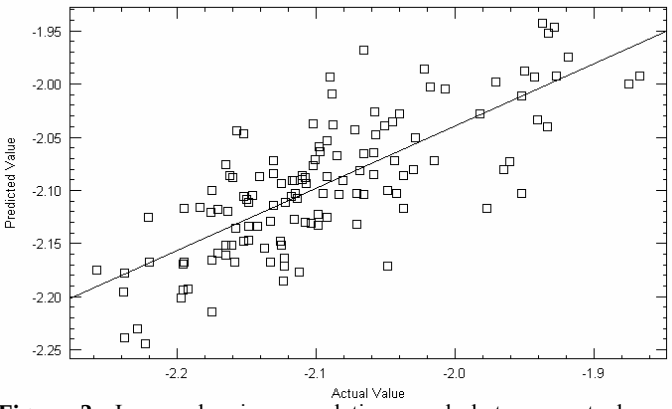

Figure 3: Image showing correlation graph between actual values and predicted values.

After deselecting the test set, regression analysis is again performed on the training set. Cross-validation was calculated using leave-one-out (LOO) technique over 1 random trial and the F to enter value being 2 and F to leave value being 1, to include the most significant variables in generating the regression model (Figure 3). The analysis resulted in 19 variables final model out of 85 independent variables considered and with the following statistical results: s value: 0.0510836, f value: 11.3275 , F probability: 1.02931e-018, Regression coefficient, r: 0.826295, r^2: 0.682764, Cross Validation,

r^2(CV): 0.823873, Residual Sum of Squares: 0.260953, Predictive Sum of Squares: 0.144879

The results showed that 19 out of 85 independent variables (torsion angles) are highly influential towards ligand binding with in CDK2 proteins. The statistically significant MLR model for training set was given below.

Log (1/dock score) = - 0.0011* Ile10 (phi)
- 0.0002* Ile10 (chi1)
- 0.0004* Glu12 (phi)
+ 0.0002* Thr14 (chi1)
- 5.5458e-005* Lys20 (chi1)
+ 8.0995e-005* Lys20 (chi2)
- 0.0024* Lys33 (phi)
+ 7.2626e-005* Lys33 (chi4)
+ 0.0001* Val64 (chi1)
- 0.0045* Phe80 (psi)
- 0.0021* Glu81 (psi)
- 0.0053* Phe82 (psi)
+ 0.0001* Phe82 (chi1)
- 0.0020* Leu83 (phi)
- 0.0004*Asp86 (chi2)
- 0.0003* Gln131 (chi3)
+ 0.0028* Asn132 (psi)
+ 0.0034* Leu134 (phi)
- 0.0029* Leu134 (chi1)
- 0.8562

The generated regression model indicates that an increase in torsion angles of Ile10 (phi), Ile10 (chi1), Glu12 (phi), Lys20 (chi1), Lys33 (phi), Phe80 (psi), Glu81 (psi), Phe82 (psi), Asp86 (chi2), Gln131 (chi3), Leu134 (chi1) contributes negatively to the activity. On the other hand, an increase in the torsion angles of Thr14 (chi1), Lys20 (chi2), Lys33 (chi1), Phe82 (chi1), Asn132 (psi), Len134 (phi) represents a positive contribution to the activity.

List of influential descriptors obtained from regression analysis are: Torsion angles: phi and chi1 of Ile10 ; phi of Glu12; chi1 of Thr14; chi1 and chi2 of Lys20 ; phi and chi4 of Lys33 ;chi1 of Val64 ; psi of Phe80 ; psi of Glu81 ; psi and chi1 of Phe82 ; phi of Leu83 ; chi2 of Asp86; chi3 of Gln131 ; psi of Asn132 ; phi and chi1 of Leu134.

Cross validation using statistical methods

A graph is plotted between predicted vs actual values and vice versa for test set data to obtain R^2 and R_0^2 (Figure 4). The calculated values of $(R^2 – R_0^2) / R^2$ = 0.0097 (which is less than 0.1) and k=0.9994 (which should between 0.85 and 1.15) shows that the results obtained are statistically significant.

CONCLUSION

Cyclin dependent kinases are known to exhibit different conformational states which affect the ligand binding within the active site region. In this study an attempt has been made by applying statistical techniques such as multiple linear regressions to study the influence of backbone and sidechain torsion angles on ligand binding. Regression analysis was carried out taking torsion angles of 23 consensus active site residues in all

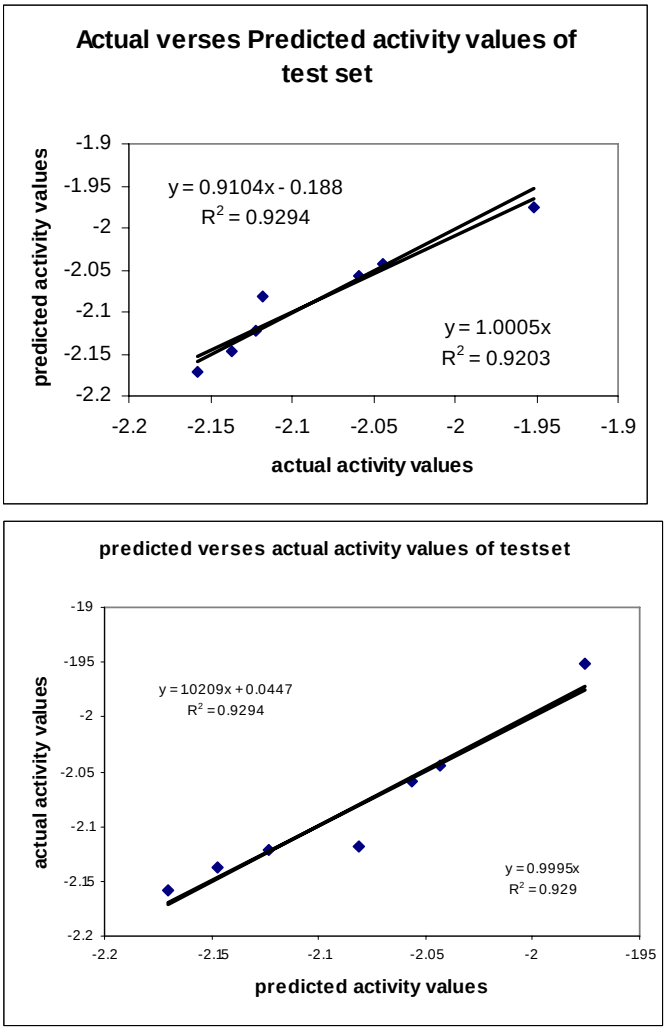

Figure 4: Predicted Vs Actual activities and vice versa of test set

135 CDK2 proteins as independent variables (descriptors) and the activity score of each protein as dependent variable to predict highly influential torsion angles. The initial analysis resulted in a 24 variable regression model with correlation coefficient r^2 =0.61 and cross validation r^2cv =0.64 respectively. After excluding seven proteins as outliers, a 19 variable final regression model on training set displayed a good correlation between actual and predicted values with correlation coefficient r^2 =0.68 and cross validation r^2cv =0.82 respectively. The model has passed the required cross validation tests (actual versus predicted and vice versa) on test set. Therefore this computational study carried out on 135 CDK2 proteins using MLR technique has revealed that 19 out of 85 torsion angles of amino acids in active site play a significant role in determining the ligand binding with in cdk2 proteins.

REFERENCES

1. Brun V, Legraverend M, Grierson DS (**2001**). Cyclin-dependent kinase (CDK) inhibitors: development of a general strategy for the construction of 2,6,9-trisubstituted purine libraries. Part 1 *Tetrahedron Lett* . *42* : 8161-8164.
2. Schang LM (**2002**). Cyclin-dependent kinases as cellular targets for antiviral drugs. *J Antimicrob Chemother.* *50* : 779-792.

3. Fischer PM, Endicott J, Meijer L (**2003**). Cyclin-dependent kinase inhibitors. *Prog Cell Cycle Res. 5* :235-248.
4. Sausville EA, Johnson J, Alley M et al (**2000**). Inhibition of CDKs as a therapeutic modality. *Ann N Y Acad Sci. 910*:207-221.
5. Ekholm SV and Reed SI (**2000**). Regulation of G(1) cyclin-dependent kinases in the mammalian cell cycle. *Curr Opin Cell Biol . 12* : 676-684.
6. Endicott JA, Noble ME, Tucker JA (**1999)**. Cyclin-dependent kinases: inhibition and substrate recognition. *Curr Opin Struct Biol. 9*: 738-744.
7. Gartel AL, Tyner AL (**2002**). The role of the cyclin-dependent kinase inhibitor p21 in apoptosis. *Mol Cancer Ther . 1:* 639-649.
8. Sherr CJ (**2000**). The Pezcoller lecture: cancer cell cycles revisited. *Cancer Res . 60*: 3689-3695.
9. http://www.cellsignal.com/reference/pathway/Cell_Cycle_G1S.html
10. Noble ME, Endicott JA, Johnson LN (**2004**). Protein kinase inhibitors: insights into drug design from structure. *Science 303*: 1800-1806
11. Canduri F and De Azevedo Jr WF (**2005)**. Structural Basis for Interaction of Inhibitors with Cyclin-Dependent Kinase 2 *Current Comput-Aided Drug Design. 1:*53-64.
12. Pindur U, Kim YS, Mehrabani F (**1999**). Advances in indolo[2,3-a]carbazole chemistry: design and synthesis of protein kinase C and topoisomerase I inhibitors. *Curr Med Chem .6:*29-69.
13. Golbraikh A and Tropsha A *(***2002)**. Beware of q2! *J. Mol. Graph. Model. 20*: 269-276.
14. Afantitis A, Melagraki G, Sarimveis H et al (**2006).** A Novel QSAR Model for Modeling and Predicting Induction of Apoptosis by 4-Aryl-4H-chromenes *Bioorg. Med. Chem. 14*: 6686-6694.
15. Kim D, Hong SI, Lee DS (**2006**). The quantitative structure-mutagenicity relationship of polycylic aromatic hydrocarbon metabolites *Mol. Sci.* 7: 556-570.

Received: 02 September 2011 Revised: 01 October 2011
Accepted: 01 October 2011 Online: 01 January 2012